Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Gene xx (2003) xxx–xxx
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

The human genome has 49 cytochrome *c* pseudogenes, including a relic of a primordial gene that still functions in mouse

Zhaolei Zhang, Mark Gerstein*

Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114, USA

Received 1 January 2003; received in revised form 2 February 2003; accepted 3 March 2003
Received by W. Makalowski

Abstract

Using a computational approach, we have identified 49 cytochrome c (cyc) pseudogenes in the human genome. Analysis of these provides a detailed description of the molecular evolution of the cyc gene. Almost all of the pseudogenes are full-length, and we have concluded that they mostly originated from independent retrotransposition events (i.e. they are processed). Based on phylogenetic analysis and detailed sequence comparison, we have further divided these pseudogenes into two groups. The first, consisting of four young pseudogenes that were dated to be between 27 and 34 Myr old, originated from a gene almost identical to the modern human cyc gene. The second group of pseudogenes is much older and appears to have descended from ancient genes similar to modern rodent cyc genes. Thus, our results support the observation that accelerated evolution in cyc sequence had occurred in the primate lineage. The oldest pseudogene in the second group, dated to be over 80 Myr old, resembles the testis-specific cyc gene in modern rodents. It is likely that the mammalian ancestor had both the somatic and the testis-specific cyc genes. While the testis-specific gene is still functional in modern rodents, the human has lost it, retaining only a pseudogene in its place. Thus, our study may have identified a pseudogene that is a dead relic of a gene that has completely died off in the human lineage.

Keywords: Cytochrome c; Pseudogenes; Genome; Evolution; Bioinformatics

1. Introduction

Cytochrome *c* (cyc) is a central component of the electron transfer chain in the cell, and is involved in both aerobic and anaerobic respiration. It is also involved in other cellular processes such as apoptosis (Kluck et al., 1997) and heme biosynthesis (Biel and Biel, 1990). It is a ubiquitous protein, found in all eukaryotes and prokaryotes. Because of its importance, relatively small size (104 amino acids in mammals) and ease of isolation, cyc has been very intensively studied. Cyc has also been used as a paradigm in the study of the evolution of protein sequence and structure (Chothia and Lesk, 1985; Wu et al., 1986; Mills, 1991). The amino acid sequences of cyc from many species are now available (Banci et al., 1999); the sequences among vertebrates are especially conserved except among primates, where acceleration in non-synonymous mutation has been observed (Evans and Scarpulla, 1988; Grossman et al., 2001).

By screening genomic DNA libraries, multiple copies of cytochrome *c* processed pseudogenes were discovered in mammalian genomes (Scarpulla et al., 1982; Scarpulla, 1984), including 11 copies in human (Evans and Scarpulla, 1988). Processed pseudogenes are disabled copies of functional genes that do not produce a functional, full-length protein (Vanin, 1985; Mighell et al., 2000; Harrison et al., 2002a). It is believed that they arose from LINE1-mediated retrotransposition, i.e. reverse-transcription of mRNA transcripts followed by integration into genomic DNA, presumably in the germ line (Kazazian and Moran, 1998; Esnault et al., 2000). They are characterized by a complete lack of introns, the presence of small flanking direct repeats and a polyadenine tract near the 3′ end (provided that they have not decayed). Existence of pseudogenes in the genome can obscure the identification

Abbreviations: cyc, cytochrome *c*; *HCS*, human somatic cytochrome *c* gene; HCP, human cytochrome *c* pseudogene; UTR, un-translated region; CDS, protein coding sequence; Myr, million years.

* Corresponding author. Tel.: +1-203-432-6105; fax: +1-360-838-7861.
E-mail address: mark.gerstein@yale.edu (M. Gerstein).

0378-1119/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00579-1

and cloning of functional genes; however, pseudogenes can also provide a fossil record of gene sequences existing at various times during evolution.

Previously, we identified over 2000 ribosomal protein (RP) pseudogenes in the human genome (Harrison et al., 2002b; Zhang et al., 2002), most of which were previously overlooked by DNA hybridization experiments. Motivated by this discovery of an unexpectedly large number of additional pseudogenes, we carried out a similar comprehensive survey on human cytochrome *c* pseudogenes. Our study provides a complete molecular record of the recent evolution of this gene and demonstrates the importance of examining pseudogenic sequences. It also demonstrates a specific instance of a gene disappearing and leaving only a fossil pseudogene in its place.

2. Materials and methods

The basic procedures of our pseudogene discovery pipeline have been previously described (Zhang et al., 2002). A brief overview is given below.

2.1. Six-frame BLAST search for raw fragment homologies

We used the human genome draft freeze of Aug 06, 2001, downloaded from Ensembl website (http://www.ensembl.org). Subsequently, all the chromosomal coordinates were based on these sequences. The amino acid sequences of the cytochrome c proteins were extracted from SWISS-PROT (Bairoch and Apweiler, 2000). Each un-masked human chromosome was split into smaller overlapping chunks of 5.1 MB, and the tblastn program of the BLAST package 2.0 (Altschul et al., 1997) was run on these sequences. The default SEG (Wootton and Federhen, 1993) low-complexity filter parameters were used in the homology search. We then picked the significant homology matches (e-value $<10^{-4}$), and reduced them for mutual overlap by selecting the matches in order of decreasing significance and removing any matches that overlapped substantially with a previously-picked match (i.e. more than ten amino acids or 30 bp).

2.2. Alignment optimization by FASTA dynamic programming

After the BLAST matches were sorted according to their starting positions on the chromosomes, they were examined, and the neighboring matches were merged if they were deemed to be part of the same pseudogene sequence. The merged matches were then extended on both sides to equal the length of the cyc gene plus 30 bp buffers. For each extended match, the human cytochrome *c* (*HCS*) amino acid sequence was re-aligned to the genomic DNA sequence using the program FASTA (Pearson, 1997). FASTA utilizes global dynamic programming that allows gaps between neighboring but not immediately adjacent matches; it also recognizes frame shifts. At this point, we had a total of 50 cyc pseudogene candidates.

2.3. Checking for exon structures

We then examined each candidate pseudogene for the existence of exon structures. One sequence on chromosome 7 was identified as the functional *HCS* gene, as its sequence, including the exons, introns and the flanking regions, matched perfectly with the previously known functional *HCS* gene. Forty-six (46) pseudogene candidates had continuous, intron-less coding regions, which suggested that they were processed pseudogenes; these sequences were labeled as 'intact' processed pseudogenes. The three remaining pseudogene candidates contained retrotransposon sequences inserted in their otherwise continuous coding regions; they were labeled as 'disrupted' processed pseudogenes. We further extended the pseudogene sequences on both sides to obtain the 5′ and 3′ un-translated (UTR) sequences.

2.4. Phylogenetic analysis and dating

Multiple sequence alignment of the pseudogenes and genes was performed using the program ClustalW (Thompson et al., 1994). MEGA2 (Kumar et al., 2001) was used for all the phylogenetic analysis. A phylogenetic tree was constructed by applying the neighbor-joining (NJ) method (Saitou and Nei, 1987; Nei and Kumar, 2000) to the protein coding regions. For each cyc pseudogene, we also calculated the nucleotide sequence divergence from the modern *HCS* gene, using Kimura's two-parameter model (Kimura, 1980), which corrected for multiple hits and also took into account different substitution rates between sites and for transitions vs. transversions. We calculated the ages of some young pseudogenes from the sequence divergence, using formula $T = D/(k)$, where D is the divergence and k is the mutation rate per year per site. A mutation rate of 1.5×10^{-9} for pseudogenes was used (Li, 1997).

