Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 7666–7670, November 1985
Genetics

Variation among human 28S ribosomal RNA genes

(sequence/evolution/structure)

IRIS LAUDIEN GONZALEZ, JEROME L. GORSKI, THOMAS J. CAMPEN, D. J. DORNEY, JEANNE M. ERICKSON, JAMES E. SYLVESTER, AND ROY D. SCHMICKEL

Department of Human Genetics, University of Pennsylvania, Philadelphia, PA 19104

Communicated by Robert E. Forster, July 12, 1985

ABSTRACT **We report the complete 5025-base sequence of the human 28S rRNA gene. Variability within the species has been demonstrated by sequencing a variable region from six separately cloned genes. This region is one of three large subunit rRNA regions that show extreme sequence and size variation among species. The interspecies differences suggest species-specific functions for these sections, while the intraspecies heterogeneity indicates differences among ribosomes. Comparison of the human gene with a partial sequence from the chimpanzee 28S gene yields divergence rates for the two species: 0.8% for conserved regions of the gene and 3.7% for a variable region. The rapid divergence rates of variable regions in the ribosomal gene may permit answers to the question of time of separation of closely related species.**

We report the sequence of a single copy of human 28S rDNA (see Fig. 3) and compare it to the 1429-base-pair sequence of the chimpanzee between the *Bam*HI sites (human bases 1405–2839) and to five other copies of the human gene to examine both long- and short-term evolutionary changes. The segments we have chosen for comparison among humans and between human and chimpanzee include a slowly evolving and a rapidly changing sequence.

A great variation in length is characteristic of the large subunit rRNAs of various organisms. This gene ranges in size from 2900 bases in prokaryotes (23S) to 3392 bases in yeast and 5025 bases in man (28S). The published sequences of several large subunit rRNA genes have shown that size variation occurs by expansion or contraction of variable joining sequences at 10–12 specific points within the molecule (1–7). [There have been corrections to the sequence published in ref. 5 (I. G. Wool, personal communication).] These variable sequences have been called "divergent" or "D" domains by Hassouna *et al.* (4), "expansion segments" by Clark *et al.* (7), and "variable regions" by Chan *et al.* (5). While some of these sequences are highly variable in all species, others are conserved in the vertebrates. These regions alternate with regions that are highly conserved in prokaryotes and eukaryotes. Secondary structure models for eukaryotic rRNA are based on a conserved "prokaryotic-like" structure, with additional helices formed by the variable sequences. The variable sequences have several distinguishing features: (*i*) they are located in the same places in all eukaryotes studied; (*ii*) their size may vary greatly among species; (*iii*) there is greatly reduced homology among several of these sequences from one species to another, in contrast to the extensive homology in the conserved regions; (*iv*) several of these segments have features similar to those of the internal transcribed spacers: a high G+C content and a low adenine content, and they contain short repeated sequences; and (*v*) they can form self-contained double-helical structures. Some of these segments diverge rapidly, as evidenced by the considerable variation seen between related species, such as mouse (4) and rat (5, 6), and human and chimpanzee (this paper). This rapid divergence is most evident when comparing sequences within the human species. Transcribed sequences show variation from one individual to another and within a single individual. This is a rather unexpected finding, because it has been assumed that all rRNA molecules are identical.

METHODS

The human 28S rRNA gene was originally cloned from fetal liver and placental tissue into bacteriophage λ and subsequently subcloned into pBR322 plasmids in two segments: plasmid pA4 contained 7000 bases and included part of the 18S gene, both internal transcribed spacers, the 5.8S gene, and the section of the 28S gene 5′ to the *Eco*RI site; plasmid $pD_{ES}19$ contained the 28S gene segment 3′ to the *Eco*RI site and 400 bases of spacer (8).

For sequencing purposes, the 28S gene was divided as defined by the vertical lines on Fig. 1, each segment being suitably digested and subcloned into M13 vectors to be sequenced by the method of Sanger *et al.* (9). The clones for the 5′ segment before the *Hin*cII site were derived from *Sau*3A digestion of a larger clone that extended from the 18S gene to the first *Bam*HI site in the 28S gene; screening was done with an already identified clone that spans the *Hin*cII site. The last bases at the 5′ end were sequenced on the long *Hin*cII clone with the help of a sequencing primer prepared from one of the other clones. The 5′ end of the gene was defined by analogy to the published mouse and rat sequences. The *Hin*cII/*Bam*HI segment was purified from a gel, subjected to partial digestion by *Hpa* II, and ligated into M13 vectors in both orientations. The *Bam*HI segment was digested with *Hpa* II, *Sau*3A, *Xma* I, and *Taq* I to obtain the subclones sequenced. The asterisk on the sequencing strategy (Fig. 1) identifies a segment sequenced within a larger clone with the help of a specially synthesized primer; this primer was also used for sequencing the same segment of five other copies of the human gene. The *Bam*HI/*Eco*RI segment was subcloned after digestions with *Xma* I, *Pvu* II, *Hin*cII, *Alu* I, and *Sst* I. The section 3′ to the *Eco*RI site was subcloned after *Rsa* I, *Sau*3A, and partial *Hpa* II digestions. The 3′ end of the gene was defined by S1 nuclease mapping (unpublished results).

