Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 9625–9629, December 1988
Evolution

The human somatic cytochrome *c* gene: Two classes of processed pseudogenes demarcate a period of rapid molecular evolution

(accelerated evolution/respiratory chain/primate lineage)

MARK J. EVANS AND RICHARD C. SCARPULLA

Department of Molecular Biology, Northwestern University Medical School, Chicago, IL 60611

Communicated by Fred Sherman, September 9, 1988 (received for review April 8, 1988)

ABSTRACT **We have isolated and determined the DNA sequences of the human somatic cytochrome *c* gene (HCS) and 11 processed pseudogenes. HCS is the functional homologue to the previously characterized rat somatic gene because it correctly encodes the human heart protein, is present in single copy in the human genome, is nearly identical in both size and intron/exon structure to rodent somatic genes, and shares a high degree of sequence homology with its rat counterpart including a well-conserved promoter region (77% over 250 nucleotides). In contrast to the rodent system, however, where the known pseudogenes all originated from a locus encoding the present day cytochrome *c*, the human pseudogenes are of two types. A predominant class of older pseudogenes came from a progenitor of HCS that encoded an ancestral form of the protein, while a second group of only a few young pseudogenes originated from a recent parent of HCS that encoded the current cytochrome *c* polypeptide. These two distinct classes of human pseudogenes provide a molecular record of the history of cytochrome *c* evolution in primates and demarcate a short period of rapid evolution of the functional gene.**

Cytochrome *c* is an essential respiratory protein present as a single genetically determined isoform in mammalian somatic tissues (reviewed in refs. 1 and 2). In contrast, mammalian genomes contain families of ≈35 different cytochrome *c* sequences (3). In the rat, nearly all of these are processed pseudogenes derived from multiple mRNAs transcribed from a single intron-containing functional somatic gene (4, 5). The only other functional genetic variant of cytochrome *c* in mammals is a testis-specific protein (6) encoded in the rat by a single-copy gene with no homology to the somatic gene outside of the coding region and no related pseudogenes (7).

The protein sequence of cytochrome *c* has been determined for numerous eukaryotic species (referenced in ref. 8), allowing the construction of an extensive evolutionary lineage (9). These sequences were instrumental in the original formulation of the molecular clock hypothesis that the rate of molecular evolution is approximately constant among different evolutionary lineages (10). In mammals, cytochrome *c* has undergone very little evolutionary change except in the primate lineage, where the protein has accumulated an unusually large number of replacement changes. Parsimony analysis has suggested that an acceleration in the rate of cytochrome *c* molecular evolution occurred early in the primate lineage between the eutherian ancestor and the monkey–hominid ancestor (9). Here the structures of the human somatic cytochrome *c* gene (HCS) and 11 intronless pseudogenes provide a unique molecular record of accelerated gene evolution.

MATERIALS AND METHODS

A human testis cDNA library (Clontech Laboratories, Palo Alto, CA) and human genomic libraries [either a fetal liver *Hae* II/*Alu* I library (11) or an embryonic fibroblast *Eco*RI library] were screened as described (12). Human HC-type pseudogenes were obtained by using the rat cytochrome *c* coding region as a hybridization probe. Human HS-type pseudogenes were isolated by using the oligonucleotide HCS-1 complementary to codons 10–16 of the functional gene HCS. All DNA sequences were determined on both strands by the dideoxy chain-termination method (13). Polyadenylylated RNA was isolated from human tissues and HeLa cells as described (7). Genomic hybridizations to human placental DNA (Sigma) were performed as described (3, 5, 14).

