Available online at www.sciencedirect.com



Gene xx (2003) xxx-xxx



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 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.

© 2003 Elsevier Science B.V. All rights reserved.

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

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.

55 \* Corresponding author. Tel.: +1-203-432-6105; fax: +1-360-838-7861.
 56 *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

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 



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 116 (RP) pseudogenes in the human genome (Harrison et al., 117 2002b; Zhang et al., 2002), most of which were previously 118 overlooked by DNA hybridization experiments. Motivated 119 by this discovery of an unexpectedly large number of addi-120 tional pseudogenes, we carried out a similar comprehensive 121 survey on human cytochrome c pseudogenes. Our study 122 provides a complete molecular record of the recent evolu-123 tion of this gene and demonstrates the importance of 124 examining pseudogenic sequences. It also demonstrates a 125 specific instance of a gene disappearing and leaving only a 126 fossil pseudogene in its place. 127

#### 130 **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, 138 2001, downloaded from Ensembl website (http://www. 139 ensembl.org). Subsequently, all the chromosomal coordi-140 nates were based on these sequences. The amino acid 141 sequences of the cytochrome c proteins were extracted from 142 SWISS-PROT (Bairoch and Apweiler, 2000). Each 143 un-masked human chromosome was split into smaller 144 145 overlapping chunks of 5.1 MB, and the tblastn program of the BLAST package 2.0 (Altschul et al., 1997) was run on 146 these sequences. The default SEG (Wootton and Federhen, 147 1993) low-complexity filter parameters were used in the 148 homology search. We then picked the significant homology 149 matches (e-value  $< 10^{-4}$ ), and reduced them for mutual 150 overlap by selecting the matches in order of decreasing 151 significance and removing any matches that overlapped 152 substantially with a previously-picked match (i.e. more than 153 ten amino acids or 30 bp). 154

156 2.2. Alignment optimization by FASTA dynamic157 programming

After the BLAST matches were sorted according to their 159 starting positions on the chromosomes, they were examined, 160 and the neighboring matches were merged if they were 161 deemed to be part of the same pseudogene sequence. The 162 merged matches were then extended on both sides to equal 163 the length of the cyc gene plus 30 bp buffers. For each 164 extended match, the human cytochrome c (HCS) amino acid 165 sequence was re-aligned to the genomic DNA sequence 166 using the program FASTA (Pearson, 1997). FASTA utilizes 167 168 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 175 existence of exon structures. One sequence on chromosome 176 7 was identified as the functional HCS gene, as its sequence, 177 including the exons, introns and the flanking regions, 178 matched perfectly with the previously known functional 179 HCS gene. Forty-six (46) pseudogene candidates had 180 continuous, intron-less coding regions, which suggested 181 that they were processed pseudogenes; these sequences 182 were labeled as 'intact' processed pseudogenes. The three 183 remaining pseudogene candidates contained retrotrans-184 poson sequences inserted in their otherwise continuous 185 coding regions; they were labeled as 'disrupted' processed 186 pseudogenes. We further extended the pseudogene 187 sequences on both sides to obtain the 5' and 3' un-translated 188 (UTR) sequences. 189

#### 2.4. Phylogenetic analysis and dating

Multiple sequence alignment of the pseudogenes 193 and genes was performed using the program ClustalW 194 (Thompson et al., 1994). MEGA2 (Kumar et al., 2001) was 195 used for all the phylogenetic analysis. A phylogenetic tree 196 was constructed by applying the neighbor-joining (NJ) 197 method (Saitou and Nei, 1987; Nei and Kumar, 2000) to the 198 protein coding regions. For each cyc pseudogene, we also 199 calculated the nucleotide sequence divergence from the 200 modern HCS gene, using Kimura's two-parameter model 201 (Kimura, 1980), which corrected for multiple hits and also 202 took into account different substitution rates between sites 203 and for transitions vs. transversions. We calculated the ages 204 of some young pseudogenes from the sequence divergence, 205 using formula T = D/(k), where D is the divergence and k 206 is the mutation rate per year per site. A mutation rate of 207  $1.5 \times 10^{-9}$  for pseudogenes was used (Li, 1997). 208

#### 3. Results

209 210

172

173

174

190

191

192

### 211 212

213

214

#### 3.1. The human cyc pseudogene population

A total of 50 cyc homology loci were identified in the 215 human genome, including 49 pseudogenes (denoted as 216 HCP) and one intron-containing functional gene (denoted as 217 HCS). The HCS gene was located on chromosome 7 218 (cytogenic band 7p15.3, see Fig. 1), the annotation was 219 confirmed by the perfect alignment of the exons, intron, and 220 the 5' and 3' regions with the previously reported nucleotide 221 sequence ((Evans and Scarpulla, 1988), GenBank ID: 222 181241). It is known that the HCS gene contains two 223 introns. The first one is 1,073 bp long and 9 bp upstream of 224

2

128

129

131

132

133

134

135

136

137

155



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 270 101 bp long and precedes the second nucleotide of the 56th 271 codon. The 49cyc pseudogene assignments were established 272 by their lack of both introns and were, in some cases, further 273 confirmed by the existence of a poly-A tail at the 3' end. 274 Most of the pseudogenes (40 of 49) were also found to 275 contain obvious disablements in their coding regions. We 276 further searched the GenBank human EST database to 277 confirm that none of these pseudogenes was expressed. 278

We named our cyc pseudogenes sequentially from HCP1

to HCP49 according to their locations on the chromosomes.

269

279

280

These pseudogenes are spread out on 18 of the 23 326 chromosomes, except 5, 10, 18, 19, 20 and 22 (see Fig. 1). 327 Fig. 2 shows the alignment of the predicted amino acid 328 sequences of these pseudogenes with the HCS protein. The 329 disablements are highlighted in gray. More detailed 330 information on these pseudogenes is provided in Table 1. 331 Except for HCP9, HCP15 and HCP30, which are disrupted 332 into two or three fragments by insertions of retrotrans-333 posons, most of the pseudogenes have continuous 334 sequences. 335