3. Results

3.1. The human cyc pseudogene population

A total of 50 cyc homology loci were identified in the human genome, including 49 pseudogenes (denoted as HCP) and one intron-containing functional gene (denoted as *HCS*). The *HCS* gene was located on chromosome 7 (cytogenic band 7p15.3, see Fig. 1), the annotation was confirmed by the perfect alignment of the exons, intron, and the 5′ and 3′ regions with the previously reported nucleotide sequence ((Evans and Scarpulla, 1988), GenBank ID: 181241). It is known that the *HCS* gene contains two introns. The first one is 1,073 bp long and 9 bp upstream of

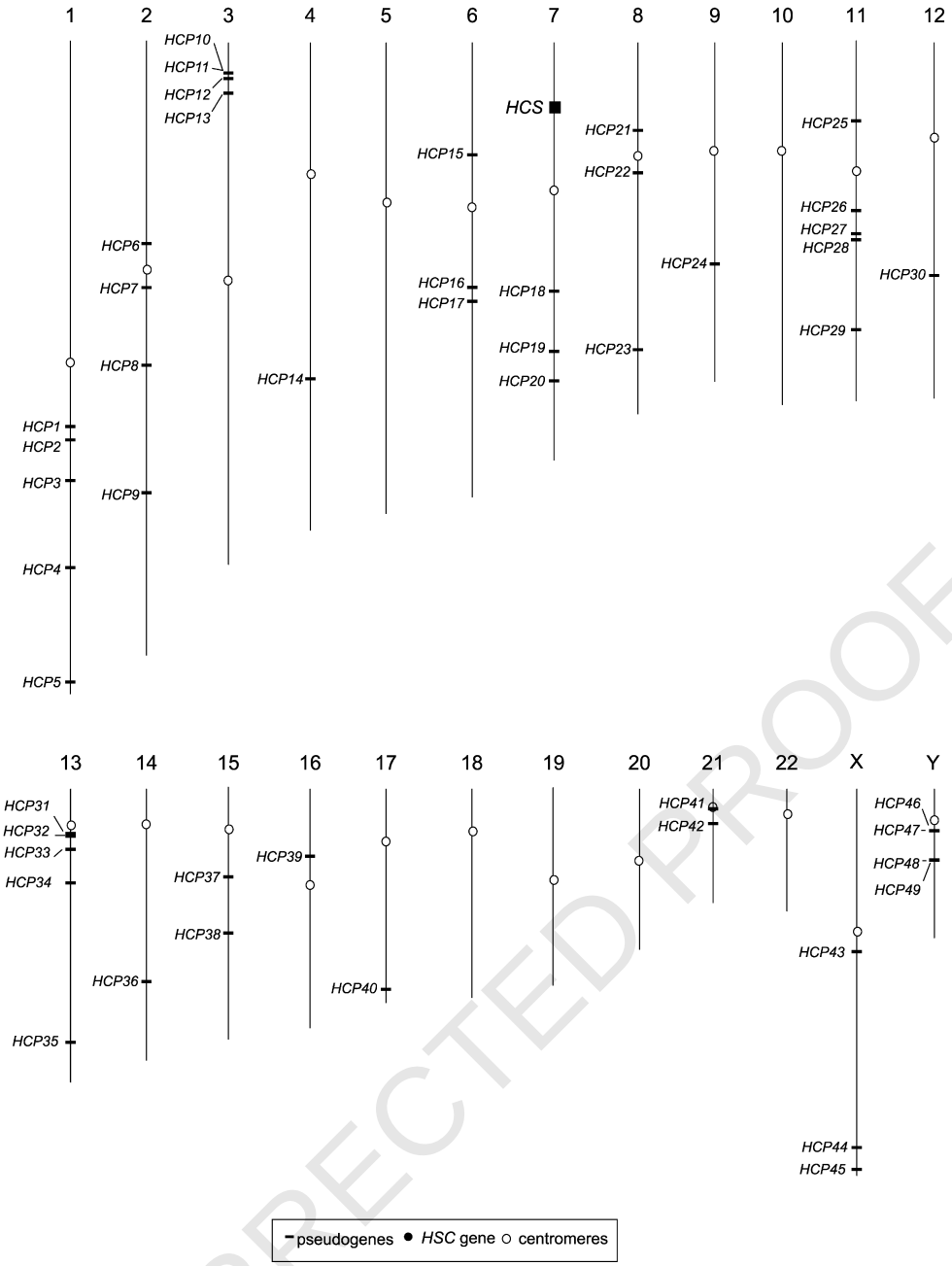


Fig. 1. A map of the cyc gene and pseudogenes in the human genome. The 24 chromosomes are shown as vertical lines. The functional HCS gene is marked as filled black square; pseudogenes marked as horizontal bars and centromeres marked as open circles.

the ATG translation initiation codon; the second intron is 101 bp long and precedes the second nucleotide of the 56th codon. The 49cyc pseudogene assignments were established by their lack of both introns and were, in some cases, further confirmed by the existence of a poly-A tail at the 3′ end. Most of the pseudogenes (40 of 49) were also found to contain obvious disablements in their coding regions. We further searched the GenBank human EST database to confirm that none of these pseudogenes was expressed.

We named our cyc pseudogenes sequentially from *HCP1* to *HCP49* according to their locations on the chromosomes. These pseudogenes are spread out on 18 of the 23 chromosomes, except 5, 10, 18, 19, 20 and 22 (see Fig. 1). Fig. 2 shows the alignment of the predicted amino acid sequences of these pseudogenes with the *HCS* protein. The disablements are highlighted in gray. More detailed information on these pseudogenes is provided in Table 1. Except for *HCP9*, *HCP15* and *HCP30*, which are disrupted into two or three fragments by insertions of retrotransposons, most of the pseudogenes have continuous sequences.