Origin of the six different human genes: A1 and A6 were cloned from placenta 1; A5 was from placenta 2; A2, A3, and A4 were cloned from the same fetal liver.

The 1.4-kilobase chimpanzee gene segment was originally cloned in bacteriophage λ and in pBR322 plasmids. The segment was removed from the vector by *Bam*HI digestion and was purified. Digestions with *Sau*3A, *Hpa* II, and *Sma* I were followed by ligation into appropriate M13 vectors.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

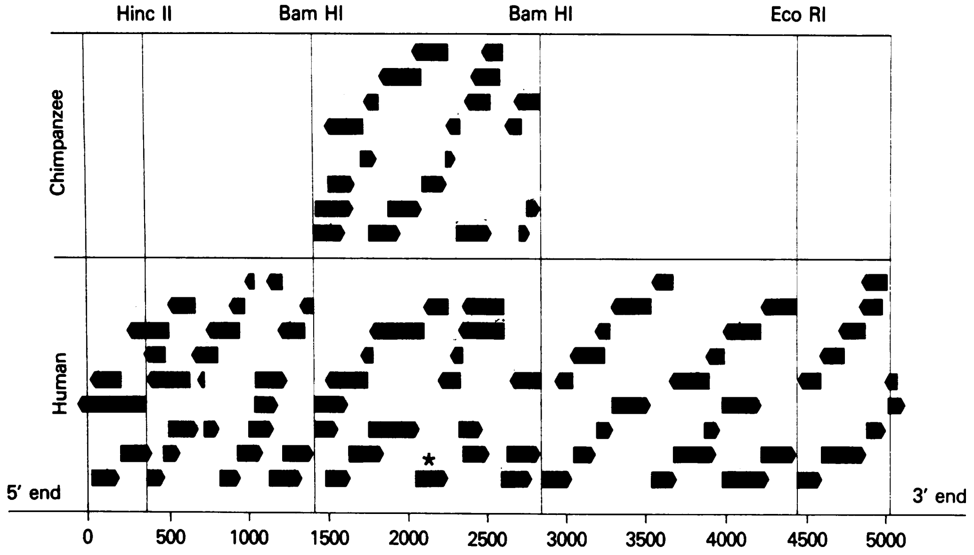

FIG. 1. Sequencing strategy for the human 28S rRNA gene. Arrows represent individual clones in M13, which are sequenced in the direction of the arrow. DNA was sequenced in each direction with overlapping segments.

RESULTS AND DISCUSSION

Differences among individually cloned human genes were found in a rapidly evolving segment that corresponds to the mouse D6 region (4). Major differences between the human and chimpanzee sequences are also found in this region. The structures for the D6 regions of humans, chimpanzees, and other organisms are shown in Fig. 2 and are characterized in Table 1. D6 is a high G+C-containing sequence that has been inserted into and replaced the top of the hairpin bases 1164–1185 of *Escherichia coli* (10, 11) or its homologue bases 1337–1363 in *Saccharomyces* (1). Thus, an expanded hairpin structure is generated with its base partially conserved. The sequence 5′ to the expanded structure is highly conserved with respect to both sequence and secondary structure. The sequence 3′ to the expanded structure forms a secondary structure very similar to *E. coli* bases 1198–1247 and *Saccharomyces* bases 1373–1430, although the nucleotide sequence is not as highly conserved.

As shown in Table 1 and in Fig. 2, *Xenopus* has a D6 region of only 41 bases, but the conserved DNA sequences at the base of the stem are present. *Physarum* has a D6 region that differs markedly from those of the other eukaryotes in base composition: it has only 53% G+C. The flanking sequences are poorly conserved, even as compared to yeast.

Six human D6 regions were sequenced and yielded the three versions presented in Fig. 2. The differences are concentrated in the segment between bases 2129 and 2166, which shows as much variation among humans as between human and chimpanzee, so that intraspecies variation and interspecies differences become one. The finding that six fragments gave three different sequences demonstrates a high frequency of variation in the rRNA gene population and suggests heterogeneity among rRNA transcripts. The differences in this region are associated with multiplication of existing sequences. For instance, at bases 2129–2149, the human has either $(GGC)_5$ or $(GGC)_7$, while the chimpanzee has $(GGC)_5$ and GGT. At bases 2154–2159, human has either $(GT)_2$ or $(GT)_3$, while chimpanzee has $(GT)_1$. These differences can be generated by unequal homologous exchange between sister chromatids. The same mechanisms can generate the differences between human and chimpanzee in the regions 2099–2116 (alternating purine/pyrimidine tract), 2173–2183 (different length of C string), and at 2209 (set of CCPu repeats in chimpanzee only). The mouse and rat sequences also show this mechanism at work, with different numbers of CGGPuA tandem repeats (4). The two published rat sequences (5, 6) are identical, except for one base at human position 2235.