RESULTS

Isolation and Characterization of the HCS Gene. The numerous processed cytochrome *c* pseudogenes in mammalian genomes (3, 5) can obscure the identification and cloning of functional cytochrome *c* genes. While screening a human testis cDNA library with the rat somatic cytochrome *c* coding region, a cloned cDNA of a partially spliced somatic cytochrome *c* transcript was isolated. The intron located in the 5′ untranslated region of this cDNA was used as a probe to distinguish the corresponding HCS gene from the numerous intronless pseudogenes in a human genomic library. The structural characteristics (summarized in Fig. 1) and nucleotide sequence of HCS (Figs. 2 and 3) establish it as the functional homologue to the previously characterized rat somatic gene RC4 (5, 12). First, HCS correctly encodes the protein purified from human heart (15). Second, the mRNA-encoding sequences are highly conserved between human and rat (80% identity) and both the coding and 3′ untranslated regions can be used to detect the numerous processed pseudogenes in the human genome (Fig. 4*B*). Third, human and rat genes are nearly identical in both size and intron/exon structure. A small intron [101 base pairs (bp), human; 105 bp, rat] interrupts glycine codon 56 while a larger intron (1073 bp, human; 796 bp, rat) is located 8 bp upstream from the translation start site. The introns detect single bands in both rat (3, 5) and human genomes (Fig. 4*B*) and as expected are highly divergent between the two genes. Finally, nearly 80% sequence homology between human and rat genes extends for 250 bp upstream from the transcription initiation site (shown to −80 in Fig. 3*A*). Multiple promoter elements have been defined in the 5′ flanking DNA of the rat gene (16), suggesting that strong selective constraints may account for this extensive structural conservation.

Two Classes of Processed Cytochrome *c* Pseudogenes in the Human Genome. In addition to the functional gene HCS, we have determined the structures of 11 intronless pseudogenes (compared to HCS in Figs. 1, 2, and 3). The initial nine clones, designated HC1, -2, -3, -4, -5, -6, -7, -8, and -10, were isolated from a human genomic library by using a coding region probe at high stringency. All but two have lost portions

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.

Abbreviation: HCS, human somatic cytochrome *c* gene.

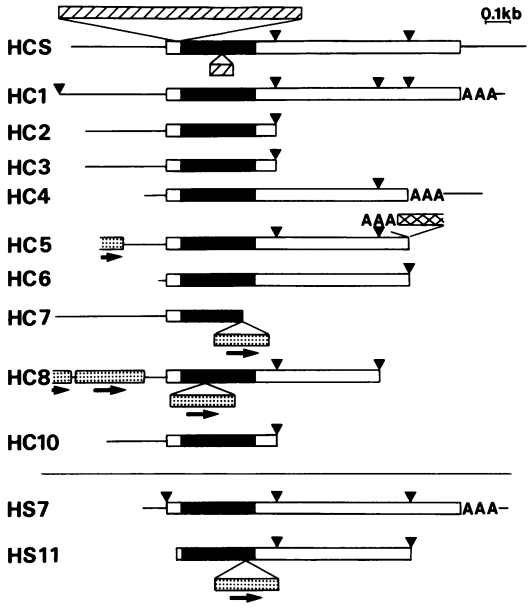

FIG. 1. Genomic structure of the human cytochrome *c* gene HCS and several processed pseudogenes. Diagrams summarize structural information revealed by sequence analysis including cytochrome *c* coding regions (solid boxes), 5′ and 3′ untranslated regions (open boxes), introns (hatched boxes), poly(A) tails (AAA), flanking regions (solid lines), *Alu* elements in the orientation indicated by arrows (stippled boxes), an unidentified pseudogene inserted in HC5 (cross-hatched box), and *Eco*RI sites (arrowheads).

of their 3′ untranslated regions as a result of cloning at internal *Eco*RI sites during library construction, and several have *Alu* element insertions within their coding regions or flanking DNA (Fig. 1). The full-length pseudogenes HC1 and HC4 terminate with poly(A) tails, are flanked by direct repeats, and correspond to the 1.4- and 1.1-kilobase (kb) somatic cytochrome *c* mRNAs detectable in human tissues and HeLa cells (Fig. 4*A*). The pseudogenes HC5 and HC6 were also apparently derived from the 1.4-kb message since they extend beyond the 1.1-kb polyadenylylation site. The 5′ ends of all of the human pseudogenes precisely correspond to the transcription initiation site (Fig. 3*A*), whereas in rat several were derived from upstream initiated transcripts (5).