The sequences of 40 of the 49 pseudogenes can be

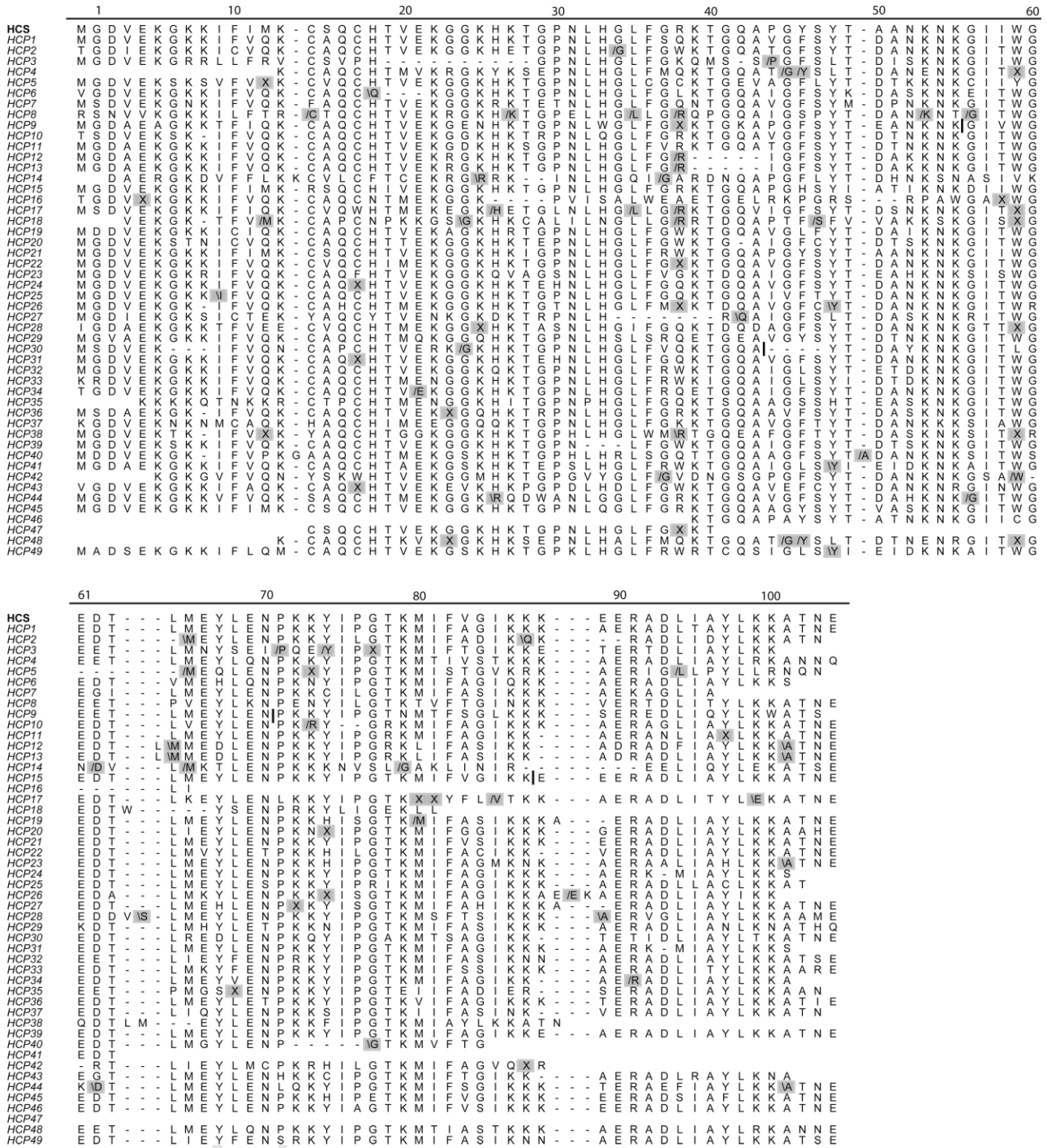


Fig. 2. Alignment of the translated amino acid sequences of the human cyc pseudogenes, together with the functional *HCS* protein sequence. In the pseudogene sequences, missing amino acids caused by truncation are left as blank, dashes '−' indicate a gap caused by DNA deletion, frame shifts and stop codons are indicated by '/', '\' and 'X'. Repeat insertions are marked as vertical bars. Apparent disablements in the pseudogenes (frame shifts and premature stop codons) are highlighted. The numbering system above the sequences is based on the *HCS* sequence.

aligned to match the entire length of the protein coding sequence (CDS) of human cyc mRNA and most of the UTR (un-translated) regions as well. The remaining nine pseudogenes: *HCP4*, *HCP16*, *HCP18*, *HCP40*, *HCP41*, *HCP42*, *HCP46*, *HCP47* and *HCP48* are truncated to various degrees at 5′ or 3′ ends. *HCP47*, the shortest one, only matches residues 14 to 40 of the HCS gene's sequence. Interestingly, this gene fragment is immediately adjacent to *HCP46* on chromosome Y, which matches HCS residues 39–104 but on the opposite chromosomal strand. It appears that *HCP46* and *HCP47* were once parts of an original complete cyc pseudogene that had undergone '5′ inversion',

Table 1
Detail information on the 49 human cytochrome *c* pseudogenes[a]

Pseudogene ID	NCBI accession number	Band (strand)	Chromosome location[b]	Amino acid identity[c] (%)	nt Divergence[d]	Comments
HCS		7p15.3 (−)	25.65M			Functional human cyc gene
HCP1	AF533162	1q21.3 (+)	151.49M	88	0.090 ± 0.020	
HCP2 (HC6)	AF533163	1q23.1 (+)	156.78M	82	0.114 ± 0.023	
HCP3	AF533164	1q24.3 (−)	172.70M	59	0.235 ± 0.042	
HCP4	AF533165	1q32.1 (−)	206.92M	73	0.138 ± 0.027	Truncated before residue 13
HCP5	AF533166	1q44 (−)	251.92M	64	0.345 ± 0.054	
HCP6	AF533167	2p12 (−)	79.59M	77	0.113 ± 0.022	
HCP7	AF533168	2q11.2 (+)	96.95M	74	0.182 ± 0.033	
HCP8	AF533169	2q14.3 (−)	127.35M	69	0.220 ± 0.037	
HCP9	AF533170	2q31.2 (−)	177.54M	79	0.473 ± 0.071	Disrupted into three fragments by Alus
HCP10	AF533171	3p25.3 (−)	11.81M	81	0.107 ± 0.022	*HCP10-13* are duplicated copies
HCP11	AF533172	3p25.3 (−)	11.87M	83	0.095 ± 0.020	See *HCP10*
HCP12	AF533173	3p25.3 (−)	14.05M	76	0.091 ± 0.020	See *HCP10*
HCP13	AF533174	3p25.1 (−)	19.78M	77	0.091 ± 0.020	See *HCP10*
HCP14	AF533175	4q28.3 (−)	131.99M	46	1.257 ± 0.253	
**HCP15 (*HS11)	AF533176	6p21.1 (+)	44.06M	92	0.041 ± 0.013	Disrupted into two fragments by Alus
HCP16[e]	AF533177	6q15 (−)	96.18M	48	N/A	Truncated after residue 58
HCP17	AF533178	6q16.1 (+)	101.62M	74	0.147 ± 0.028	
HCP18	AF533179	7q21.3 (−)	97.54M	52	0.610 ± 0.117	Truncated after residue 81
HCP19	AF533180	7q31.32 (+)	121.29M	83	0.117 ± 0.023	
HCP20	AF533181	7q32.1 (−)	132.78M	76	0.142 ± 0.026	
**HCP21 (HS7)*	AF533182	8p12 (−)	34.34M	95	0.034 ± 0.011	
HCP22	AF533183	8q11.22 (−)	51.04M	83	0.108 ± 0.022	
HCP23	AF533184	8q24.12 (−)	120.57M	76	0.164 ± 0.029	
HCP24 (HC3)	AF533185	9q22.32 (+)	86.92M	83	0.103 ± 0.021	*HCP24* and *31* are duplicated copies
HCP25 (HC10)	AF533186	11p13 (−)	31.51M	83	0.094 ± 0.020	
HCP26	AF533187	11q13.2 (−)	66.72M	75	0.148 ± 0.027	
HCP27	AF533188	11q13.4 (+)	75.78M	68	0.173 ± 0.031	
HCP28	AF533189	11q14.1 (+)	78.18M	71	0.173 ± 0.031	
HCP29	AF533190	11q22.3 (−)	113.54M	72	0.140 ± 0.026	
HCP30	AF533191	12q21.32 (+)	91.41M	72	0.177 ± 0.032	Disrupted into two fragments
HCP31 (HC3)	AF533192	13q12.11 (−)	17.40M	83	0.103 ± 0.021	See HCP24
HCP32	AF533193	13q12.11 (−)	18.53M	79	0.123 ± 0.024	*HCP32*, *41* and *49* are duplicated copies
HCP33 (HC4)	AF533194	13q12.12 (−)	23.63M	77	0.129 ± 0.025	
HCP34	AF533195	13q14.11 (+)	36.85M	86	0.088 ± 0.019	
HCP35 (HC8)	AF533196	13q32.3 (−)	99.53M	64	0.256 ± 0.042	
HCP36 (HC5)	AF533197	14q24.3 (+)	75.65M	82	0.108 ± 0.022	
HCP37	AF533198	15q15.1 (+)	35.24M	70	0.226 ± 0.039	
HCP38 (HC2)	AF533199	15q22.2 (+)	57.39M	61	0.287 ± 0.049	
HCP39 (HC1)	AF533200	16p12.1 (−)	26.36M	84	0.093 ± 0.020	
HCP40 (HC7)	AF533201	17q25.3 (+)	78.64M	68	0.174 ± 0.036	Truncated after residue 84
HCP41	AF533202	21q11.2 (−)	7.75M	71	0.198 ± 0.044	Truncated after residue 63, see also *HCP32*
HCP42	AF533203	21q21.1 (−)	13.53M	58	0.458 ± 0.084	Truncated after residue 88
HCP43	AF533204	Xq13.1 (−)	63.83M	75	0.170 ± 0.031	
HCP44	AF533205	Xq27.3 (+)	140.82M	77	0.124 ± 0.024	
**HCP45*	AF533206	Xq28 (+)	149.47M	91	0.052 ± 0.014	
**HCP46*[f]	AF533207	Yq11.221 (−)	16.31M	92	0.047 ± 0.014	Truncated before residue 39
HCP47[f]	AF533208	Yq11.221 (+)	16.31M	96	–	Residues 14–40 only.
HCP48	AF533209	Yq12 (+)	27.78M	71	0.165 ± 0.031	Truncated before residue 13
HCP49	AF533210	Yq12 (+)	27.93M	71	0.176 ± 0.031	See *HCP32*