In spite of the overall similarity, the D6 region of the two primates accumulates a relatively large number of differences when compared to the rest of the 1429 bases between the two *Bam*HI sites in the 28S gene (Fig. 3). In D6, there are 41/198 (20.7%) differences when counted base by base and 15/198 (7.5%) differences when one scores by the minimal number of deletions, insertions, or base changes between the two segments. There are 5 differences for the rest of the fragment; 3 of these are clustered in a small (20-base) variable region corresponding to the mouse D4 region. Two are in the remaining 1211 conserved bases, giving 0.16% differences. Thus, we find two distinctly different rates of change: a 0.16% rate of change for a conserved area and a 7.5–20% change for the variable sequence. (The small variable region will not be included.) Comparisons between globin gene exons and introns, and between globin gene "amino acid replacement mutations" and "silent mutations" (12) also show different rates of change data. In the one case, introns showed much higher divergence than exons. In the second case, silent mutations in the translated regions far exceeded replacement mutations.

The apparent difference in frequency of silent and amino acid replacement mutations is due to their effects on fitness: one type of mutation can be retained, the other cannot always be tolerated. It is not possible to use the same rationale to explain the divergence rate differences between the conserved and variable sequences of the ribosomal genes. A single ribosomal gene inactivation would not affect the fitness of the organism because there are 300 ribosomal genes. Even the loss of 20% of the human rRNA genes by a Robertsonian translocation shows no phenotypic effect. In *Drosophila*, >50% of the ribosomal genes must be lost for a phenotypic effect to be noticeable (13). The model of concerted evolution may be invoked to explain both the high level of conservation and the high level of divergence (14). According to this model, similar genes in an organism can correct a mutation by unequal homologous exchange on sister chromatids or by gene conversion (15, 16). Both mechanisms require repeated gene sequences for the correction to take place. In instances where repeated functioning genes are adjacent to each other, as in the case of α1- and α2-globin, the genes maintain identity via the correction mechanism (17). Divergence can result from the same mechanism, since unequal homologous ex-