323

324

325

The sequences of 40 of the 49 pseudogenes can be 336

Z. Zhang, M. Gerstein / Gene xx (2003) xxx-xxx

	1		10	20		30		40	50	60 30	03
HCS HCP1	M G D V E M G D V E	K G K K K G K K	IFIMK- IFVQK-	C S Q C H T V E C A Q C H T V E	К G G К Н К G G К Н	K T G P N L K T G P N L	HGLF	F G R K T G Q F S Q K T G Q	A P G Y S Y T - A A N K A V G F S Y T - D A N K	NKGIIWG 30	93 04
HCP2 HCP3	T G D I E M G D V E	K G K K K G R R	ICVQK- LLFRV-	C A Q C H T V E C S V P H	KGGKH	E T G P N L G P N L	H/GLF HGLF	GWKTGQ GKQMS-	A T G F S Y T - D A N K S /P G F S L T - D I S K	NKGITWG NKGIIWG 30	95
HCP4 HCP5 HCP6	M G D V E V G D V E	кзкз	VFVXK- IFVQK-		KGGKH	K T G P N L K T G P N L K T G P N L	HGLG	C G C K T G Q F G L K T G Q	A	NKCIIYG NKEITWG 30	96
HCP7 HCP8	M S D V E R S N V V	K G N K K G K K	I F V Q K -	FAQCHTV /CTQCHTV	K G G K R K R G K H /	K T E T N L K T G P E L	HGLI	G Q N T G Q G /R Q P G Q	AVGFSYM-DPNK AIGSPYT-DAN/K	NKGITWG NT/GITWG 30	97
HCP9 HCP10 HCP11	TSDVE	KSK- KGKK	I F I Q K - I F V Q K - I F V Q K -		KGENH KGGKH KGDKH	K T R P N L K T R P N L K S G P N L	QGLI	- G R K I G K - G R K T G Q - V R K T G Q	A	NKGITWG 30	98
HCP12 HCP13	MGDAE	KGKK	I F V Q K -	CAQCHTVE	K R G K H K R G K H	K T G P N L K T G P N L	HGLI	G /R	-   G F S Y T - D A K K -   G F S Y T - D A K K	NKGITWG 30	99
HCP14 HCP15 HCP16		KGKK	V F F L K K I F I M K - I F V O K -	C V L C F T C E R S Q C H I V E C A O C N T M E	K R G IR R K G G K H K G G K -	K I N L K T G P N L P V I	HGQ HGLF SALV	/G A R D N Q     G R K T G Q     V F A F T G F	A P G F L Y T - D H N K A P G H S Y I - A T I K I R K P G R S R P	NKDIIWG AWGAXWG 40	.00
HCP17 HCP18	M S D V E	KGKK KGK-	IFIQK- TFV/MK-	C V Q W H T M H C A P C N P K H	K E G K /H G S \G K H	E T G L N L K C A L I L	HG/LI	G /R K T G Q G /R R T D Q	V I G F S Y T - D S N K A P T F /S F V - V A K K	NKGITXG SKGISXG 40	.01
HCP19 HCP20 HCP21		K G K K K S T N K G K K	I C I Q K - I C V Q K - I F I M K -	CAQCHIVE	K A G K H K G G K H K G G K H	K	HGLF	- G W K T G Q = G W K T G - = R W K T G Q	A V G L F Y I - D A I K A I G F C Y T - D T S K A P G Y S Y T - A A N K	NKGITWG NKGITWG 40	.02
HCP22 HCP23	MGDVE	KGKK	I F V Q K - I F V Q K -	C V Q C H I M I C A Q F H T V I	K G G K H K G G K Q	K T G P N L V A G S N L	HGLF	G X K T G Q V G K T D Q	A V G F S Y T - D A N K A I G F S Y T - E A H K	NKGITWG NKSISWG 40	03
HCP24 HCP25 HCP26		KGKK KGK K	FVQK- FVQK-	CAQCHTVE	KGGKH KGGKH	K T G P N L K T G P N L K T G T N L	HGLF	= G Q K T G Q = G Q K T G Q = M X K T D Q	A V G F S Y I - D A N K A I V F T Y T - D A N K A V G F C IY T - D A N K	NKGITWG NKGITWR 40	04
HCP27 HCP28		KGKS	ICTEK- TFVEE-	Y A Q C Y T V E C V Q C H T M E	N K G K D K G G X H	K T R P N L K T A S N L	HGIE	GQKTDQ	A I G F S L T - D A S K D A G F S Y T - D A N K	NKRITWG NKGTTXG 40	05
HCP29 HCP30 HCP31	MSDVE	K G K K	IFVQK- IFVQK-	CAPCHTVE	RK/GKH KGGKH	K T G P N L K T G P N L K T E H N L	HGLF	= V Q K T G Q = G Q K T G Q	A V G F S F T - D T N K A Y T - D A Y K A V G F S Y T - D A N K	NKGITLG 40	06
HCP32 HCP33	MGDVEKRDVE	KGKK	I F V Q K -	CAQCHTVE	K G G K Q N G G K H	K T G P N L K T G P N L	HGLE	F R W K T G Q	A I G L S Y T - E T D K A I G L S Y I - D T D K	NKGITWG 40	07
HCP34 HCP35 HCP36	MSDAE	KKKKQ KKKKQ	T N K K R - I F V Q K -	C T P C H T M E C A Q C H T V E	N G G K H K X G Q H	N T G P N L I T G P N P K T R P N L	HGLF	= G Q K T S Q . = G R K T G Q .	A	NKGITWG 40	08
HCP37 HCP38	KGDVE MGDVE		MCAQK-	HAQCHIME	EGGQQ	K T G P N L K T G P H L K T G P N	HGLF	F G K K T G Q V M IR T G Q E	A V G F T Y T - D A N K A F G F T Y T - D A S K A - G F S Y T - D T S K	KKSIAWG NKSITXR 40	09