[a] The class 1 pseudogenes are indicated by *; the rest of the pseudogenes are class 2.
[b] Chromosomal coordinates of the pseudogene in Mb (million base pair).
[c] Amino acid sequence identity between translated cyc pseudogene and *HCS* sequence.
[d] Nucleotide sequence divergence and its standard error, $d \pm$ SE, between pseudogenes and modern *HCS* gene. The standard errors were computed analytically.
[e] *HCP16* is too incomplete to compute the sequence divergence.
[f] *HCP46* and *HCP47* were merged together in calculating the divergence.

which is common for LINE1-mediated retrotransposition (Ostertag and Kazazian, 2001). In some of the following discussions, the *HCP47* sequence was merged into *HCP46* to form a complete cyc pseudogene sequence.

Fig. 3 shows the sequence alignment of the 5′ UTR regions of 45 human cyc pseudogenes and the *HCS* mRNA; the 5′ flanking regions were also included for the pseudogenes. The four pseudogenes that are truncated near the 5′ end are not included in the alignment. The two downward arrows mark the start of the *HCS* mRNA sequence and the ATG translation initiation codon. As can be seen, most of these pseudogenes have retained the nearly intact 5′ UTR sequence. This high degree of sequence preservation is a little surprising, as it has been known that LINE1-mediated reverse-transcription has a low efficiency and often leads to 5′ truncation and thus incomplete insertion of mRNA transcripts into the genome.

As outlined in bold in Fig. 3, three groups of the pseudogenes share almost identical 5′ flanking sequences. This indicated that the pseudogene sequences within each group arose from genomic duplications of an original pseudogene, rather than from independent reverse-transcription events, and that the sequences had, therefore, retained the flanking sequence of the original pseudogene. The pseudogenes in the first group (*HSP10*, *HSP11*, *HSP12* and *HSP13*) were located very close to each other on chromosome 3 (see Table 1 and Fig. 1). This suggested an intra-chromosome sequential duplication event. The two other groups (the first consisting of *HCP31* and *HCP24* and the other consisting of *HCP32*, *HCP41* and *HCP49*) appeared to have resulted from inter-chromosomal duplications. Such extensive segmental duplications in the human genome have been described recently (Bailey et al., 2002).

By screening human cDNA and genomic libraries, Evans and Scarpulla (Evans and Scarpulla, 1988) previously reported 11 human cyc pseudogenes, which were named *HC1-HC6*, *HS7*, *HC8*, *HC9*, *HC10*, and *HS11*. We were able to unambiguously assign ten of these eleven sequences to a single pseudogene in our pseudogene set as indicated in the leftmost column on Table 1. The remaining one, *HC3*, has identical to a pair of duplicated pseudogenes: *HCP24* and *HCP31*. Therefore, in addition to the previously reported 11, we discovered 37 new cyc pseudogenes in the human genome.

3.2. Phylogenetic analysis

We were interested in tracing the origin of these cyc pseudogenes and placing them into the context of evolution. Fig. 4 shows the phylogenetic tree constructed by applying the neighbor-joining (NJ) method (Saitou and Nei, 1987; Nei and Kumar, 2000) to the protein-coding regions of human cyc pseudogenes and the functional cyc genes from human, rat, mouse, chicken and fruitfly. Rodents have two cyc genes in their genomes: the somatically expressed genes (*CYCS_RAT*, *CYCS_MOUSE*) and the testis-specific genes (*CYCT_RAT*, *CYCT_MOUSE*). These testis-specific cyc genes are only expressed during spermatogenesis (Virbasius and Scarpulla, 1988). Compared with their somatic counterparts, they have different exon structures and differ at 14–15 amino acid positions. Fruitfly also has two cyc genes, *FLY_DC4* and *FLY_DC3*, which differ at 32 amino acid positions (Limbach and Wu, 1985); it was believed that they diverged about 520 Myr (million years) ago (Wu et al., 1986). *FLY_DC4* has a much higher expression level in the cell than *FLY_DC3*, and was used to root the phylogenetic tree.

As expected, the two fruitfly genes were clearly separated from the vertebrate sequences. Also, the chicken gene and the rodent testis-specific genes were placed close to each other and distant from the mammalian somatic genes and the majority of the human pseudogenes (except *HCP9*). It was postulated that these tissue-specific cyc genes arose from duplication of an ancestral cyc gene (Limbach and Wu, 1985) and the estimated divergence time of these genes from somatic genes was close to the divergence time of birds and mammals (Mills, 1991).

Table 1 lists the nucleotide sequence divergences between each cyc pseudogene and the modern *HCS* gene calculated according to Kimura's two-parameter model (Kimura, 1980). Sequence divergence, or the number of nucleotide substitutions between sequences, is a measure of evolutionary distance between two sequences. In this case, the divergence values were correlated with the ages of the pseudogenes, i.e. the approximate time when each pseudogene was inserted into the genome. It might be expected that, on average, the older pseudogenes should have greater divergence than the younger ones. However, special care has to be taken in comparing divergence of pseudogenes, as they contain not only the accumulated mutations in the pseudogene sequences after they were inserted into the genome, but also the sequence differences in the functional genes from which they originated. It is rather tempting to estimate the age of a pseudogene by simply dividing the divergence by a constant nucleotide substitution rate. However, we believe such a simplified calculation should not be applied here for the cyc pseudogenes, as it assumes that the pseudogenes all originated from the same ancestral cyc gene and same mRNA transcript. As will be discussed in Section 3.3, this is certainly not true for the cyc pseudogenes.

3.3. Two classes of cyc pseudogenes

Based on a comparison of the pseudogene sequences with the modern *HCS* gene and consensus mammalian cyc sequence, Evans and Scarpulla (Evans and Scarpulla, 1988) divided their 11 human cyc pseudogenes into two classes. The predominant class of older pseudogenes (denoted as 'class 2', nine members) appeared to have originated from an ancient progenitor of the cyc gene, and the remaining two pseudogenes (class 1, *HS7* and *HS11*) were younger and