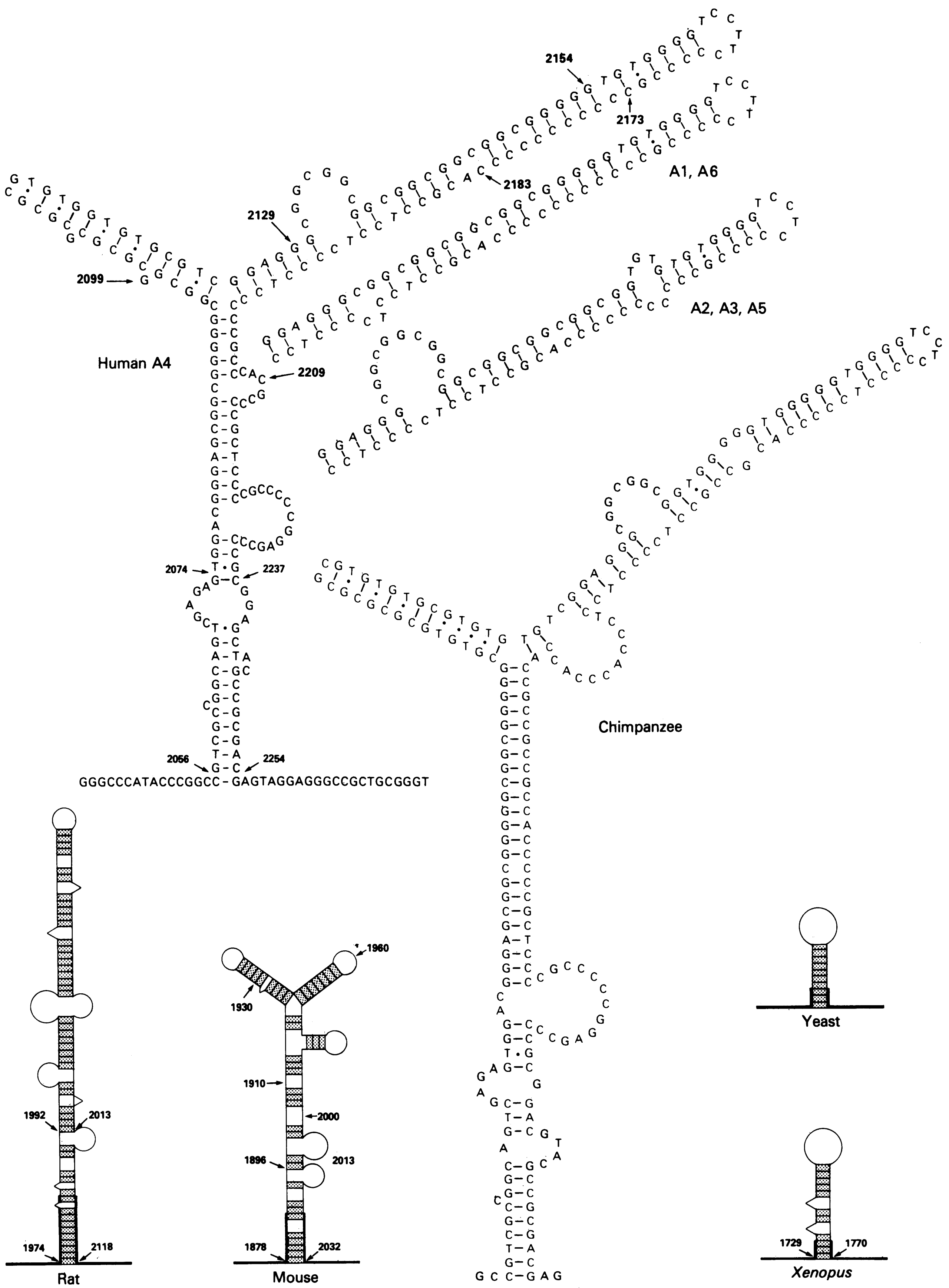

FIG. 2. Postulated secondary structures for the D6 regions. Thicker lines at base of stems indicate conserved regions. The structures were drawn so as to maximize intramolecular base pairing, which is indicated with shaded boxes.

change yields unequal reciprocal products. This can operate on whole genes, such as the rRNA tandem repeats, resulting in higher or lower numbers of repeats. It can also operate on short repeats present within a gene, leading to expansion or

Table 1. Comparison of D6 regions

	Start	End	Length, bases	% G+C	% A
			Complete hairpin		
Human	2056	2254	198	84	6.5
Chimpanzee	—	—	211	82.5	7.1
Mouse	1878	2032	154	83.1	6.5
Rat	1974	2118	144	87	6.9
Xenopus	1730	1770	41	80.5	7.3
Physarum	1392	1452	60	53	25
Yeast	1336	1363	27	55.5	18.5
			Inserted region		
Human	2074	2237	163	86	4.3
Chimpanzee	—	—	176	84.1	5.1
Mouse	1896	2013	117	86	3.3
Rat	1992	2099	107	90.7	3.7

contraction of a gene region; these expanding/contracting regions are more likely to be present in sequences that do not code for proteins and that do not interfere with RNA structure.

One can use the differences in the conserved regions to calculate the unit evolutionary periods (UEP) of the sequences studied. The UEP has been defined as the number of years for fixing 1% divergence (18). For the 1211-base-pair sequence, the two differences divided by the combined target area (2422) give a divergence of 0.08%, over the 5 million years since human–chimpanzee separation. From this, one obtains a UEP of 62.5 million years for this rDNA fragment. This is a lower divergence rate and higher UEP than obtained for the human and chimpanzee duplicated α-globin genes, which is 0.45% and 11 million years, respectively (19). The divergence rate in the rDNA variable regions studied here is much higher: (*i*) Among human genes, the variable region is restricted to 33 bases (because at least two GGC repeats must be intact to allow unequal homologous recombination), so that the divergence rate is 3/66 = 0.45%. (*ii*) Between the human and chimpanzee genes it is 15/396 = 3.7%, yielding a UEP of 1.3 million years. The comparison of human to chimpanzee sequences shows a closer molecular relationship than any other known species comparison.

In comparing the rodents to humans, we find 12 differences between mouse and human and 14 differences between rat and human in the 1211-base conserved sequence, to give divergences of 0.5% and 0.57%. Seven differences are shared by rat and mouse. Curiously, there are 13 differences in the conserved sequence of mouse and rat, yielding a divergence of 0.53%. This would indicate that the evolutionary split of mouse and rat is as ancient as that of the rodent and primate. This conclusion could be incorrect because of the effect of different generation times of the two kinds of mammal and the differences in types of mutations may not be comparable.