HCP40 HCP41	MDDVE	KGK- KGKK	IFVPKG IFVQK-		K G S K H K G S K H	K T G P H L K T E P S L	HRLS	S G Q T T G Q R W K T G Q	A A G F S Y T A D A N K A I G L S IY I - E I D K	NKSITWS 41 NKAITWG 41	10
HCP42 HCP43 HCP44	VGDVE	KGKG	V F V Q N - I F A Q K - V F V O K -	Y S K W H T V E C A Q X H T V E S A Q C H T M E	KGGMH KEVKH	K T G P G V K P G P D L O D W A N I		G V D N G S G W K T G Q G R K T G O	G P G F S Y T - D A N K A V E F C Y T - D A N K A V G F S Y T - D A H K	NKGSAW- NRGINWG 41	11
HCP45 HCP46	MGDVE	E K G K K	ÎFÎMK-	ĊŚQĊHŦVI	KĞĞKH	ΚΤĞΡΝĽ	QGLI	G R K T G Q K T G Q	A A G Y S Y T - V A S K A P A Y S Y T - A T N K	NKGITWG 41 NKGIICG 41	12
HCP47 HCP48 HCP49	MADSE	ЕКСКК	K-	C S Q C H T V E C A Q C H T K V C A Q C H T V E	K G G K H K X G K H K G S K H	K T G P N L K S E P N L K T G P K L	HGLE	- G X K T - M Q K T G Q - R W R T C Q	A T <mark>/G/Y</mark> SLT-DTNE SIGLSWI-EIDK	NRGITXG 41	13
									_	4	14
	61		70	8	0		90		100	41	15
HCS HCP1	EDT EDT	- L M E	Y L E N P K Y L E N P K	K Y I P G T K N K Y I P G T K N		K K K K K K	E E R A		K K A T N E K K A T N E	41	17
HCP3 HCP4	EET	L M N	Y S E I /P Q Y L Q N P K	ENIPSTK	I I F T G I I T I V S T	ККЕ ККЕ ККК	TERT		K K R K A N N Q	4	.18
HCP5 HCP6 HCP7	EDT	/ME 	Q	X Y I P G T K N N Y I P G T K N K C I I G T K N	S T G V     F A G         F A S	K R K Q K K K K K		I G /L L P Y L A D L I A Y L A G L I A	L R N Q N K K S	4	19
HCP8 HCP9	EET - ·	- PVE - LME	Y L K N P E Y L E N P K	NYILGTK KYIPGTNM	V F T G I I T F S G L	N К К К К К	VER		K K A T N E K W A T S	42	20
HCP10 HCP11 HCP12			Y L E N P K Y L E N P K D L E N P K	KY I PGRKI KY I PGRKI		ккк КК КК		A N L I A X L A N L I A X L A D F I A Y L	K K A T N E K K A T N E K K A T N E	42	21
HCP13 HCP14	EDT N/DV	- L M M E		KYIPGRKI KKNVSL/G	IFASI KLINI	K K R	ADRA		K K VA T N E E K A T S E	42	22
HCP15 HCP16 HCP17	EDT - ·	L M E L I L K E	YLENLK	KY I PGTK	XYFLN	ткк	AERA		EKATNE	42	23
HCP18 HCP19 HCP20	EDTW- EDT	- L M E	Y S E N P R Y L E N P K Y I E N P K	K Y L I G E K I K H I S G T K I N X I P G T K N		ккка	- E R /		ККАТЛЕ	42	24
HCP21 HCP22	EDT	- L M E	Y L E N P K Y L E T P K	K Y I P G T K N K H I L G T K N	I I F V S I I I F A C I	ккк кк	E E R A V E R A		K K A T N E K K A T N E	42	25
HCP23 HCP24 HCP25		L M E L M E	Y L E N P K Y L E N P K Y L E S P K	K H I P G I K M K Y I P G T K M K Y I P R I K M	F A G M     F A G       F A G	K N K K K K K K K		A A L I A H L K - M I A Y L A D L L A C L	K K VA I N E K K S K K A T	42	26
HCP26 HCP27	EDA	L M K	Y L E N P K H L E N P X	K X I S G T K N		ККАЕ/ЕК КККА	A E R /		K K K K A T N E	42	27
HCP28 HCP29 HCP30		- LME - LMH - LRE	Y L E N P K Y L E T P K D L E N P K			ккк ккк кк	AERA		K N A T H Q T K A T N E	42	28
HCP31 HCP32	EDT	L M E	Y L E N P K Y F E N P R	K Y I P G T K N K Y I P G T K N		K K K K N N	AERA		K K S K K A T S E	42	29
HCP33 HCP34 HCP35	EDT EET	L M E P M G	Y V E N P K S X E N P K	KYIPGTKI KYIPGTE	IFAGI	K K K K E R				43	30
HCP36 HCP37 HCP38		- L M E	Y L E T P K Y L E N P K V L E N P K	KYIPGTKV KSIPGTK		ККК NК К А Т N	T E R A V E R A	A D L I A Y L A D L I A Y L	ККАТІЕ ККАТN	4:	31
HCP39 HCP40	EDT - ·	- L M E	Y L E N P K Y L E N P -	KYIPGTKN \GTKN	I I F A G I I V F T G	KKE	AERA	ADLIAYL	ККАТЛЕ	4:	32 32
HCP41 HCP42 HCP43	E D T - R T E G T		Ү L M C P K Ү L E N H K			Q X R K K	AFR		KNA	43	33 34
HCP44 HCP45	K ND T	- LME	YLENPK	K Y I P G T K N K H I P E T K N	I F S G I I F V G I	ккк ккк	TERA	A E F I A Y L A D S I A F L		42 A'	.35
HCP46 HCP47 HCP48	EDT	L M E ' L M E '	T L E N P K Y L Q N P K	KYIAGTKI	ITIASTI	ккк ККК	AERA		RKANNE	4.	.36
HCP49	EDT	- L I É '	YFENSR	KYIPGTKI	IIFASI	K N N	AERA	ADLIAYL	ККАТЅЕ	43	37
Fig. 2. Align	ment of the	translated	amino acid	sequences of the l	iuman cyc ps	eudogenes,	together	with the function	onal HCS protein sequence	. In the pseudogene 4	38