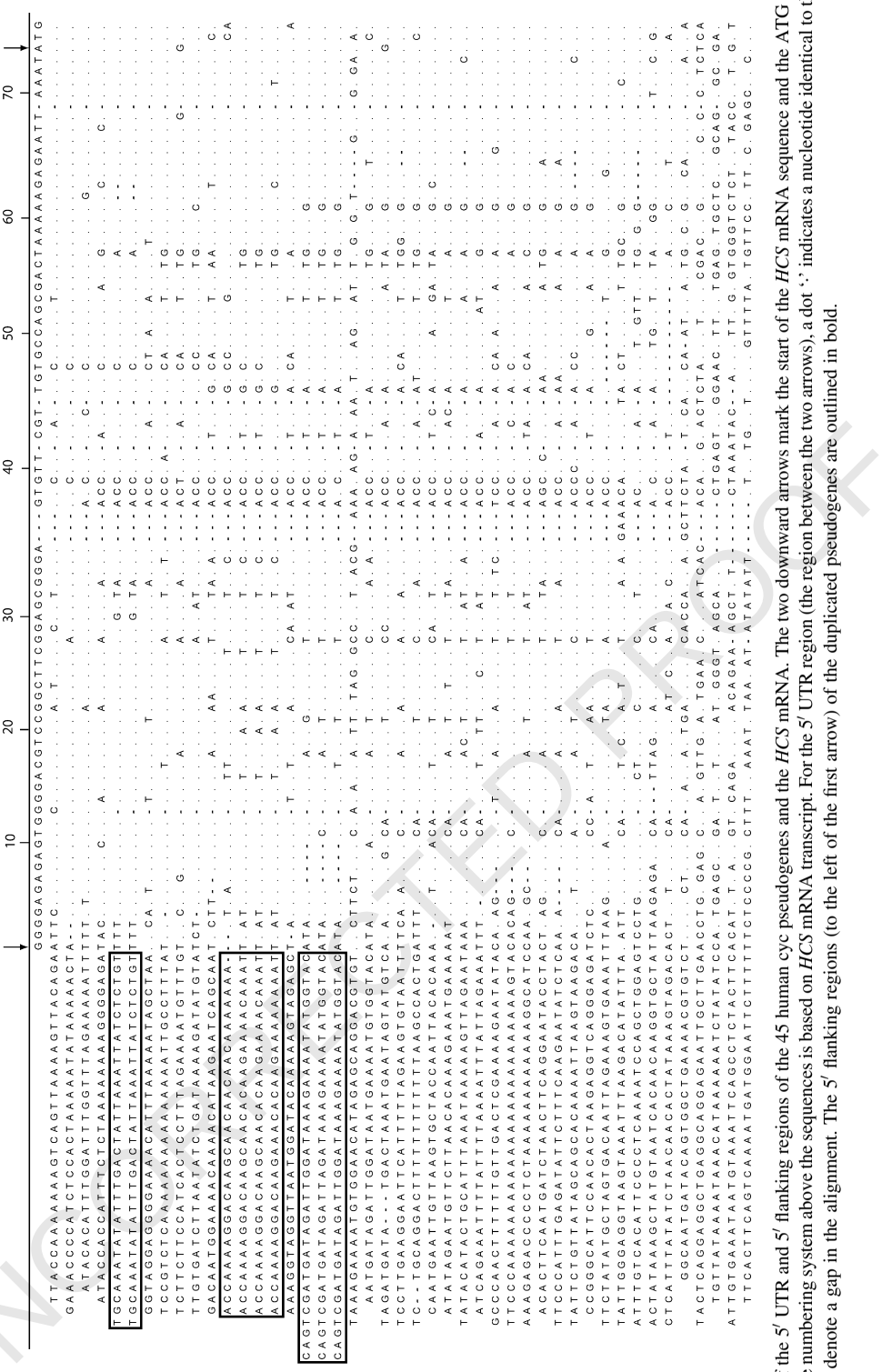
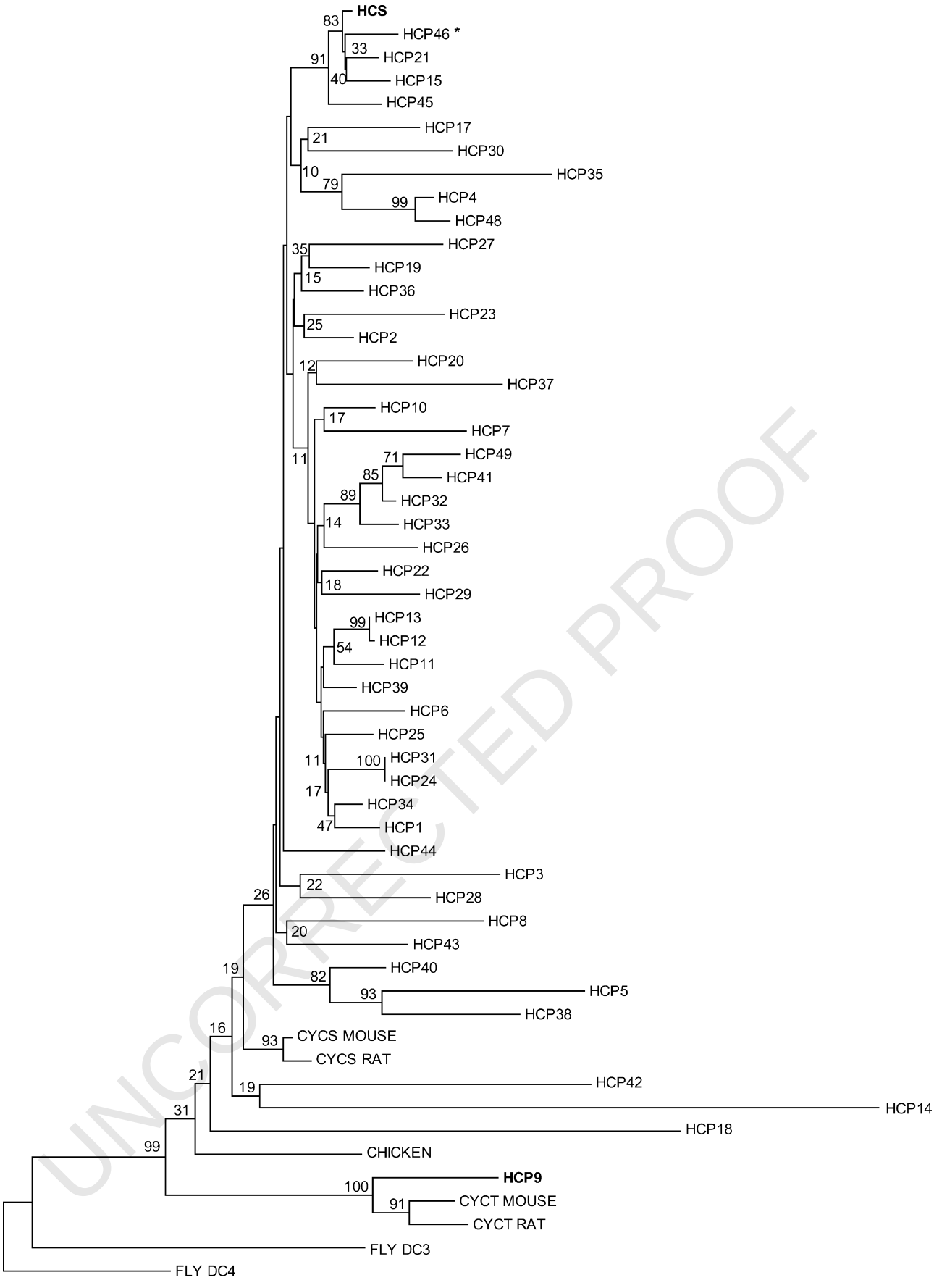


Fig. 3. Alignment of the 5′ UTR and 5′ flanking regions of the 45 human cyc pseudogenes and the *HCS* mRNA. The two downward arrows mark the start of the *HCS* mRNA sequence and the ATG translation initiation codon. The numbering system above the sequences is based on *HCS* mRNA transcript. For the 5′ UTR region (the region between the two arrows), a dot '·' indicates a nucleotide identical to that in *HCS* mRNA; dashes '–' denote a gap in the alignment. The 5′ flanking regions (to the left of the first arrow) of the duplicated pseudogenes are outlined in bold.



originated from a gene similar to the present HCS gene. This classification clearly accords with the phylogenetic tree shown in Fig. 4, as the topology of the tree suggests that the majority of human cyc pseudogenes were not direct descendents of present-day *HCS* gene.