A comparison of secondary structures for D6 regions of rRNA genes other than human and chimpanzee has been published (4); all were postulated to share the same basic structure, which gradually becomes larger in the higher organisms. We favor a different structure of the large subunit D6 region. The secondary structures of the D6 regions of six species (Fig. 2) show a partially conserved stem. The first 4 base pairs of each species' hairpin are homologous to those of yeast. Homology extends farther between primate and rodent D6 structures, which are also identical for the next 4 bases and markedly homologous for the following 6 base pairs. Above these 14 base pairs, in both primate and rodent structures, is a single-stranded region that is A+G-rich, as is characteristic of the single-stranded ends of rRNA hairpins. Beyond this region are highly divergent expansion sequences, which can form straight or branched double-stranded structures. Human and chimpanzee share a similar structure, with a Y-shaped expansion segment. The stems are nearly identical up to human base 2085, after which the length and structures differ. The left branch of the "Y" shows some

```
CGCGACCTCA GATCAGACGT GGCGACCCGC TGAATTTAAG CATATTAGTC AGCGGAGGAA   60
AAGAAACTAA CCAGGATTCC CTCAGTAACG GCGAGTGAAC AGGGAAGAGC CCAGCGCCGA  120
ATCCCCGCCC CGCGGGGCGC GGGACATGTG GCGTACGGAA GACCCGCTCC CCGGCGCCGC  180
TCGTGGGGGG CCCAAGTCCT TCTGATCGAG GCCCAGCCCG TGGACGGTGT GAGGCCGGTA  240
GCGGCCGGCG CGCGCCCGGG TCTTCCCGGA GTCGGGTTGC TTGGGAATGC AGCCCAAAGC  300
GGGTGGTAAA CTCCATCTAA GGCTAAATAC CGGCACGAGA CCGATAGTCA ACAAGTACCG  360
TAAGGGAAAG TTGAAAAGAA CTTTGAAGAG AGAGTTCAAG AGGGCGTGAA ACCGTTAAGA  420
GGTAAACGGG TGGGGTCCGC GCAGTCCGCC CGGAGGATTC AACCCGGCGG CGGGTCCGGC  480
CGTGTCGGCG GCCCGGCGGA TCTTTCCCGC CCCCCGTTCC TCCCGACCCC TCCACCCGCC  540
CTCCCTTCCC CCGCCGCCCC TCCTCCTCCT CCCCGGAGGG GGCGGGCTCC GGCGGGTGCG  600
GGGGTGGGCG GGCGGGGCCG GGGGTGGGGT CGGCGGGGGA CCGTCCCCCG GACCGGCGAC  660
CGGCCGCCGC CGGGCGCATT TCCAGGCGGT GCGCCGCGAC CGGCTCCGGG ACGGCTGGGA  720
AGGCCCGGCG GGGAAGGTGG CTCGGGGGGC CCCGTCCGTC CGTCCGTCCT CCTCCTCCCC  780
CGTCTCCGCC CCCCGGCCCC GCGTCCTCCC TCGGGAGGGC GCGCGGGTCG GGGCGGCGGC  840
GGCGGCGGCG GTGGCGGCGG CGGCGGGGGC GGCGGGACCG AAACCCCCCC CGAGTGTTAC  900
AGCCCCCCCG GCAGCAGCAC TCGCCGAATC CCGGGGCCGA GGGAGCGAGA CCCGTCGCCG  960
CGCTCTCCCC CCTCCCGGCG CCCACCCCCG CGGGAATCCC CGCGAGGGGG GTCTCCCCCG 1020
GCGCGGCGCC GGCGTCTCCT CGTGGGGGGG CCGGGCCACC CCTCCCACGG CGCGACCGCT 1080
CTCCCACCCC TCCTCCCCGC GCCCCCGCCC CGGCGACGGG GGGGGTGCCG CGCGCGGGTC 1140
GGGGGGCGGG GCGGACTGTC CCCAGTGCGC CCCGGGCGGG TCGCGCCGTC GGGCCCGGGG 1200
GAGGTTCTCT CGGGGCCACG CGCGCGTCCC CCGAAGAGGG GGACGGCGGA GCGAGCGCAC 1260
GGGGTCGGCG GCGACGTCGG CTACCCACCC GACCCGTCTT GAAACACGGA CCAAGGAGTC 1320
TAACACGTGC GCGAGTCGGG GGCTCGCACG AAAGCCGCCG TGGCGCAATG AAGGTGAAGG 1380
CCGGCGCGCT CGCCGGCCGA GGTGGGATCC CGAGGCCTCT CCAGTCCGCC GAGGGGCACC 1440
                                 ▲
ACCGGCCCGT CTCGCCCGCC GCGCCGGGGA GGTGGAGCAC GAGCGCACGT GTTAGGACCC 1500
GAAAGATGGT GAACTATGCC TGGGCAGGGC GAAGCCAGAG GAAACTCTGG TGGAGGTCCG 1560
TAGCGGTCCT GACGTGCAAA TCGGTCGTCC GACCTGGGTA TAGGGGCGAA AGACTAATCG 1620
AACCATCTAG TAGCTGGTTC CCTCCGAAGT TTCCCTCAGG ATAGCTGGCG CTCTCGCAGA 1680
                                                     C
CCCGACGCAC CCCCGCCACG CAGTTTTATC CGGTAAAGCG AATGATTAGA GGTCTTGGGG 1740
           A   C C
CCGAAACGAT CTCAACCTAT TCTCAAACTT TAAATGGGTA AGAAGCCCGG CTCGCTGGCG 1800
TGGAGCCGGG GTGGAATGCG AGTGCCTAGT GGGCCACTTT TGGTAAGCAG AACTGGCGCT 1860
GCGGGATGAA CCGAACGCCG GGTTAAGGCG CCCGATGCCG ACGCTCATCA GACCCCAGAA 1920
AAGGTGTTGG TTGATATAGA CAGCAGGACG GTGGCCATGG AAGTCGGAAT CCGCTAAGGA 1980