Aside from dispersed mutational defects all of the previously characterized rat pseudogenes clearly encode the rat somatic protein (5). In contrast, alignment of the nine HC pseudogenes with HCS showed that these pseudogenes did not originate from transcripts encoding the current human cytochrome *c* but encoding a protein nearly identical to the consensus of nonprimate mammalian cytochromes *c* (9). At each of seven amino acid positions (Fig. 2, codons 11, 12, 15, 46, 50, 58, and 83) that distinguish the human protein from other mammalian cytochromes *c*, the nonprimate residue is encoded by a base difference present in all nine HC pseudogenes. For example, at codon 11 all HC pseudogenes encode the valine common to most other mammalian proteins instead of the isoleucine encoded by HCS. Only at codons 44 and 89 do nucleotides present in the HC pseudogenes result in a residue different from either HCS or the mammalian consensus. In addition, the HC pseudogenes share four synonymous differences with HCS (codons 20, 30, 36, and

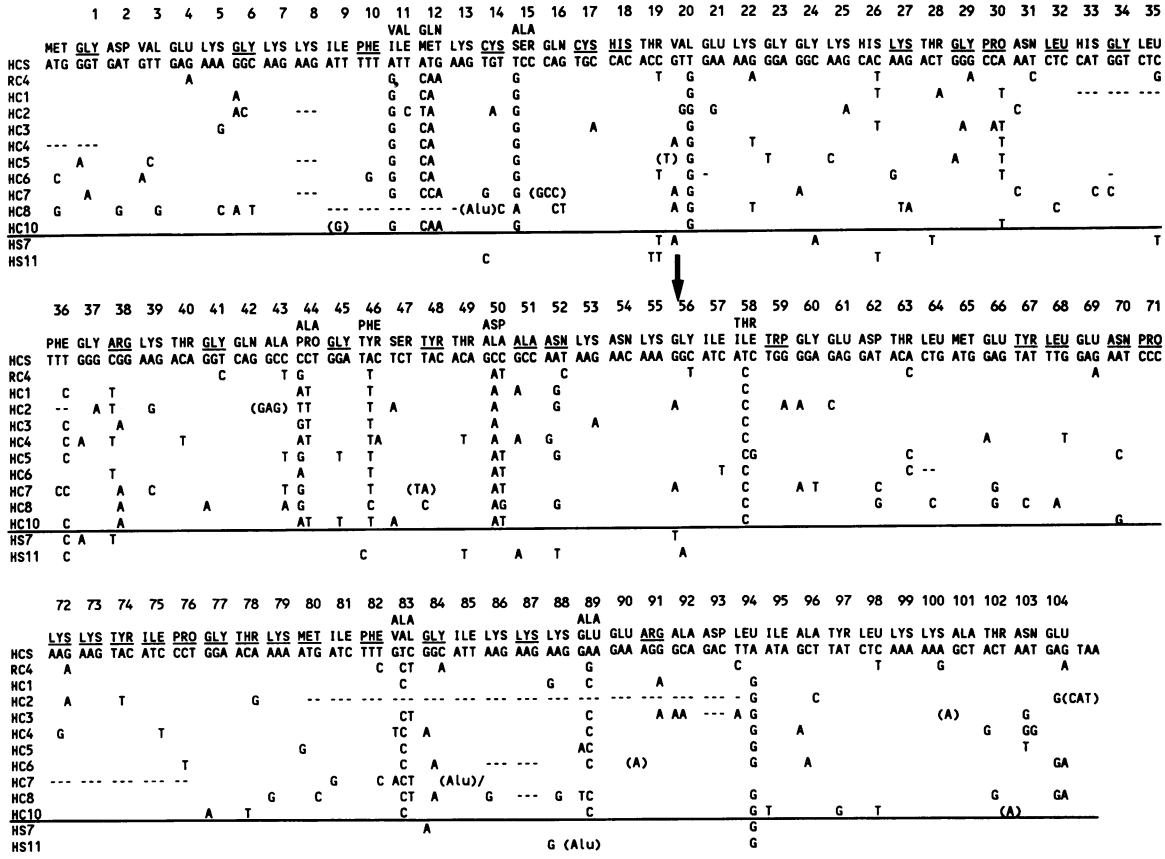

FIG. 2. Nucleotide sequence comparison of coding regions from the human gene HCS, the rat gene RC4, and 11 human pseudogenes (HC1, HC2, HC3, HC4, HC5, HC6, HC7, HC8, and HC10; HS7 and HS11). Only nucleotide differences with HCS are shown for RC4 and the processed pseudogenes. Deletions and insertions of nucleotides relative to HCS are indicated by dashes and enclosure within parentheses (open parentheses located below the nucleotide preceding the insertion), respectively. The numbered amino acid sequence of HCS is presented above the nucleotide sequence, with differences between the HC pseudogene consensus sequence protein and the HCS protein indicated. The position of the coding region intron is indicated by an arrow. Amino acids absolutely invariant throughout all known cytochromes *c* (8) are underlined.