As we have obtained a comprehensive set of 49 cyc pseudogenes (48 if *HCP46* and *HCP47* are considered as one), which is considerably larger than the set of 11 pseudogenes previously analyzed. We wanted to test whether the previous classification still held true. This goal was best achieved by comparing sequences at the informative codon positions where mutations had occurred during recent evolution. Cyc is a highly conserved protein among eukaryotes; the pair-wise amino acid sequence identities range from 45% between mammals and yeast to 93.3% between chicken and mouse. An accelerated rate of amino acid changes has occurred on the primate lineage leading to the human ancestor, as there were amino acid changes at nine positions since the split between *Rattus* and *Homo* (Grossman et al., 2001). Among these positions (11, 12, 15, 44, 46, 50, 58, 83, 89), none belongs to the 'conservatively substituted' category (Banci et al., 1999), which suggests that they are probably not directly involved in the electron-transfer process. Fig. 5 compares the human pseudogenes and the functional genes from human, rodents and chicken at these codon positions. For each position, the sequences that have the same amino acid type as *HCS* are shown in pink. Also, for eight of the nine positions, a dominant amino acid type exists among the pseudogenes; the positions that share this dominant amino acid type are highlighted in light green. For position 44, both Val and Ile are dominant amino acid types, so both are highlighted.

It is obvious from the alignment that at all nine codon positions, the majority of the human pseudogenes share an amino acid type that is different with the HCS gene. Four pseudogenes, *HCP15*, *HCP21*, *HCP45* and *HCP46*, have the highest sequence identity with *HCS* at these positions, and they were selected and labeled as class 1 and the rest of the pseudogenes were grouped into class 2. Note that we used the same nomenclature as used by previous investigators (Evans and Scarpulla, 1988; Grossman et al., 2001).

A more detailed look at these codon positions follows. At position 11, 31 of the 48 human pseudogenes have residue Val and codon GTT or GTC; in contrast, the HCS gene and three class 1 pseudogenes (*HCP15*, *HCP21* and *HCP45*) have residue Ile and codon ATT. Interestingly, as in most of the class 2 pseudogenes, the somatic rodent and chicken genes also have Val and GTT/GTC at position 11. The same pattern also occurs at positions 12, 15, 46, 50, 58 and 83, where the majority of the class 2 pseudogenes share the same amino acid type with rodent somatic genes, and the class 1 pseudogenes share a different amino acid type with the *HCS* gene. At position 44, there is no predominant amino acid type among the pseudogenes, as Ile occurs 14 times and Val occurs 12 times; however, the HCS gene and three class 1 pseudogenes (HCP15, HCP2 and HCP46) have Pro at the position. This particular position has obviously gone through very rapid changes in recent evolution, since rodent somatic cyc genes have residue Ala at the position, which occurred only six times among the pseudogenes. At position 89, the predominant amino acid among the pseudogenes is Ala (occurring 24 times), which is different from both human and rodent somatic genes: HCS and all class 1 pseudogenes have Glu and the rodent somatic genes have Gly.

The sequence comparison shown in Fig. 5 strongly supports the notion that the human cyc pseudogenes originated from a functional gene that had undergone significant changes during the mammalian evolution. The four pseudogenes in class 1 appear to be from a gene that is identical to the modern *HCS* gene, while the class 2 pseudogenes are much older and have more resemblance to rodent somatic genes. Although we divided the pseudogenes into two classes, it is important to note that gene evolution was a gradual process, and our classification in no way implies any dramatic changes in the biochemical function and gene structure. Our classification is in very good agreement with the phylogenetic analysis, as the four class 1 pseudogenes were found in a separate branch together with the *HCS* gene at the top of the tree (Fig. 4). Furthermore, as shown in Fig. 3, the 5′ UTR sequences of *HCP15*, *HCP21* and *HCP45* also have the fewest number of substitutions compared with *HCS* mRNA sequences. Given the conclusion that the four class 1 pseudogenes and the modern *HCS* gene share the same origin, it is possible to actually date these pseudogenes based on their sequence divergence. Using the formula $T = D/(k)$, where D is the divergence and k is the mutation rate per year per site, the ages for the pseudogenes were estimated to be 27 ± 8 Myr for *HCP15*, 23 ± 7 Myr for *HCP21*, 34 ± 9 Myr for *HCP45* and 31 ± 9 Myr for *HCP46*. A mutation rate of 1.5×10^{-9} per site per year for pseudogenes was used (Li, 1997). In comparison, the divergence time of human from gibbons is believed to be 20 to 30 Myr ago (Lander et al., 2001). The much lower number of pseudogenes in class 1 compared with the number in class 2 is consistent with the observed decline of retrotransposition activity during the last 40 Myr in the human genome (Lander et al., 2001).

Fig. 4. Phylogenetic tree of the human cyc pseudogenes. The tree is constructed using the software MEGA2 (Kumar et al., 2001) on the protein-coding regions, and it is rooted by the fruitfly *FLY_DC4* gene sequence. (**HCP47* is merged into *HCP46*). Percentage bootstrap values (based on 1000 replications) supporting each node are also indicated.