GTGTGTAACA ACTCACCTGC CGAATCAACT AGCCCTGAAA ATGGATGGCG CTGGAGCGTC 2040
GGGCCCATAC CCGGCCGTCG CCGGCAGTCG AGAGTGGACG GGAGCGGCGG GGGCGGCGGC 2100
GCGCGCGCGC GTGTGGTGTG CGTCGGAGGG CGGCGGCGGC GGCGGCGGCG GGGGTGTGGG 2160
GTCCTTCCCC CGCCCCCCCC CCCACGCCTC CTCCCCTCCT CCCGCCCACG CCCCGCTCCC 2220
CGCCCCCGGA GCCCCGCGGA GCTACGCCGC GACGAGTAGG AGGGCCGCTG CGGTGAGCCT 2280
TGAAGCCTAG GGCGCGGGCC CGGGTGGAGG CCGCCGCAGG TGCAGATCTT GGTGGTAGTA 2340
GCAAATATTC AAACGAGAAC TTTGAAGGCC GAAGTGGAGA AGGGTTCCAT GTGAACAGCA 2400
GTTGAACATG GGTCAGTCGG TCCTGAGAGA TGGGCGAGCG CCGTTCCGAA GGGACGGGCG 2460
ATGGCCTCCG TTGCCCTCGG CCGATCGAAA GGGAGTCGGG TTCAGATCCC CGAATCCGGA 2520
GTGGCGGAGA TGGGCGCCGC GAGGCGTCCA GTGCGGTAAC GCGACCGATC CCGGAGAAGC 2580
CGGCGGGAGC CCCGGGGAGA GTTCTCTTTT CTTTGTGAAG GGCAGGGCGC CCTGGAATGG 2640
GTTCGCCCCG AGAGAGGGGC CCGTGCCTTG GAAAGCGTCG CGGTTCCGGC GGCGTCCGGT 2700
GAGCTCTCGC TGGCCCTTGA AAATCCGGGG GAGAGGGTGT AAATCTCGCG CCGGGCCGTA 2760
CCCATATCCG CAGCAGGTCT CCAAGGTGAA CAGCCTCTGG CATGTTGGAA CAATGTAGGT 2820
AAGGGAAGTC GGCAAGCCGG ATCCGTAACT TCGGGATAAG GATTGGCTCT AAGGGCTGGG 2880
TCGGTCGGGC TGGGGCGCGA AGCGGGGCTG GGCGCGCGCC GCGGCTGGAC GAGGCGCGCG 2940
CCCCCCCCAC GCCCGGGGCA CCCCCCTCGC GGCCCTCCCC CGCCCCACCC GCGCGCGCCG 3000
CTCGCTCCCT CCCCACCCCG CGCCCTCTCT CTCTCTCTCT CCCCCGCTCC CCGTCCTCCC 3060
CCCTCCCCGG GGGAGCGCCG CGTGGGGGCG CGGCGGGGGG AGAAGGGTCG GGGCGGCAGG 3120
GGCCGCGCGG CGGCCGCCGG GGCGGCCGGC GGGGGCAGGT CCCCGCGAGG GGGGCCCCGG 3180
GGACCCGGGG GGCCGGCGGC GGCGCGGACT CTGGACGCGA GCCGGGCCCT TCCCGTGGAT 3240
CGCCCCAGCT GCGGCGGGCG TCGCGGCCGC CCCCGGGGAG CCCGGCGGCG GCGCGGCGCG 3300
CCCCCCACCC CCACCCCACG TCTCGGTCGC GCGCGCGTCC GCTGGGGGCG GGAGCGGTCG 3360
GGCGGCGGCG GTCGGCGGGC GGCGGGGCGG GGCGGTTCGT CCCCCCGCCC TACCCCCCCG 3420
GCCCCGTCCG CCCCCCGTTC CCCCCTCCTC CTCGGCGCGC GGCGGCGGCG GCGGCAGGCG 3480
GCGGAGGGGC CGCGGGCCGG TCCCCCCCGC CGGGTCCGCC CCCGGGGCCG CGGTTCCGCG 3540
CGCGCCTCGC CTCGGCCGGC GCCTAGCAGC CGACTTAGAA CTGGTGCGGA CCAGGGGAAT 3600
CCGACTGTTT AATTAAAACA AAGCATCGCG AAGGCCCGCG GCGGGTGTTG ACGCGATGTG 3660
ATTTCTGCCC AGTGCTCTGA ATGTCAAAGT GAAGAAATTC AATGAAGCGC GGGTAAACGG 3720
CGGGAGTAAC TATGACTCTC TTAAGGTAGC CAAATGCCTC GTCATCTAAT TAGTGACGCG 3780
CATGAATGGA TGAACGAGAT TCCCACTGTC CCTACCTACT ATCCAGCGAA ACCACAGCCA 3840
AGGGAACGGG CTTGGCGGAA TCAGCGGGGA AAGAAGACCC TGTTGAGCTT GACTCTAGTC 3900
TGGCACGGTG AAGAGACATG AGAGGTGTAG AATAAGTGGG AGGCCCCCGG CGCCCCCCCG 3960
GTGTCCCCGC GAGGGGCCCG GGGCGGGGTC CGCGGCCCTG CGGGCCGCCG GTGAAATACC 4020
ACTACTCTGA TCGTTTTTTC ACTGACCCGG TGAGGCGGGG GGGCGAGCCC GAGGGGCTCT 4080
CGCTTCTGGC GCCAAGCGCC CGCCCGGCCG GGCGCGACCC GCTCCGGGGA CAGTGCCAGG 4140
TGGGGAGTTT GACTGGGGCG GTACACCTGT CAAACGGTAA CGCAGGTGTC CTAAGGCGAG 4200
CTCAGGGAGG ACAGAAACCT CCCGTGGAGC AGAAGGGCAA AAGCTCGCTT GATCTTGATT 4260
TTCAGTACGA ATACAGACCG TGAAAGCGGG GCCTCACGAT CCTTCTGACC TTTTGGGTTT 4320
TAAGCAGGAG GTGTCAGAAA AGTTACCACA GGGATAACTG GCTTGTGGCG GCCAAGCGTT 4380
CATAGCGACG TCGCTTTTTG ATCCTTCGAT GTCGGCTCTT CCTATCATTG TGAAGCAGAA 4440
TTCGCCAAGC GTTGGATTGT TCACCCACTA ATAGGGAACG TGAGCTGGGT TTAGACCGTC 4500
GTGAGACAGG TTAGTTTTAC CCTACTGATG ATGTGTTGTT GCCATGGTAA TCCTGCTCAG 4560
TACGAGAGGA ACCGCAGGTT CAGACATTTG GTGTATGTGC TTGGCTGAGG AGCCAATGGG 4620
GCGAAGCTAC CATCTGTGGG ATTATGACTG AACGCCTCTA AGTCAGAATC CCGCCCAGGC 4680
GAACGATACG GCAGCGCCGC GGAGCCTCGG TTGGCCTCGG ATAGCCGGTC CCCCGCCTGT 4740
CCCCGCCGGC GGGCCGCCCC CCCCTCCACG CGCCCCGCCG CGGGAGGGCG CGTGCCCCGC 4800
CGCGCGCCGG GACCGGGGTC CGGTGCGGAG TGCCCTTCGT CCTGGGAAAC GGGGCGCGGC 4860
CGGAAAGGCG GCCGCCCCCT CGCCCGTCAC GCACCGCACG TTCGTGGGGA ACCTGGCGCT 4920
AAACCATTCG TAGACGACCT GCTTCTGGGT CGGGGTTTCG TACGTAGCAG AGCAGCTCCC 4980
TCGCTGCGAT CTATTGAAAG TCAGCCCTCG ACACAAGGGT TTGTC                5025
```