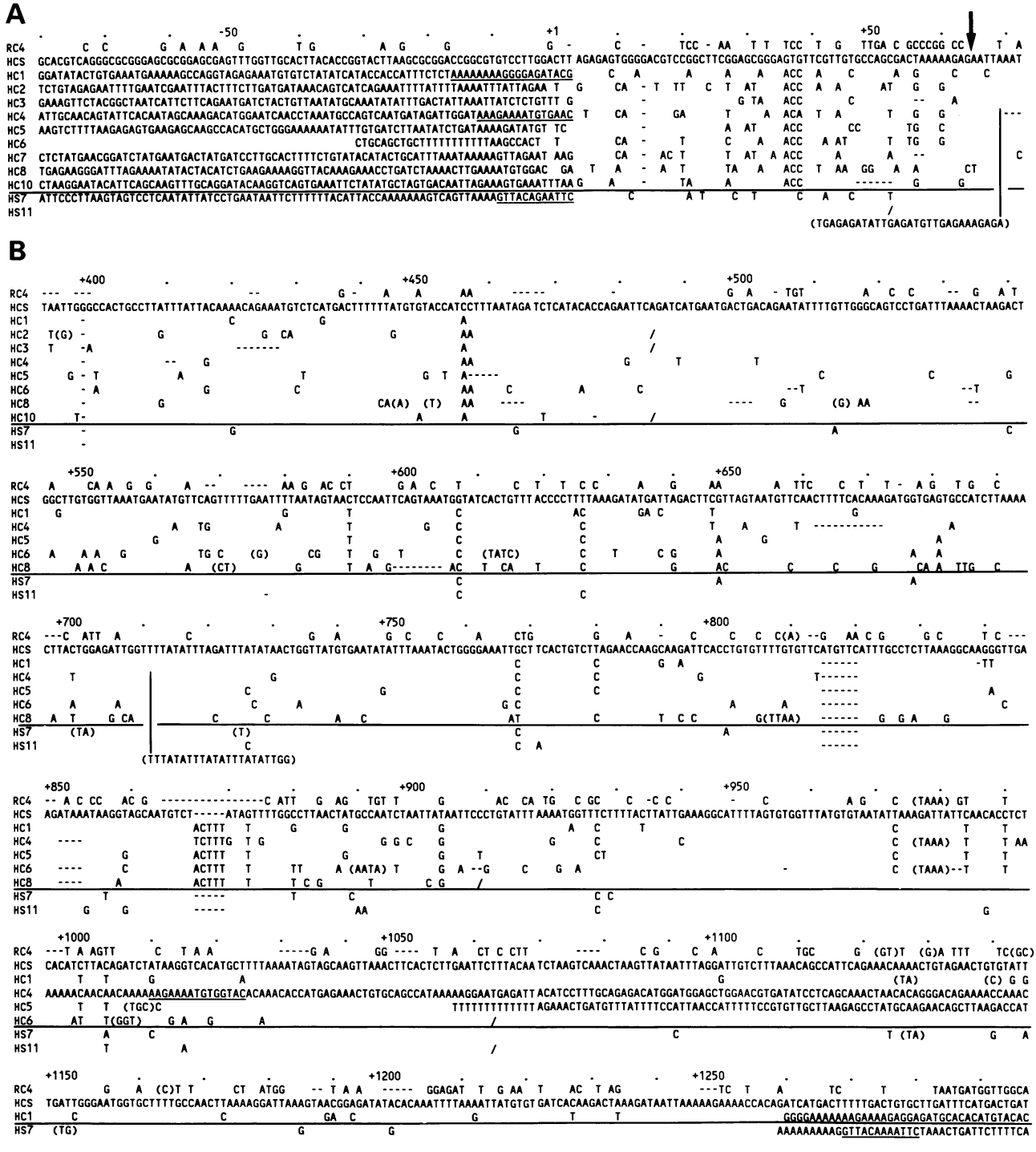

FIG. 3. Nucleotide sequence comparison of 5′ and 3′ noncoding regions of the human gene HCS, the rat gene RC4, and the human pseudogenes. (*A*) Comparison of the 5′ noncoding and flanking regions. (*B*) Comparison of the 3′ noncoding and flanking regions. The numbering system above the sequences is based on S1 nuclease mapping of the rat gene RC4 (5). Deletions and insertions are indicated as in Fig. 2. Slashes indicate the *Eco*RI sites demarcating the end of the Charon 4A phage insert of several pseudogenes except in the 5′ noncoding region of HS11 where the slash indicates the end of the pseudogene sequenced region. Direct repeats flanking the pseudogenes are underlined. The position of the intron in the 5′ noncoding region is denoted by the downward arrow.

94) along with many shared nucleotide differences, insertions, and deletions in their 5′ and 3′ untranslated regions (Fig. 3). Nearly all of these differences are also found in the functional rat somatic gene RC4. Therefore, the nine HC pseudogenes were derived not from HCS but from an ancient progenitor of the cytochrome *c* gene encoding a protein more similar to those found in nonprimate mammals.

To determine whether any of the numerous human cytochrome *c* pseudogenes were derived from a progenitor encoding the current human protein, genomic hybridizations were performed with oligonucleotide probes complementary to codons 10–16 of either HCS (HCS-1, CTGGGAACACT-TCATAATAAA) or the HC pseudogene consensus (HCP-1, CTGGGCACACTTCTGAACAAA). Under conditions of specific hybridization to