		11	12	15	44	46	50	58	83	89
	HCS	I ATT	M ATG	S TCC	P CCT	Y TAC	A GCC	I ATC	V GTC	E GAA
(HS11)	HCP15	I ATT	M ATG	S TCC	P CCT	H CAC	A GCC	I ATC	V GTC	E GAA
(HS7)	HCP21	I ATT	M ATG	S TCC	P CCT	Y TAC	A GCC	I ATC	V GTC	E GAA
Class I	HCP45	I ATT	M ATG	S TCC	A GCT	Y TAC	V GTT	T ACC	V GTC	E GAA
	HCP46	-	-	S TCC	P CCT	Y TAT	A GCC	I ATT	V GTC	E GAA
Class 2	HCP1	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAT	I ATC	A GCC	A GCA
(HC6)	HCP2	V GTT	Q CAG	A GCC	T ACT	F TTC	D GAT	T ACC	A GCC	-
	HCP3	F TTC	R AGA	S TCA	/ CT	F TTC	D GAC	I ATC	T ACA	T ACA
	HCP4	-	-	A GCC	T ACT	/ AT	D GAC	T ACC	V GTC	A GCA
	HCP5	V GTT	X TAG	V GTC	A GCT	F TTC	D GAC	I ATC	T ACT	A GCA
	HCP6	V GTT	Q CAG	A GCC	I ATT	F TTC	D GAT	T ACC	A GCT	A GCA
	HCP7	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAC	T ACT	A GCC	A GCA
	HCP8	F TTC	T ACA	T ACC	I ATT	S TCC	D GAC	T ACC	T ACT	V GTA
	HCP9	I ATT	Q CAG	A GCT	P CCA	F TTT	E GAG	V GTA	S TCT	S AGT
	HCP10	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAC	T ACC	A GCC	A GCA
	HCP11	V GTT	Q CAG	A GCC	T ACT	F TTC	D GAC	T ACC	A GCC	A GCA
	HCP12	V GTT	Q CAG	A GCC	I ATT	F TTC	D GAC	T ACC	A GCC	A GCA
	HCP13	V GTT	Q CAG	A GCC	I ATT	F TTC	D GAC	T ACC	A GCC	A GCA
	HCP14	F TTT	L CTT	V GTC	P CCT	F TTT	D GAT	I ATT	I ATT	-
	HCP16	V GTT	Q CAG	A GCC	R CGT	P CCC	-	X TAG		
	HCP17	I ATT	Q CAG	V GTC	I ATT	F TTC	D GAT	T ACC	F TTT	A GCA
	HCP18	V GTC	/ AG	A GCC	P CCC	F TTC	V GTT	S TCT		
	HCP19	I ATT	Q CAG	A GCC	V GTT	L TTA	D GAT	T ACC	A GCC	-
	HCP20	V GTT	Q CAG	A GCC	I ATT	F TTC	D GAC	T ACC	G GGC	G GGA
	HCP22	V GTT	Q CAG	V GTC	V GTT	F TTC	D GAT	T ACC	A GCC	V GTA
	HCP23	V GTT	Q CAG	A GCC	I ATT	F TTC	E GAG	S TCC	A GCA	A GCA
(HC3)	HCP24	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAC	T ACC	A GCT	A GCA
(HC10)	HCP25	V GTT	Q CAA	A GCC	I ATT	F TTC	D GAT	T ACC	A GCC	A GCA
	HCP26	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAC	T ACC	A GCT	A GCA
	HCP27	T ACT	E GAG	A GCC	I ATT	F TTC	D GAT	T ACC	A GCC	-
	HCP28	V GTT	E GAG	V GTC	A GCT	F TTC	D GAT	T ACC	T ACC	A GCA
	HCP29	V GTT	Q CAG	A GCC	V GTT	Y TAC	D GAC	T ACC	A GCC	A GCA
	HCP30	V GTT	Q CAG	A GCA	-	-	D GAT	T ACC	A GCT	T ACA
(HC3)	HCP31	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAC	T ACC	A GCT	A GCA
	HCP32	V GTT	Q CAG	A GCC	I ATT	L TTA	E GAG	T ACC	A GCC	A GCA
(HC4)	HCP33	V GTT	Q CAG	A GCC	I ATT	L TTA	D GAC	T ACC	S TCC	A GCA
	HCP34	V GTT	Q CAG	A GCC	I ATT	F TTC	D GAT	T ACC	A GCC	A GCA
(HC8)	HCP35	K AAA	K AAA	T ACC	A GCT	S TCC	E GAG	T ACC	A GCT	S TCA
(HC5)	HCP36	V GTT	Q CAG	A GCC	A GCT	F TTC	D GAT	T ACG	A GCC	T ACA
	HCP37	A GCT	Q CAG	A GCC	V GTT	F TTC	D GAC	A GCC	A GCC	V GTA
(HC2)	HCP38	V GTC	X TAG	A GCC	F TTT	F TTC	D GAC	T ACC	Y TAT	-
(HC1)	HCP39	V GTT	Q CAG	A GCC	I ATT	F TTC	D GAC	T ACC	A GCC	A GCA
(HC7)	HCP40	V GTT	P CCA	A GCC	A GCT	F TTC	D GAT	T ACC	T ACT	
	HCP41	V GTT	Q CAG	A GCC	I ATT	L TTA	E GAG	T ACC		
	HCP42	V GTT	Q CAG	S AGC	P CCT	F TTC	D GAT	A GCC	A GCT	
	HCP43	A GCT	Q CAA	A GCC	V GTT	F TTC	D GAT	N AAC	T ACT	A GCA
	HCP44	V GTT	Q CAG	A GCC	V GTT	F TTC	D GAT	T ACC	S TCC	T ACA
	HCP48	-	-	A GCC	T ACT	/ AT	D GAC	T ACC	A GCC	A GCA
	HCP49	L CTT	Q CAG	A GCC	I ATT	L TTA	E GAG	T ACC	A GCC	A GCA
	CYCS MOUSE	V GTT	Q CAG	A GCC	A GCT	F TTC	D GAT	T ACC	A GCT	G GGA
	CYCS RAT	V GTT	Q CAA	A GCC	A GCT	F TTC	D GAT	T ACC	A GCT	G GGA
	CYCT MOUSE	V GTT	Q CAG	A GCT	P CCA	F TTT	D GAT	I ATC	A GCT	S AGT
	CYCT RAT	I ATT	Q CAG	A GCG	P CCA	F TTT	D GAT	I ATC	A GCT	S AGT
	CHICKEN	V GTC	Q CAG	S TCC	E GAG	F TTC	D GAT	T ACT	A GCG	S TCT
	FLY DC4	V GTG	Q CAG	A GCC	A GCC	F TTC	D GAC	T ACC	A GCC	N AAC
	FLY DC3	V GTG	Q CAG	A GCC	A GCG	Y TAC	D GAT	T ACC	A GCA	E GAG

3.4. HCP9 resembles rat testis-specific cyc gene

As shown in Fig. 4, pseudogene *HCP9* appears to be very old, as it was placed near the root of the tree. This pseudogene also has one of the largest sequence divergences from the modern *HCS* gene at 0.473 ± 0.071 per site per Myr (Table 1). Furthermore, it is disrupted into three fragments by two DNA insertions, both containing many retrotransposons. As discussed earlier, it is difficult to calculate the age of the pseudogenes based on sequence divergence; however, in this case we could actually deduce a lower boundary for the age of HCP9 by estimating the age of the retrotransposons contained in the inserted sequences. Using the RepeatMasker program (Smit, AFA & Green, P, URL:http://repeatmasker.genome.washington.edu/), several LTR sequences of MalR and ERVL types and several Alu sequences of AluJo and AluJb types were identified. It has been estimated that in the human genome, LTR/MalR and LTR/ERVL species had died out about 40 Myr ago (Smit, 1993; Cordonnier et al., 1995). The AluJo and AluJb sequences were ancient Alu species that were last active at around 81 Myr ago (Mighell et al., 1997; Smit, 1999). These facts indicated that HCP9 was inserted into the genome at least 80 Myr ago, which was before the divergence between human and prosimians (55–80 Myr) and after the estimated time for eutherian mammalian radiation (~100 Myr) (Lander et al., 2001). This particular pseudogene must have been inherited from a mammalian ancestor long before primate lineage emerged.

The phylogenetic tree also placed *HCP9* on a separate branch together with two testis-specific rodent genes. To better understand the origin of this ancient cyc pseudogene, we compared the nucleotide sequences between *HCP9* and the human and rodent cyc genes at the diagnostic codon positions where the somatic and testis-specific rodent cyc genes have different amino acids (Fig. 6). At ten of the thirteen positions, *HCP9* shares identical amino acid and almost identical codons with the testis-specific rat cyc gene (*CYCT_RAT*) rather than with the somatic rodent cyc genes. Hence the result from sequence comparison was consistent with what was inferred from the phylogenetic analysis: that the human pseudogene *HCP9* had a common origin with the rodent testis-specific cyc genes.

The testis-specific cyc genes are found only in rat and

Fig. 5. Sequence alignment at nine codon positions of the human cyc pseudogenes and the functional cyc genes from human, rodents and chicken. For each codon position, the sequences that have the same amino acid with the *HCS* gene are shown in pink; the sequences that share the same amino acid with the majority of the human pseudogenes are shown in green. The pseudogenes were divided into two classes based on their sequence identity with the *HCS* gene. A blank at a codon position indicates a missing sequence caused by truncation, and dashes '−' indicate gaps caused by DNA deletion. Frame shifts and stop codons are indicated by '/', '\' and 'X'. (**HCP47* sequence is merged into *HCP46*).

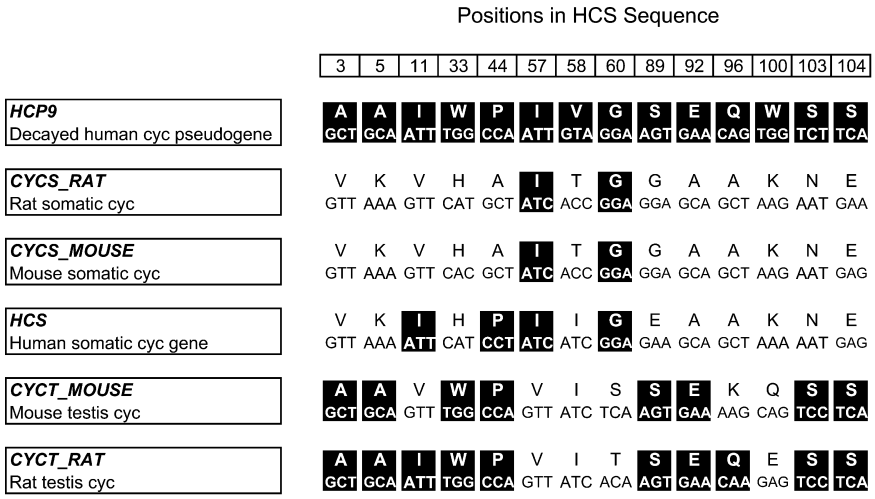


Fig. 6. Sequence comparison between pseudogene *HCP9* and the somatic and testis-specific cyc genes from rodents and human at selected codon positions. The positions where *HCP9* and the rodent testis-specific genes share an identical amino acid are highlighted.

mouse and possibly in bull and rabbit (Kim and Nolla, 1986), but not in human or other primates. It is likely that a functional cyc gene similar to the modern rodent testis-specific gene existed in the genome of an ancient mammalian ancestor. While modern rodents have kept the functional gene, humans only retained the pseudogene and lost the functional copy. It has been reported that none of the rodent cyc pseudogenes discovered so far originated from the testis-specific genes (Wu et al., 1986), however, all of these pseudogenes were discovered by genomic hybridization experiments instead of by computationally scanning the genome. As with the human genome, we expect many more cyc pseudogenes, and possibly testis-specific cyc pseudogenes, to be discovered in the mouse after the complete mouse genome sequence becomes available.