FIG. 3. Sequence of the human 28S rRNA gene. The two *Bam*HI sites that define the segment used for human–chimpanzee comparison are underlined at bases 1405 and 2839. The broken line indicates the variable D6 region. Differences within this DNA segment are given in Fig. 2. Differences between human and chimpanzee in the *Bam*HI fragment are indicated in bold characters below the human sequence. ▲ under a letter indicates that base is not found in the chimpanzee.

sequence difference between human and chimpanzee, while maintaining the same length. The right branch contains the region that is variable among human genes. The mouse and rat structures are not interchangeable. The two rodents differ noticeably in this sequence and this is reflected in the postulated secondary structures.

The large subunit rRNAs show many structural arrangement and processing differences among the diverse kingdoms. An RNA sequence between the 5.8S and the 28S gene (internal transcribed spacer) is found in eukaryotes and in some mitochondria, but not in prokaryotes, and it is removed from the ribosomal transcript by processing; the resulting rRNAs are not ligated (20, 21). In the processing of *Tetrahymena* rRNA, an internal sequence is removed and the remaining sequences are ligated by an autocatalytic process (22). In *Drosophila*, the presence of a long internal sequence appears to prevent transcription and causes gene inactivation (23). *Physarum* large subunit genes contain introns that are spliced out (2). Mitochondria and chloroplasts are believed to be derived from prokaryotic ancestors that have intronless genes; their rRNA genes, often shorter than those of prokaryotes, retain introns (24, 25). Since the archetype ribosome is not available for study, it is not possible to determine whether a given sequence has been present from earliest times or if it is a late development. One can view the evolution of the large subunit rRNA in either of two ways: *(i)* It is the result of growth, starting from a minimal ancestral form that combined several translation-related functions into one structure. The growth only occurred at specific sites that would not interfere with the ribosome function. *(ii)* It evolved to a minimal and a maximal form from an intermediate-sized prototype, which was a conglomerate of shorter functional molecules joined by "spacers" (26). In this second view, the prokaryotic genes may be a streamlined version of the bulkier archetype. The eukaryotic line has not eliminated the joining sequences. In eukaryotes, and in the "prokaryote-like" organelles, it appears that there is no strong distinction between internal transcribed spacers, self-splicing introns, enzymatically spliced introns, inactivating inserts, or "expansion segments." It is likely that all of these represent ancient joining sequences and that a given type of internal sequence may change classification as a consequence of evolutionary changes. We favor the second model for the origin of variable joining segments such as D6: the very ancient origin and the freedom to mutate have led to the great differences found between species, with respect to both sequence and secondary structure.