linearized control plasmids (Fig. 4*D Lower*), HCP-1 detected a complex pattern of genomic fragments similar to that observed with the full coding region (Fig. 4*B*). In contrast, HCS-1 hybridized to only a few genomic restriction fragments including 4.2-kb *Eco*RI and 5.7-kb *Pst* I fragments corresponding to the HCS gene itself

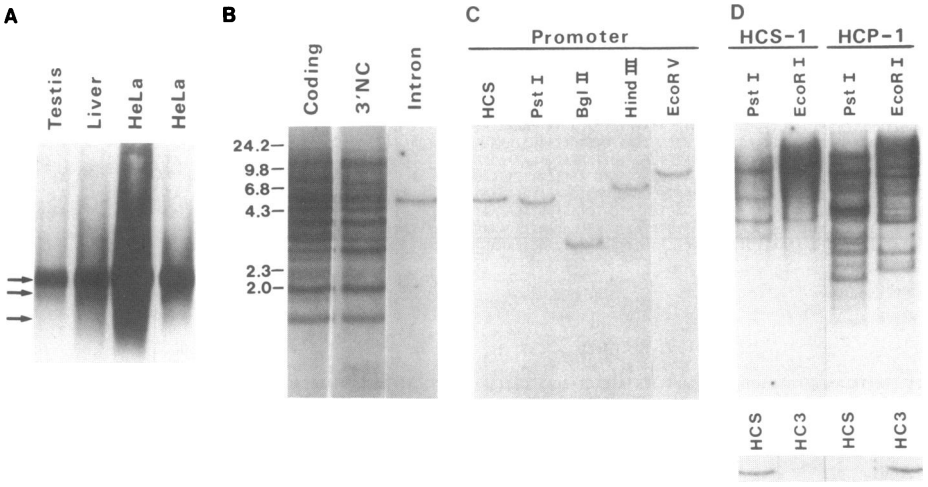

FIG. 4. HCS mRNAs and genomic organization. (*A*) Hybridization analysis of human cytochrome *c* mRNAs. Human testis (15 μg), liver (10 μg), and HeLa (10 μg) polyadenylylated RNAs were hybridized with a HCS 3′ noncoding region (+479 to +1062) probe. Arrows indicate the positions of the 1.4-, 1.1-, and 0.7-kb rat somatic cytochrome *c* mRNAs (4). The filter was exposed for 72 hr except for the 12-hr exposure depicted in the last HeLa lane. (*B*) *Pst* I-digested human genomic DNA (10 μg per lane) was hybridized with a HCS coding region probe (+19 to +484), 3′ noncoding (NC) region probe (+479 to +1062), or a large intron probe as indicated. (*C*) A HCS promoter region probe (−274 to +18) was hybridized to a single-copy gene equivalent (66 pg) of a *Pst* I-digested plasmid containing HCS and 10 μg of human genomic DNA digested with various restriction enzymes (*Pst* I, *Bgl* II, *Hin*dIII, and *Eco*RV). (*D*) Two oligonucleotides, HCS-1 and HCP-1, were hybridized at 50°C or 55°C, respectively, to 10 μg of human genomic DNA digested with either *Pst* I or *Eco*RI. (*Lower*) Single-copy gene equivalents of linearized plasmids containing either the HCS gene or the HC3 pseudogene hybridized with both oligonucleotides.