3.5. Online database

The pseudogene sequences described here have been deposited to GenBank with accession numbers: AF533162–AF533210. The data and results discussed in this report can be accessed online at http://bioinfo.mbb.yale.edu/genome/pseudogene/human-cyc/orhttp://pseudogene.org/.

4. Discussion

The 49 cyc pseudogenes we describe here present an evolutionary record of the human cytochrome c gene; our findings strongly support the hypothesis that this gene has evolved at a very rapid rate in the recent human lineage. The sequence information we report here will not only aid researchers to design better *HCS*-specific probes to avoid pseudogene complications, but will also be very useful in calibrating and estimating various evolutionary and phylogenetic models. The discovery of the common origin between pseudogene *HCP9* and the rodent testis-specific cyc genes will also improve our understanding of the relationship between gene expression and cell development.

Traditionally, most of the pseudogenes reported in literature were discovered by screening a genomic library using DNA hybridization techniques. As has been demonstrated in this study and other reports, such experiments often overlook the bulk of the pseudogene population. The discovery of such a great number of cytochrome c pseudogenes also raises the question as to the total number of pseudogenes in the human genome; such an estimate is important in the accurate prediction and annotation of functional genes. Differentiation between functional genes and disabled pseudogenes in genome annotation has proven to be a challenging and difficult task. For instance, it was suggested that in the *Caenorhabditis elegans* genome a fifth of annotated genes could be pseudogenes (Mounsey et al., 2002). With the advent of the complete human genome sequence, a systematic and comprehensive survey of pseudogenes is much needed, not only to provide better functional gene annotation, but also to extend our understanding of the evolution of genes and genomes as a whole. We also did a preliminary survey in the recently published mouse draft genome sequence (Waterston et al., 2002) and detected about 40 cytochrome *c* processed pseudogenes. However, the relative low quality of the mouse sequence did not allow for detailed comparison between these two sets of pseudogenes.

Acknowledgements

MG acknowledges NIH grant 2P01GM54160-04. Z.Z. thanks Dr. Paul Harrison for comments on the manuscript and Dr. Duncan Milburn and Nat Echols for computational help.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E., 2002. Recent segmental duplications in the human genome. Science 297, 1003–1007.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45–48.

Banci, L., Bertini, I., Rosato, A., Varani, G., 1999. Mitochondrial cytochromes c: a comparative analysis. J. Biol. Inorg. Chem. 4, 824–837.

Biel, S.W., Biel, A.J., 1990. Isolation of a *Rhodobacter capsulatus* mutant that lacks c-type cytochromes and excretes porphyrins. J. Bacteriol. 172, 1321–1326.

Chothia, C., Lesk, A.M., 1985. Helix movements and the reconstruction of the haem pocket during the evolution of the cytochrome c family. J. Mol. Biol. 182, 151–158.

Cordonnier, A., Casella, J.F., Heidmann, T., 1995. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. J. Virol. 69, 5890–5897.

Esnault, C., Maestre, J., Heidmann, T., Human, L.I.N.E., 2000. retro-transposons generate processed pseudogenes. Nat. Genet. 24, 363–367.

Evans, M.J., Scarpulla, R.C., 1988. The human somatic cytochrome c gene: two classes of processed pseudogenes demarcate a period of rapid molecular evolution. Proc. Natl. Acad. Sci. USA 85, 9625–9629.

Grossman, L.I., Schmidt, T.R., Wildman, D.E., Goodman, M., 2001. Molecular evolution of aerobic energy metabolism in primates. Mol. Phylogenet. Evol. 18, 26–36.

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., Gerstein, M., 2002a. A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. J. Mol. Biol. 316, 409–419.

Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., Gerstein, M., 2002b. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res. 12, 272–280.

Kazazian, H.H. Jr, Moran, J.V., 1998. The impact of L1 retrotransposons on the human genome. Nat. Genet. 19, 19–24.

Kim, I.C., Nolla, H., 1986. Antigenic analysis of testicular cytochromes c using monoclonal antibodies. Biochem. Cell Biol. 64, 1211–1217.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kluck, R.M., Bossy-Wetzel, E., Green, D.R., Newmeyer, D.D., 1997. The release of cytochrome c from mitochondria: a primary site for Bcl-2 regulation of apoptosis. Science 275, 1132–1136.

Kumar, S., Tamura, K., Jakobsen, I.B., Nei, M., 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, W.-H., 1997. Molecular Evolution, Sinauer Associates, Sunderland, MA.

Limbach, K.J., Wu, R., 1985. Characterization of two Drosophila melanogaster cytochrome c genes and their transcripts. Nucleic Acids Res. 13, 631–644.

Mighell, A.J., Markham, A.F., Robinson, P.A., 1997. Alu sequences. FEBS Lett. 417, 1–5.

Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F., 2000. Vertebrate pseudogenes. FEBS Lett. 468, 109–114.

Mills, G.C., 1991. Cytochrome c: gene structure, homology and ancestral relationships. J. Theor. Biol. 152, 177–190.

Mounsey, A., Bauer, P., Hope, I.A., 2002. Evidence Suggesting That a Fifth of Annotated *Caenorhabditis elegans* Genes May Be Pseudogenes. Genome Res. 12, 770–775.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics, Oxford University Press, Oxford, New York.

Ostertag, E.M., Kazazian, H.H. Jr, 2001. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. 11, 2059–2065.

Pearson, W.R., 1997. Comparison of DNA sequences with protein sequences. Genomics 46, 24–36.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Scarpulla, R.C., 1984. Processed pseudogenes for rat cytochrome c are preferentially derived from one of three alternate mRNAs. Mol. Cell Biol. 4, 2279–2288.

Scarpulla, R.C., Agne, K.M., Wu, R., 1982. Cytochrome c gene-related sequences in mammalian genomes. Proc. Natl. Acad. Sci. USA 79, 739–743.

Smit, A.F., 1993. Identification of a new, abundant superfamily of mammalian LTR- transposons. Nucleic Acids Res. 21, 1863–1872.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. 19, 253–272.

Virbasius, J.V., Scarpulla, R.C., 1988. Structure and expression of rodent genes encoding the testis-specific cytochrome *c*. Differences in gene structure and evolution between somatic and testicular variants. J. Biol. Chem. 263, 6791–6796.

Waterston, R.H., Lindblad-Toh, K., Birney, E., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17, 149–163.

Wu, C.I., Li, W.H., Shen, J.J., Scarpulla, R.C., Limbach, K.J., Wu, R., 1986. Evolution of cytochrome c genes and pseudogenes. J. Mol. Evol. 23, 61–75.

Zhang, Z., Harrison, P., Gerstein, M., 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res. 12, 1466–1482.