The ribosomal gene permits both short-term and long-term evolutionary studies: the comparison of human and chimpanzee sequences has shown that evolution can proceed at different rates in different sections of a transcribed gene. The variable joining segments of the ribosomal gene can be used to study evolutionary relationships of very closely related species or of systematic differences within a species: here they show the close relationship between normal variation within a species and the changes associated with speciation. If some of the variable sequences can influence gene expression and function, the mechanism of unequal homologous exchange could produce rapid and dramatic phenotype differences, with little alteration of the sequence content. It is possible that generalized regulatory rearrangements could be responsible for the large phenotypic differences between human and chimpanzee, while the genotype differences appear to be very small (27, 28).

This work was supported in part by National Institutes of Health Grant HD 16930 to R.D.S. I.L.G. and J.L.G. were supported by National Institutes of Health Postdoctoral Training Grant GM 07511. J.L.G. is now supported by National Institutes of Health Physician Scientist Award K-11-HD 00610.

1. Veldman, G. M., Klootwijk, J., de Regt, V. C. H. F., Planta, R., Branlant, C., Krol, A. & Ebel, J.-P. (1981) *Nucleic Acids Res.* **9**, 6935–6952.
2. Otsuka, T., Nomiyama, H., Yoshida, H., Kukita, T., Kuhara, S. & Sakaki, Y. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3163–3167.
3. Ware, V. C., Tague, B. W., Clark, C. G., Gourse, R. L., Brand, R. C. & Gerbi, S. A. (1983) *Nucleic Acids Res.* **11**, 7796–7817.
4. Hassouna, N., Michot, B. & Bachellerie, J.-P. (1984) *Nucleic Acids Res.* **12**, 3563–3599.
5. Chan, Y.-L., Olvera, J. & Wool, I. G. (1983) *Nucleic Acids Res.* **11**, 7819–7831.
6. Hadjiolov, A. A., Georgiev, O. I., Nosikov, V. V. & Yavachev, L. P. (1984) *Nucleic Acids Res.* **12**, 3877–3894.
7. Clark, C. G., Tague, B. W., Ware, V. C. & Gerbi, S. A. (1984) *Nucleic Acids Res.* **12**, 6197–6220.
8. Erickson, J. M., Rushford, C. L., Dorney, D. J., Wilson, G. N. & Schmickel, R. D. (1981) *Gene* **16**, 1–9.
9. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
10. Noller, H. F., Kop, J. A., Wheaton, V., Brosius, J., Gutell, R. R., Kopylov, A. M., Dohme, F. & Herr, W. (1981) *Nucleic Acids Res.* **9**, 6167–6189.
11. Branlant, C., Krol, A., Machatt, M. A., Pouyet, J., Ebel, J.-P., Edwards, K. & Kossel, H. (1981) *Nucleic Acids Res.* **9**, 4303–4320.
12. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., De Riel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 653–668.
13. Shermoen, A. W. & Kiefer, B. I. (1975) *Cell* **4**, 275–280.
14. Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O. & Zimmer, E. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7323–7327.
15. Tartof, K. D. (1975) *Annu. Rev. Genet.* **9**, 355–365.
16. Slightom, J. L., Blechl, A. E. & Smithies, O. (1980) *Cell* **21**, 627–638.
17. Liebhaber, S. A., Goosens, M. & Kan, Y. W. (1981) *Nature (London)* **290**, 26–29.
18. Wilson, A. C., Carlson, S. S. & White, T. J. (1977) *Annu. Rev. Biochem.* **46**, 573–639.
19. Liebhaber, S. A. & Begley, K. (1983) *Nucleic Acids Res.* **11**, 8915–8929.
20. Perry, R. P. (1976) *Annu. Rev. Biochem.* **45**, 605–629.
21. Seilhammer, J. J., Glutell, R. R. & Cummings, D. J. (1984) *J. Biol. Chem.* **259**, 5173–5181.
22. Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. & Cech, T. R. (1982) *Cell* **31**, 147–157.
23. Long, E. O. & Dawid, I. B. (1979) *Cell* **18**, 1185–1196.
24. Rochaix, J.-D. & Darlix, J.-L. (1982) *J. Mol. Biol.* **159**, 383–395.
25. Bos, J. L., Heyting, C., Borst, P., Arnberg, A. C. & Van Bruggen, E. F. J. (1978) *Nature (London)* **275**, 336–337.
26. Blake, C. (1983) *Nature (London)* **306**, 535–537.
27. King, M.-C. & Wilson, A. C. (1975) *Science* **188**, 107–116.
28. Sibley, C. G. & Ahlquist, J. E. (1984) *J. Mol. Evol.* **20**, 2–15.