(Fig. 4*D*). Since any pseudogenes formed from HCS having more than two base changes within the HCS-1 oligonucleotide region would remain undetected under these hybridization conditions, two additional HCS-specific oligonucleotides (codons 42–48 and 45–51) were hybridized to human genomic DNA. No additional bands were detected by using either oligonucleotide, confirming the scarcity of HCS-like sequences in the human genome (data not shown). Using the oligonucleotide HCS-1, two nonallelic processed pseudogenes, HS7 and HS11, were isolated from a human genomic library (Fig. 1). As shown in Fig. 2, none of the nonsynonymous differences found in the HC class of pseudogenes is present in the two HS pseudogenes, although two of the shared synonymous differences with HCS are observed (codons 36 and 94). Similarly, the noncoding regions of the HS pseudogenes very closely resemble HCS except for a few positions where the shared differences in the HC pseudogenes are maintained (Fig. 3; +607, +768, +815–820, and +927). Therefore, the HS pseudogenes originate from a recent progenitor of HCS that correctly encoded the current human cytochrome *c* and are thereby distinguished from the much more abundant HC pseudogenes.

Evolutionary Divergence of Human Cytochrome *c* Pseudogenes. The HS class of pseudogenes can be accounted for by mRNAs transcribed from a recent functional progenitor of HCS. However, we find no evidence for a second functional human gene that can account for the predominant class of HC pseudogenes. Since the human HCS and rat RC4 promoters are highly conserved, it is likely that a second human somatic gene resulting from a duplication in the primate lineage would also have a very similar promoter. Hybridization of the HCS upstream region to human genomic DNA digested with several enzymes revealed only a single band in each lane (Fig. 4*C*) under conditions in which the rat somatic cytochrome *c* gene was detected. In addition, the HCS gene and the HC pseudogenes contain a *Dra* III polymorphism located within the coding region at codons 18–20 (CACACCGTT in HCS, CACACCGTG in the HC pseudogenes), yet the promoter also hybridized to a single band in *Dra* III-digested human genomic DNA (data not shown). Finally, the comparable signal intensities of the human genomic DNA bands and a single-copy equivalent of the *Pst* I-digested plasmid containing the HCS gene (Fig. 4*C*) is consistent with a single gene encoding human somatic cytochrome *c*. Thus, the HC pseudogenes likely originated from a progenitor of HCS more ancient than the HS progenitor. If this is the case, all of the HC pseudogenes should be more divergent from HCS than either of the HS pseudogenes. As shown in Table 1, even when the shared HC differences are excluded, each HC pseudogene is more divergent from HCS than are the HS pseudogenes with no overlap in divergence values between the two distinct pseudogene types.

DISCUSSION

Here the HCS gene family is shown to comprise a single functional gene and numerous intronless pseudogenes. The

Table 1. Divergence analysis of the human pseudogenes

Pseudogene	Divergence, *d*	Estimated age, years × 10^{-6}
HC1	0.032 ± 0.006	25 ± 5
HC2	0.097 ± 0.017	75 ± 13
HC3	0.042 ± 0.010	32 ± 8
HC4	0.053 ± 0.008	41 ± 6
HC5	0.040 ± 0.007	31 ± 5
HC6	0.073 ± 0.009	56 ± 7
HC7	0.110 ± 0.023	85 ± 18
HC8	0.131 ± 0.014	100 ± 11
HC10	0.039 ± 0.010	30 ± 8
HS7	0.022 ± 0.005	17 ± 4
HS11	0.019 ± 0.008	15 ± 6