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 '/', 'V' 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.443Interestingly, this gene fragment is immediately adjacent to444*HCP46* on chromosome Y, which matches HCS residues44539–104 but on the opposite chromosomal strand. It appears446that *HCP46* and *HCP47* were once parts of an original447complete cyc pseudogene that had undergone '5' inversion',448

### 

### Z. Zhang, M. Gerstein / Gene xx (2003) xxx-xxx

#### Table 1

Detail information on the 49 human cytochrome c pseudogenes<sup>a</sup> 

7-150		(%)		
/p15.3	(-) 25.65M			Functional human cyc gene
AF533162 1q21.3	(+) 151.49M	88	$0.090 \pm 0.020$	, ,
AF533163 1q23.1	(+) 156.78M	82	$0.114 \pm 0.023$	
AF533164 1q24.3	(-) 172.70M	59	$0.235 \pm 0.042$	
AF533165 1q32.1	(-) 206.92M	73	$0.138\pm0.027$	Truncated before residue 13
AF533166 1q44 (	-) 251.92M	64	$0.345 \pm 0.054$	
AF533167 2p12 (	-) 79.59M	77	$0.113 \pm 0.022$	
AF533168 2q11.2	(+) 96.95M	74	$0.182 \pm 0.033$	
AF533169 2q14.3	(-) 127.35M	69 70	$0.220 \pm 0.037$	Dismonto di inte di mas for emergito has Alere
AF533170 2q31.2 AF533171 2p25.3	(-) $1/7.34M$	/9 91	$0.4/3 \pm 0.0/1$ 0.107 ± 0.022	HCP10, 12 are duplicated copies
AF533172 3p25.3	(-) 11.87M	83	$0.107 \pm 0.022$ 0.095 ± 0.020	See HCP10
AF533173 3p25.3	(-) 14.05M	76	$0.093 \pm 0.020$ $0.091 \pm 0.020$	See HCP10
AF533174 3p25.1	(-) 19.78M	77	$0.091 \pm 0.020$	See <i>HCP10</i>
AF533175 4a28.3	(-) 131.99M	46	$1.257 \pm 0.253$	
I) AF533176 6p21.1	(+) 44.06M	92	$0.041 \pm 0.013$	Disrupted into two fragments by Alus
AF533177 6q15 (	-) 96.18M	48	N/A	Truncated after residue 58
AF533178 6q16.1	(+) 101.62M	74	$0.147 \pm 0.028$	
AF533179 7q21.3	(-) 97.54M	52	$0.610 \pm 0.117$	Truncated after residue 81
AF533180 7q31.3	2 (+) 121.29M	83	$0.117 \pm 0.023$	
AF533181 7q32.1	(-) 132.78M	76	$0.142 \pm 0.026$	
AF533182 8p12 (	-) 34.34M	95	$0.034 \pm 0.011$	
AF533183 8q11.2	2 (-) 51.04M	83	$0.108 \pm 0.022$	
AF533184 8q24.1	2(-) 120.5/M	76	$0.164 \pm 0.029$ 0.102 ± 0.021	UCD24 and 21 are duplicated copies
AF555185 9q22.5	$(-) \qquad 31.51M$	83	$0.103 \pm 0.021$ 0.004 ± 0.020	<i>HCP24</i> and <i>51</i> are duplicated copies
AF533187 11a13	2(-) 66.72M	75	$0.094 \pm 0.020$ 0.148 + 0.027	
AF533188 11q13.	4(+) 75.78M	68	$0.173 \pm 0.031$	
AF533189 11q14.	1 (+) 78.18M	71	$0.173 \pm 0.031$	
AF533190 11q22.	3 (-) 113.54M	72	$0.140 \pm 0.026$	
AF533191 12q21.	32 (+) 91.41M	72	$0.177 \pm 0.032$	Disrupted into two fragments
AF533192 13q12.	11 (-) 17.40M	83	$0.103 \pm 0.021$	See HCP24
AF533193 13q12.	11 (-) 18.53M	79	$0.123 \pm 0.024$	HCP32, 41 and 49 are duplicated copies
AF533194 13q12.	12 (-) 23.63M	77	$0.129 \pm 0.025$	
AF533195 13q14.	11 (+) 36.85M	86	$0.088 \pm 0.019$	
AF533196 13q32.	3 (-) 99.53M	64	$0.256 \pm 0.042$	
AF533197 14q24.	5 (+) /5.65M	82	$0.108 \pm 0.022$ 0.226 ± 0.020	
AF533100 15022	2(+) 53.24M 2(+) 57.30M	70 61	$0.220 \pm 0.039$ 0.287 ± 0.040	
AF533200 16n12	1(-) 26 36M	84	$0.093 \pm 0.020$	
AF533201 17a25.	3 (+) 78.64M	68	$0.174 \pm 0.036$	Truncated after residue 84
AF533202 21q11.	2 (-) 7.75M	71	$0.198 \pm 0.044$	Truncated after residue 63, see also HCP32
AF533203 21q21.	l (-) 13.53M	58	$0.458 \pm 0.084$	Truncated after residue 88
AF533204 Xq13.1	(-) 63.83M	75	$0.170\pm0.031$	
AF533205 Xq27.3	(+) 140.82M	77	$0.124\pm0.024$	
AF533206 Xq28 (	+) 149.47M	91	$0.052\pm0.014$	
AF533207 Yq11.2	21 (-) 16.31M	92	$0.047 \pm 0.014$	Truncated before residue 39
AF533208 Yq11.2	21 (+) 16.31M	96 71	-	Residues 14–40 only.
AF533209 Yq12 (	+) 27.78M	71	$0.165 \pm 0.031$	Truncated before residue 13
AF533210 Yq12 (	+) 27.93M	/1	$0.1/6 \pm 0.031$	See HCP32
AF533 AF533 AF533 AF533 AF533 AF533 AF533 AF533 AF533 aF533 AF533	203         21q21.1           204         Xq13.1           205         Xq27.3           206         Xq28 (-           207         Yq11.2           208         Yq11.2           209         Yq12 (-           210         Yq12 (-           are indicated by         s of the pseudogo	204 $Xq13.1(-)$ 13.35M 204 $Xq13.1(-)$ 63.83M 205 $Xq27.3(+)$ 140.82M 206 $Xq28(+)$ 149.47M 207 $Yq11.221(-)$ 16.31M 208 $Yq11.221(+)$ 16.31M 209 $Yq12(+)$ 27.78M 210 $Yq12(+)$ 27.93M are indicated by *; the rest of the pseudog 5 of the pseudogene in Mb (million base p	203 $21q_{21,1}(-)$ $13.5M$ $38$ 204 $Xq_{13,1}(-)$ $63.83M$ $75$ 205 $Xq_{27,3}(+)$ $140.82M$ $77$ 206 $Xq_{28}(+)$ $149.47M$ $91$ 207 $Yq_{11,221}(-)$ $16.31M$ $92$ 208 $Yq_{11,221}(+)$ $16.31M$ $96$ 209 $Yq_{12}(+)$ $27.78M$ $71$ 210 $Yq_{12}(+)$ $27.93M$ $71$ are indicated by *; the rest of the pseudogenes are class 2. $36$ of the pseudogene in Mb (million base pair).	$\begin{array}{cccccccccccccccccccccccccccccccccccc$



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 565 regions of 45 human cyc pseudogenes and the HCS mRNA; 566 the 5' flanking regions were also included for the pseudo-567 genes. The four pseudogenes that are truncated near the 5'568 end are not included in the alignment. The two downward 569 arrows mark the start of the HCS mRNA sequence and the 570 ATG translation initiation codon. As can be seen, most of 571 these pseudogenes have retained the nearly intact 5' UTR 572 sequence. This high degree of sequence preservation is a 573 little surprising, as it has been known that LINE1-mediated 574 reverse-transcription has a low efficiency and often leads to 575 5' truncation and thus incomplete insertion of mRNA 576 transcripts into the genome. 577

As outlined in bold in Fig. 3, three groups of the pseudo-578 genes share almost identical 5' flanking sequences. This 579 indicated that the pseudogene sequences within each group 580 arose from genomic duplications of an original pseudogene, 581 rather than from independent reverse-transcription events, 582 and that the sequences had, therefore, retained the flanking 583 sequence of the original pseudogene. The pseudogenes in 584 the first group (HSP10, HSP11, HSP12 and HSP13) were 585 located very close to each other on chromosome 3 (see 586 Table 1 and Fig. 1). This suggested an intra-chromosome 587 sequential duplication event. The two other groups (the first 588 consisting of HCP31 and HCP24 and the other consisting of 589 HCP32, HCP41 and HCP49) appeared to have resulted 590 from inter-chromosomal duplications. Such extensive seg-591 mental duplications in the human genome have been 592 593 described recently (Bailey et al., 2002).

By screening human cDNA and genomic libraries, Evans 594 and Scarpulla (Evans and Scarpulla, 1988) previously 595 reported 11 human cyc pseudogenes, which were named 596 HC1-HC6, HS7, HC8, HC9, HC10, and HS11. We were 597 able to unambiguously assign ten of these eleven sequences 598 to a single pseudogene in our pseudogene set as indicated in 599 the leftmost column on Table 1. The remaining one, HC3, 600 has identical to a pair of duplicated pseudogenes: HCP24 601 and HCP31. Therefore, in addition to the previously 602 reported 11, we discovered 37 new cyc pseudogenes in 603 604 the human genome.

#### 606 3.2. Phylogenetic analysis

605

607

We were interested in tracing the origin of these cyc 608 pseudogenes and placing them into the context of evolution. 609 Fig. 4 shows the phylogenetic tree constructed by applying 610 the neighbor-joining (NJ) method (Saitou and Nei, 1987; 611 Nei and Kumar, 2000) to the protein-coding regions of 612 human cyc pseudogenes and the functional cyc genes from 613 human, rat, mouse, chicken and fruitfly. Rodents have two 614 cyc genes in their genomes: the somatically expressed genes 615 616 (CYCS\_RAT, CYCS\_MOUSE) and the testis-specific genes

(CYCT\_RAT, CYCT\_MOUSE). These testis-specific cyc 617 genes are only expressed during spermatogenesis (Virbasius 618 and Scarpulla, 1988). Compared with their somatic 619 counterparts, they have different exon structures and differ 620 at 14-15 amino acid positions. Fruitfly also has two cyc 621 genes, FLY\_DC4 and FLY\_DC3, which differ at 32 amino 622 acid positions (Limbach and Wu, 1985); it was believed that 623 they diverged about 520 Myr (million years) ago (Wu et al., 624 1986). FLY\_DC4 has a much higher expression level in the 625 cell than *FLY\_DC3*, and was used to root the phylogenetic 626 tree. 627

As expected, the two fruitfly genes were clearly sepa-628 rated from the vertebrate sequences. Also, the chicken gene 629 and the rodent testis-specific genes were placed close to 630 each other and distant from the mammalian somatic genes 631 and the majority of the human pseudogenes (except HCP9). 632 It was postulated that these tissue-specific cyc genes arose 633 from duplication of an ancestral cyc gene (Limbach and 634 Wu, 1985) and the estimated divergence time of these genes 635 from somatic genes was close to the divergence time of 636 birds and mammals (Mills, 1991). 637

Table 1 lists the nucleotide sequence divergences 638 between each cyc pseudogene and the modern HCS gene 639 calculated according to Kimura's two-parameter model 640 (Kimura, 1980). Sequence divergence, or the number of 641 nucleotide substitutions between sequences, is a measure of 642 evolutionary distance between two sequences. In this case, 643 the divergence values were correlated with the ages of the 644 pseudogenes, i.e. the approximate time when each pseudo-645 gene was inserted into the genome. It might be expected 646 that, on average, the older pseudogenes should have greater 647 divergence than the younger ones. However, special care 648 has to be taken in comparing divergence of pseudogenes, as 649 they contain not only the accumulated mutations in the 650 pseudogene sequences after they were inserted into the 651 genome, but also the sequence differences in the functional 652 genes from which they originated. It is rather tempting to 653 estimate the age of a pseudogene by simply dividing the 654 divergence by a constant nucleotide substitution rate. 655 However, we believe such a simplified calculation should 656 not be applied here for the cyc pseudogenes, as it assumes 657 that the pseudogenes all originated from the same ancestral 658 cyc gene and same mRNA transcript. As will be discussed 659 in Section 3.3, this is certainly not true for the cyc 660 pseudogenes. 661

#### 3.3. Two classes of cyc pseudogenes

Based on a comparison of the pseudogene sequences 665 with the modern HCS gene and consensus mammalian cyc 666 sequence, Evans and Scarpulla (Evans and Scarpulla, 1988) 667 divided their 11 human cyc pseudogenes into two classes. 668 The predominant class of older pseudogenes (denoted as 669 'class 2', nine members) appeared to have originated from 670 an ancient progenitor of the cyc gene, and the remaining two 671 pseudogenes (class 1, HS7 and HS11) were younger and 672

662

663

664

			10	20	30	40	50	60	70	
		*		1		I			1	+
HCS		GGGGAG/	A G A G T G G G G A	A C G T C C G G C T <sup>-</sup>	T C G G A G C G G G A	G T G T T - C G T - T	GTGCCAGCGA	CTAAAAAGAGA	АТТ - ААА	TATG
HCP21	T T A C C A A A A A A A G T C A G T T A A A A G T T A C	А G A A T T C	C	A. T	С Т	C A	. С Т		<del>.</del>	
HCP15	G A A A C C C C A T C T C C A C T A A A A	АСТА			. A	C	. C		<del>.</del>	
HCP45	A T A C A C A T T T G G A T T T G G T T T A G A A A A	АТТТТТ. Т		A		A. C C	. C	G		
HCP39	A T A C C A C C A T T T C T C T A A A A	<u>A G A</u> T A C	C A	A	. A A	A C C A	. C A .	. G C	. C	
HCP31	Т G C A A A T A T A T T T G A C T A T T A A A T T A T C T C	Т G Т Т Т Т			G . T A	A C C	. C	. A <b></b>		
HCP24	T G C A A A T A T A T T T G A C T A T T A A A T T A T C T C	т G Т Т Т Т			G. T.A	A C C	. C	. A <b></b>	•	
HCP22	G G T A G G A G G G G A A A A C A T T T A A A A A T A G C	ТАА СА. Т.	Т	т	A	A C C A	. C T . A A	T		
HCP1	Т С С Б Т С Т С А А А А А А А А А А А А А А А Т Т Б С С Т	ТТАТ		. т	. A T T.	A C C . A	. C A T	Τ G		
HCP6	T C T T C T T C C A T C A C T C C T T A G A A A A T G T	ттбт с (	3	A	. A A	A C T A	. C A T	Τ G	G	. G
HCP36	Т Т G Т G A Т C Т Т A A Т A Т C Т G A Т A A A A G A T A T	GТАТСТ			. A A T	A C C		T G C	•	
HCP16	д а с а а т д д а а а с а а а с а т а а д а т с а д	СААТ. СТТ		A AA	. Т ТА. А.	A C C T G	. са т	ΑΑΤ	<del>.</del>	C.
HCP10	A C C A A A A G G A C A A G C A A C A A C A A C A A A A	ААТ.А.		гт	тт.с.	A C C	G. C.C		•	C A
HCP13	A C C A A A A G G A C A A G C A A C A C	аатт. ат		ТА.А	тт.с.	ACC T G	. c	т G		
HCP12	A C C A A A A G G A C A A G C A A C A C	ΑΑΤΤ.ΑΤ		т А. А	ттс.	A C C T G	. c	т д		
HCP11	A C C A A A A G G A C A A G A A A C A C	ΑΑΤΤ.ΑΤ		ΤΑ.Α	тт.с.	ACC G		т д С	т.	
HCP28	AAAGGTAGGTTAATGGATACAAAAGTAG	AGCT A	т	ТА	. CA. AT	A C C T A	. CA T	. A		A
HCP49	C A G T C G A T G A T A G A T T G G A T A A A G A A A A	ТАСАТА		A G	Τ	A C C T A	Т	T.G. G.		
HCP41	C A G T C G A T G A T A G A T T A G A T A A A G A A A A	тасата -	c .	A T	Т	ACC - T - A	т	TG G	-	
HCP32		ТАСАТА		тт	т		т	те е		
HCP33		TGT CTTCI		Δ ΤΤ ΤΔG	000 T 400	- A A A A G - A A A		т с с т	6 - 6 (	ι 
HCP34		ΤΑΓΑΤΑ	A -	ат т	C A A			TC C T	-	C
UCD42		GTCA A	G CA-	т	. C		· · · · · · · · · · · · · · · · · · ·	та с		
HCR20		010A.A						T C C C		
HCF 23		CTCATTT.		<b>n n</b> T	· ^ · · · ^ · · · ·	ACC		TGG G • •		
HCF2		CAAA T	CA		. C			TGG		
HCF17				· · · · · · · · · · · · · · · · · · ·	. CA T			1 A G . C		
HCP20								AG	<del>.</del>	
HCP40					. I AI A.		· · · · · · · · · · · · · · · · · · ·	. A G	•	<b>U</b>
HCP38		A A A I I I			. I AI	ACCAA	Al		•	
HCP19	GCCCAACITITIGIIGACICGAAAAGAAT	A I A C A . A G - 🦷	••••••			I C C A A		. A G G	· · · <del>·</del> · · · ·	
HCP27	T T C C C A A A A A A A A A A A A A A A	GTACACAG	· • · · C · • · · ·		. T T	A C C C A		. A G	· · · <del>·</del> · · · ·	
HCP23		A I C C A A . G C -	•••••	A I	. I A I	ACC IA A		. C G	<del>.</del>	
HCP26	A C A C T T C A A T G A T C T A A C T T C A G G A A T A C	СТАСТ. АС	· · · · C · - · · ·	A	. T TA	A G C . C A	AA.	T G G A .	· · · <del>·</del> · · ·	
HCP7	T T C C C A T T T G A G A T A T T C T T C A A G A A T A T	СТСАА. А	C A		. T A	A C C A A	ΑΑ.	. A G A .	<del>.</del>	
HCP44	Т А Т Т С Т G Т Т А Т А G С А G С А С А А А Т Т А А G Т А	А G A C A Т .	A . <b>-</b>	A T	. C	A C C C A A		. A G	•	С
HCP5	C C G G G C A T T C C A A C A C T A A G A G G T C A G G G	А G А T C T C	C C - A	. Т А А	. T	A C C T A	GA.	. A G		
HCP25	Т Т С Т А Т А Т G С Т А G Т G А С А А Т Т А G А А А G Т G А	А А Т Т Т А А G	A <del>.</del>		. A	A C C		. G G		
HCP30	Т А Т Т G G G A G G T A A G T A A A T T T A A G A C A T A T	ТТА. АТТ	C A	. T. C A. T	. T A. A.	GAAACA	СТ Т.	T G C G	<b>-</b> . C .	
HCP3	АТТТ G T C A C A T T C C C C T C A A A A T C C A G C T G	GАGТССТБ		ст с	. C T	A C A A	Т. СТТ	T G . G . G	<del>.</del>	
HCP8	А С Т А Т А А А G С Т А Т G Т А А Т С А С А А С А А G G T G	СТАТТААБАС	GA CA	ТТА <b>Б А</b>	. CA	A . C A A	ТСТ	ΤΑ	Т	C.G.
HCP37	С Т С А Т Т Т А Т А Т С Т А А С А А С А С	GАСАСТ Т.	C A	A T . C .	. A A. C	A C C T		. A C T .	<del>.</del>	A .
HCP42	G G C A A T G A T A C A G T G G C T G A A A A C G T	GТСТСТ	Г СА А	А. Т С А	. C A C C A A .	GCTTCTA T. CA.	. C A - A T A .	T G . C . G C A .	•	ΑΑ
HCP35	T        A        C        T        A	Т G A A C C T G . 🤇	G A G . C A	GTTG. A. TGA	A. C A T C A C	A C A G . A C T C	ТАТС	G A C G	С. С-С. 1	СТСА
HCP18	Т G T T A T A A A A T A A A C A T A A A A A	АТССА. ТБАС	G C G A . T	. T A T. G G 🤅	3 T . A G C A	C T G A G T G G	БААС. ТТ ТС	AG.TGCTCG	C A G G C	C. GA.
HCP14	А Т Т G Т G А А А Т А А Т G Т А А А Т Т С А G C C T C T A C	ТТСАСАТ. Т.	A GT. CA	G A A C A G /	АА-АGСТ. Т	C T A A A T A C	A TT. G. G	TGGGTCTCT	ТАСС 1	. G. T
HCP9	T T C A C T T C A G T C A A A A T G A T G G A A T T C T	тттттстсс	ССС. СТТТ	. A A A T . T A A . /	ат - атататт -	•••• T T T T.	G T T T T A .	TGTTCC.TT.C	. GAGC	С