Mean number of substitutions per nucleotide site, *d* ± SD, calculated by the method of Jukes and Cantor (17). To obtain a conservative divergence estimate for the HC pseudogenes, the base changes present in all of the HC pseudogenes were assumed to represent mutations in the gene and not in each pseudogene, with allowance for either adenine or guanine as the first base of codon 44 and at position +647 in the 3′ noncoding region. Divergence values were derived from the entire coding region plus the 3′ noncoding region to nucleotide position +991 except for the 3′ truncated clones HC2, -3, -7, and -10, where the maximum number of sites to the 3′ end of each clone was utilized. Ages were calculated as described (18), assuming a constant rate of 1.3×10^{-9} nucleotide changes per site per year (19).

human multipseudogene family resembles the rat system in that both have a similar genomic copy number, utilize multiple mRNAs polyadenylylated at alternative sites, and frequently associate with other forms of repetitive element thought to arise through retrotransposition (4, 5, 20). The remarkable difference in the human system is that almost none of the pseudogenes was derived from a gene encoding the current form of human cytochrome *c*. While the consensus of rat pseudogenes clearly encodes rat cytochrome *c* the predominant class of HC pseudogenes encodes a protein closely resembling that found in nonprimate mammalian lineages. Only the HS pseudogenes specify the human cytochrome *c* encoded by HCS. Interestingly, at a few synonymous and 3′ noncoding region positions the HS and HC pseudogenes share nucleotide differences from HCS as if the HS pseudogenes were derived from a recent intermediate between the HC pseudogene progenitor and the current HCS gene.

Both the gene dosage experiment and the absence of a *Dra* III site expected to be present in the parent gene for the HC pseudogenes are consistent with the initial characterization of a single human somatic protein (15). The striking differences between rodent somatic and testicular genes (7) make it unlikely that the human pseudogenes are the products of a human testis-specific cytochrome *c* gene. Although our collection is clearly biased toward younger pseudogenes, there is no overlap between the time of formation of the HC and HS isolates as expected if the two classes originated from a single gene at different times. Therefore, the likely origin of the HC pseudogenes is an ancestral form of HCS.

The structures of the pseudogenes may be considered as a unique molecular record of the history of the somatic cytochrome *c* gene in the primate lineage. Using a recent estimate for the neutral rate of evolution in the higher primates of 1.3×10^{-9} substitutions per site per year (19), the ages span a period of 15–100 million years (Table 1). As predicted by the near identity of human (15), chimpanzee (21), and rhesus monkey (22) protein sequences, the HS pseudogene sequences indicate that cytochrome *c* has not changed in the last 15 million years within the human lineage, although two synonymous substitutions have occurred. In contrast, during the time between the formation of the youngest HC pseudogenes and the HS pseudogenes the cytochrome *c* gene acquired 10 nonsynonymous base changes resulting in 9 amino acid replacements and 2 synonymous changes. Although the time period during which these changes occurred cannot be precisely determined, the estimated ages of the youngest HC pseudogenes and the HS pseudogenes (Table 1) suggest that it did not greatly exceed 15 million years. Using the method of Li *et al.* (23), this time frame yields approximate nonsynonymous and synonymous divergence rates of 2.7×10^{-9} and 2.2×10^{-9} substitutions per nucleotide site per year, respectively, during this period. In contrast, since the formation of the HS pseudogenes the nonsynonymous and synonymous divergence rates of the HCS gene have been 0.0 and 2.2×10^{-9}, respectively. Although only approximate, the relative values suggest that the rate of change of the cytochrome *c* gene at synonymous positions was similar during these two periods, whereas the rate of nonsynonymous substitution deviated from clock-like behavior.

The molecular events governing the rapid transition of the primate cytochrome *c* molecule to its current human form remain unknown. The relaxation of negative selection has been suggested to accelerate the rate of protein evolution of recently duplicated genes, resulting in similar nonsynonymous and synonymous divergence values (23), as observed above. We find no evidence to support the existence of a duplicate cytochrome *c* gene in the human genome. The absence of any overlap in divergence values between the two pseudogene classes argues against the simultaneous existence of two active genes. Nevertheless, we cannot rule out the possibility of a duplication followed by a deletion in the human lineage. In addition, the HCS amino acid changes may not be selectively neutral since biochemical evidence indicates an enhanced maximal reaction rate of primate cytochrome *c* with primate cytochrome *c* oxidase as compared to nonprimate cytochrome *c* oxidase (24). Furthermore, the evolutionary rate of the mitochondrial-encoded cytochrome *c* oxidase subunit II, which directly interacts with cytochrome *c* (25), was also elevated in the primate lineage (26, 27) prior to the Old World monkey/human divergence (28) and has since slowed. Further refinement of the times of acceleration for cytochrome *c* and subunit II evolution will be necessary to determine whether these two events occurred concomitantly or whether one preceded the other.