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.

# 

Z. Zhang, M. Gerstein / Gene xx (2003) xxx-xxx

1

728	727	726	725	774	722	721	720	719	718	717	716	715	714	712	711	710	709	708	707	706	705	704	703	702	701	700	699	8698	697	909 CK0	605	693	692	691	690	689	889	687	000	684	683	682	681	680	679	678	677	676	675	674



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 902 903 pseudogenes (48 if HCP46 and HCP47 are considered as one), which is considerably larger than the set of 11 904 pseudogenes previously analyzed. We wanted to test 905 whether the previous classification still held true. This 906 goal was best achieved by comparing sequences at the 907 908 informative codon positions where mutations had occurred 909 during recent evolution. Cyc is a highly conserved protein 910 among eukaryotes; the pair-wise amino acid sequence 911 identities range from 45% between mammals and yeast to 912 93.3% between chicken and mouse. An accelerated rate of 913 amino acid changes has occurred on the primate lineage 914 leading to the human ancestor, as there were amino acid 915 changes at nine positions since the split between Rattus 916 and Homo (Grossman et al., 2001). Among these positions 917 (11, 12, 15, 44, 46, 50, 58, 83, 89), none belongs to the 918 'conservatively substituted' category (Banci et al., 1999), 919 which suggests that they are probably not directly involved 920 in the electron-transfer process. Fig. 5 compares the human 921 pseudogenes and the functional genes from human, rodents 922 and chicken at these codon positions. For each position, 923 the sequences that have the same amino acid type as HCS 924 are shown in pink. Also, for eight of the nine positions, a 925 dominant amino acid type exists among the pseudogenes; 926 the positions that share this dominant amino acid type 927 are highlighted in light green. For position 44, both Val 928 and Ile are dominant amino acid types, so both are 929 highlighted.

930 It is obvious from the alignment that at all nine codon 931 positions, the majority of the human pseudogenes share an 932 amino acid type that is different with the HCS gene. Four 933 pseudogenes, HCP15, HCP21, HCP45 and HCP46, have 934 the highest sequence identity with HCS at these positions, 935 and they were selected and labeled as class 1 and the rest of 936 the pseudogenes were grouped into class 2. Note that we 937 used the same nomenclature as used by previous investi-938 gators (Evans and Scarpulla, 1988; Grossman et al., 2001). 939

A more detailed look at these codon positions follows. At 940 position 11, 31 of the 48 human pseudogenes have residue 941 Val and codon GTT or GTC; in contrast, the HCS gene and 942 three class 1 pseudogenes (HCP15, HCP21 and HCP45) 943 have residue Ile and codon ATT. Interestingly, as in most of 944 the class 2 pseudogenes, the somatic rodent and chicken 945 genes also have Val and GTT/GTC at position 11. The same 946 pattern also occurs at positions 12, 15, 46, 50, 58 and 83, 947

where the majority of the class 2 pseudogenes share the 953 same amino acid type with rodent somatic genes, and the 954 class 1 pseudogenes share a different amino acid type with 955 the HCS gene. At position 44, there is no predominant 956 amino acid type among the pseudogenes, as Ile occurs 14 957 times and Val occurs 12 times; however, the HCS gene and 958 three class 1 pseudogenes (HCP15, HCP2 and HCP46) have 959 Pro at the position. This particular position has obviously 960 961 gone through very rapid changes in recent evolution, since 962 rodent somatic cyc genes have residue Ala at the position, which occurred only six times among the pseudogenes. At 963 964 position 89, the predominant amino acid among the pseudo-965 genes is Ala (occurring 24 times), which is different from 966 both human and rodent somatic genes: HCS and all class 1 967 pseudogenes have Glu and the rodent somatic genes have 968 Glv.