We thank Gretchen Evans for technical assistance and Drs. W.-H. Li and M. Goodman for computer programs and helpful discussions. We acknowledge the Northwestern University Biotechnology Facilities for synthesis of oligonucleotides and the Cooperative Human Tissue Network for human tissue samples. This work was supported by Grant GM32525 from the National Institutes of Health and Grant 83-34 from the American Cancer Society, Illinois Division.

1. Sherman, F. & Stewart, J. W. (1971) *Annu. Rev. Genet.* **5**, 257–296.
2. Margoliash, E. & Schejter, A. (1966) *Adv. Protein Chem.* **21**, 114–286.
3. Scarpulla, R. C., Agne, K. M. & Wu, R. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 739–743.
4. Scarpulla, R. C. & Wu, R. (1983) *Cell* **32**, 473–482.
5. Scarpulla, R. C. (1984) *Mol. Cell. Biol.* **4**, 2279–2288.
6. Hennig, B. (1975) *Eur. J. Biochem.* **55**, 167–183.
7. Virbasius, J. V. & Scarpulla, R. C. (1988) *J. Biol. Chem.* **263**, 6791–6796.
8. Hampsey, D. M., Das, G. & Sherman, F. (1986) *J. Biol. Chem.* **261**, 3259–3271.
9. Baba, M. L., Darga, L. L., Goodman, M. & Czelusniak, J. (1981) *J. Mol. Evol.* **17**, 197–213.
10. Margoliash, E. (1963) *Proc. Natl. Acad. Sci. USA* **50**, 672–679.
11. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
12. Scarpulla, R. C., Agne, K. M. & Wu, R. (1981) *J. Biol. Chem.* **256**, 6480–6486.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
14. Thein, S. L. & Wallace, R. B. (1986) in *Human Genetic Diseases: A Practical Approach*, ed. Davies, K. E. (IRL, Washington, DC), pp. 33–50.
15. Matsubara, H. & Smith, E. L. (1963) *J. Biol. Chem.* **238**, 2732–2753.
16. Evans, M. J. & Scarpulla, R. C. (1988) *Mol. Cell. Biol.* **8**, 35–41.
17. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–123.
18. Wu, C.-I., Li, W.-H., Shen, J. J., Scarpulla, R. C., Limbach, K. J. & Wu, R. (1986) *J. Mol. Evol.* **23**, 61–75.
19. Koop, B. F., Goodman, M., Xu, P., Chan, K. & Slightom, J. L. (1986) *Nature (London)* **319**, 234–238.
20. Scarpulla, R. C. (1985) *Nucleic Acids Res.* **13**, 763–775.
21. Margoliash, E. & Fitch, W. M. (1968) *Ann. N.Y. Acad. Sci.* **151**, 359–381.
22. Rothfus, J. A. & Smith, E. L. (1965) *J. Biol. Chem.* **240**, 4277–4283.
23. Li, W.-H., Luo, C.-C. & Wu, C.-I. (1985) in *Molecular Evolutionary Genetics*, ed. MacIntyre, R. J. (Plenum, New York), pp. 1–94.
24. Osheroff, N., Speck, S. H., Margoliash, E., Veerman, E. C. I., Wilms, J., Konig, B. W. & Muijsers, A. O. (1983) *J. Biol. Chem.* **258**, 5731–5738.
25. Bisson, R., Azzi, A., Gutweniger, H., Colonna, R., Montecucco, C. & Zanotti, A. (1978) *J. Biol. Chem.* **253**, 1874–1880.
26. Brown, G. G. & Simpson, M. V. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3246–3250.
27. Cann, R. L., Brown, W. M. & Wilson, A. C. (1984) *Genetics* **106**, 479–499.
28. Ramharack, R. & Deeley, R. (1987) *J. Biol. Chem.* **262**, 14014–14021.