9

969 The sequence comparison shown in Fig. 5 strongly sup-970 ports the notion that the human cyc pseudogenes originated 971 from a functional gene that had undergone significant 972 changes during the mammalian evolution. The four pseudo-973 genes in class 1 appear to be from a gene that is identical to 974 the modern HCS gene, while the class 2 pseudogenes are 975 much older and have more resemblance to rodent somatic 976 genes. Although we divided the pseudogenes into two 977 classes, it is important to note that gene evolution was a 978 gradual process, and our classification in no way implies any 979 dramatic changes in the biochemical function and gene 980 structure. Our classification is in very good agreement with 981 the phylogenetic analysis, as the four class 1 pseudogenes 982 were found in a separate branch together with the HCS gene 983 at the top of the tree (Fig. 4). Furthermore, as shown in 984 Fig. 3, the 5' UTR sequences of HCP15, HCP21 and HCP45 985 also have the fewest number of substitutions compared with 986 HCS mRNA sequences. Given the conclusion that the four 987 class 1 pseudogenes and the modern HCS gene share the 988 same origin, it is possible to actually date these pseudogenes 989 based on their sequence divergence. Using the formula 990 T = D/(k), where D is the divergence and k is the mutation 991 rate per year per site, the ages for the pseudogenes were 992 estimated to be  $27 \pm 8$  Myr for *HCP15*,  $23 \pm 7$  Myr for 993 HCP21,  $34 \pm 9$  Myr for HCP45 and  $31 \pm 9$  Myr for 994 *HCP46.* A mutation rate of  $1.5 \times 10^{-9}$  per site per year 995 for pseudogenes was used (Li, 1997). In comparison, the 996 997 divergence time of human from gibbons is believed to be 998 20 to 30 Myr ago (Lander et al., 2001). The much lower 999 number of pseudogenes in class 1 compared with the 1000 number in class 2 is consistent with the observed decline of 1001 retrotransposition activity during the last 40 Myr in the 1002 human genome (Lander et al., 2001). 1003

948 949 950

951

952

1004 1005

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.

Z. Zhang, M. Gerstein / Gene xx (2003) xxx-xxx

15 44 46 50 58 83 89 11 12 1009 1010 M S P Y A I V E ATG TCC CCT TAC GCC ATC GTC GAA ATT HCS 1011 I M S P H A I V E ATT ATG TCC CCT CAC GCC ATC GTC GAA I M S P Y A I V E (HS11) HCP15 1012 ATT ATG TCC CCT TAC GCC ATC GTC GAA (HS7) HCP21 1013 I M S A Y V T V E ATT ATG TCC GCT TAC GTT ACC GTC GAA HCP45 1014 Class I TCC CCT TAT GCC ATT GTC GAA HCP46 1015 Class 2 V Q A V F D I A A GTT CAG GCC GTT TTC GAT ATC GCC GCA 1016 HCP1 1017 (HC6) GTT CAG GCC ACT TTC GAT ACC GCC HCP2 F R S TTC AGA TCA 1018 CT TTC GAC ATC ACA ACA НСР3 AT GAC ACC GTC GCA GCC ACT 1019 HCP4 TAG GTC GCT TTC GAC ATC ACT GCA 1020 HCP5 GTT GTT CAG GCC ATT TTC GAT ACC GCT GCA 1021 HCP6 V Q A V F D T A A GTT CAG GCC GTT TTC GAC ACT GCC GCA 1022 HCP7 F T T I S D T T V TTC ACA ACC ATT TCC GAC ACC ACT GTA HCP8 1023 ATT CAG GCT CCA TTT GAG GTA TCT AGT 1024 HCP9 V Q A V F D T A A GTT CAG GCC GTT TTC GAC ACC GCC GCA 1025 HCP10 V Q A T F D T A A GTT CAG GCC ACT TTC GAC ACC GCC GCA 1026 HCP11 V Q A I F D T A A GTT CAG GCC ATT TTC GAC ACC GCC GCA HCP12 1027 GTT CAG GCC ATT TTC GAC ACC GCC GCA HCP13 1028 TTT CTT GTC CCT TTT GAT ATT ATT HCP14 1029 GTT CAG GCC CGT CCC HCP16 TAG 1030 D ATT CAG GTC ATT TTC GAT ACC TTT GCA HCP17 1031 GTC AG GCC CCC TTC GTT TCT HCP18 1032 ATT CÃG GCC GTT TTA GAT ACC GCC HCP19 1033 V Q A I F D T G G GTT CAG GCC ATT TTC GAC ACC GGC GGA HCP20 1034 GTT CAG GTC GTT TTC GAT ACC GCC GTA HCP22 1035 V Q A I F E S A A GTT CAG GCC ATT TTC GAG TCC GCA GCA HCP23 1036 V Q A V F D T A A GTT CAG GCC GTT TTC GAC ACC GCT GCA (HC3) HCP24 1037 GTT CAA GCC ATT TTC GAT ACC GCC GCA (HC10) HCP25 1038 V Q A V F D T A A GTT CAG GCC GTT TTC GAC ACC GCT GCA HCP26 1039 ACT GAG GCC ATT TTC GAT ACC GCC HCP27 1040 HCP28 1041 HCP29 1042 HCP30 1043 V Q A V F D T A A GTT CAG GCC GTT TTC GAC ACC GCT GCA (HC3) HCP31 1044 V Q A I L E T A A GTT CAG GCC ATT TTA GAG ACC GCC GCA HCP32 1045 V Q A I L D T S A GTT CAG GCC ATT TTA GAC ACC TCC GCA (HC4) HCP33 V Q A I F D T A A GTT CAG GCC ATT TTC GAT ACC GCC GCA 1046 HCP34 1047 AÃA AÃA ACC GCT TCC GÃG ACC GCT TCA (HC8)HCP35 V Q A A F D T A T GTT CAG GCC GCT TTC GAT ACG GCC ACA 1048 (HC5) HCP36 1049 GCT CAG GCC GTT TTC GAC GCC GCC GTA HCP37 GTC TAG GCC TTT TTC GAC ACC TAT 1050 (HC2) HCP38 1051 GTT CAG GCC ATT TTC GAC ACC GCC GCA (HC1) HCP39 GTT CCA GCC GCT TTC GAT ACC ACT 1052 (HC7) HCP40 V Q A I L E T GTT CAG GCC ATT TTA GAG ACC 1053 HCP41 V Q S P F D A A GTT CAG AGC CCT TTC GAT GCC GCT HCP42 1054 A Q A V F D N T A GCT CAA GCC GTT TTC GAT AAC ACT GCA HCP43 1055 V Q A V F D T S T GTT CAG GCC GTT TTC GAT ACC TCC ACA HCP44 1056 A T / D T A A GCC ACT AT GAC ACC GCC GCA HCP48 1057 L Q A I L E T A A CTT CAG GCC ATT TTA GAG ACC GCC GCA HCP49 1058 V Q A A F D T A GTT CAG GCC GCT TTC GAT ACC GCT 1059 CYCS MOUSE GGA GTT CAA GCC GCT F D T A G TTC GAT ACC GCT GGA 1060 CYCS RAT 1061 CYCT MOUSE 1062 CYCT RAT 1063 CHICKEN 1064 FLY DC4

3.4. HCP9 resembles rat testis-specific cyc gene

As shown in Fig. 4, pseudogene HCP9 appears to be very 1067 old, as it was placed near the root of the tree. This pseudo-1068 gene also has one of the largest sequence divergences from 1069 the modern HCS gene at  $0.473 \pm 0.071$  per site per Myr 1070 1071 (Table 1). Furthermore, it is disrupted into three fragments by two DNA insertions, both containing many retrotrans-1072 posons. As discussed earlier, it is difficult to calculate the 1073 1074 age of the pseudogenes based on sequence divergence; 1075 however, in this case we could actually deduce a lower boundary for the age of HCP9 by estimating the age of 1076 1077 the retrotransposons contained in the inserted sequences. 1078 Using the RepeatMasker program (Smit, AFA & Green, P, 1079 URL:http://repeatmasker.genome.washington.edu/), several 1080 LTR sequences of MalR and ERVL types and several Alu 1081 sequences of AluJo and AluJb types were identified. It has 1082 been estimated that in the human genome, LTR/MalR and 1083 LTR/ERVL species had died out about 40 Myr ago (Smit, 1084 1993; Cordonnier et al., 1995). The AluJo and AluJb 1085 sequences were ancient Alu species that were last active at 1086 around 81 Myr ago (Mighell et al., 1997; Smit, 1999). These 1087 facts indicated that HCP9 was inserted into the genome at 1088 least 80 Myr ago, which was before the divergence between 1089 human and prosimians (55-80 Myr) and after the estimated 1090 time for eutherian mammalian radiation ( $\sim 100 \text{ Myr}$ ) 1091 (Lander et al., 2001). This particular pseudogene must 1092 have been inherited from a mammalian ancestor long before 1093 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 genes (*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

1112 Fig. 5. Sequence alignment at nine codon positions of the human cyc 1113 pseudogenes and the functional cyc genes from human, rodents and 1114 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 1115 same amino acid with the majority of the human pseudogenes are shown in 1116 green. The pseudogenes were divided into two classes based on their 1117 sequence identity with the HCS gene. A blank at a codon position indicates 1118 a missing sequence caused by truncation, and dashes '-' indicate gaps 1119 caused by DNA deletion. Frame shifts and stop codons are indicated by '/', '\' and 'X'. (\*HCP47 sequence is merged into HCP46). 1120

V Q A A Y D T A E GTG CAG GCC GCG TAC GAT ACC GCA GAG

FLY DC3

1065 1066

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

Z. Zhang, M. Gerstein / Gene xx (2003) xxx-xxx



1137 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. 1138

1140 mouse and possibly in bull and rabbit (Kim and Nolla, 1141 1986), but not in human or other primates. It is likely that a 1142 functional cyc gene similar to the modern rodent testis-1143 specific gene existed in the genome of an ancient mammal-1144 ian ancestor. While modern rodents have kept the functional 1145 gene, humans only retained the pseudogene and lost the 1146 functional copy. It has been reported that none of the rodent 1147 cyc pseudogenes discovered so far originated from the 1148 testis-specific genes (Wu et al., 1986), however, all of these 1149 pseudogenes were discovered by genomic hybridization 1150 experiments instead of by computationally scanning the 1151 genome. As with the human genome, we expect many more 1152 cyc pseudogenes, and possibly testis-specific cyc pseudo-1153 genes, to be discovered in the mouse after the complete 1154 mouse genome sequence becomes available. 1155

#### 1156 3.5. Online database 1157

1139

1158

1159

1160

1161

1162

1163

1164

1165

1167

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 1166

The 49 cyc pseudogenes we describe here present an 1168 evolutionary record of the human cytochrome c gene; our 1169 findings strongly support the hypothesis that this gene has 1170 evolved at a very rapid rate in the recent human lineage. The 1171 sequence information we report here will not only aid 1172 researchers to design better HCS-specific probes to avoid 1173 pseudogene complications, but will also be very useful in 1174 1175 calibrating and estimating various evolutionary and phylo-1176 genetic 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.

1199 Traditionally, most of the pseudogenes reported in 1200 literature were discovered by screening a genomic library 1201 using DNA hybridization techniques. As has been demon-1202 strated in this study and other reports, such experiments 1203 often overlook the bulk of the pseudogene population. The 1204 discovery of such a great number of cytochrome c pseudo-1205 genes also raises the question as to the total number of 1206 pseudogenes in the human genome; such an estimate is 1207 important in the accurate prediction and annotation of 1208 functional genes. Differentiation between functional genes 1209 and disabled pseudogenes in genome annotation has proven 1210 to be a challenging and difficult task. For instance, it was 1211 suggested that in the Caenorhabditis elegans genome a fifth 1212 of annotated genes could be pseudogenes (Mounsey et al., 1213 2002). With the advent of the complete human genome 1214 sequence, a systematic and comprehensive survey of 1215 pseudogenes is much needed, not only to provide better 1216 functional gene annotation, but also to extend our under-1217 standing of the evolution of genes and genomes as a whole. 1218 We also did a preliminary survey in the recently published 1219 mouse draft genome sequence (Waterston et al., 2002) and 1220 detected about 40 cytochrome c processed pseudogenes. 1221 However, the relative low quality of the mouse sequence did 1222 not allow for detailed comparison between these two sets of 1223 pseudogenes. 1224

#### Acknowledgements

MG acknowledges NIH grant 2P01GM54160-04. Z.Z. 1229 thanks Dr. Paul Harrison for comments on the manuscript 1230 and Dr. Duncan Milburn and Nat Echols for computational 1231 help. 1232

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1226 1227 1228

12

Z. Zhang, M. Gerstein / Gene xx (2003) xxx-xxx

#### 1233 **References**

1234

1280

1281

1282

1283

1284 1285

1286

1287

1288

- 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.
- 1248 Chothia, C., Lesk, A.M., 1985. Helix movements and the reconstruction of 1249 the haem pocket during the evolution of the cytochrome c family.
  1250 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. retrotransposons 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.
- 1257 Grossman, L.I., Schmidt, T.R., Wildman, D.E., Goodman, M., 2001.
   1258 Molecular evolution of aerobic energy metabolism in primates. Mol.
   1259 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.
- 1263 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.
- 1270 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.
- 1275 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, 1289 MA. 1290
- 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 1293 Lett. 417, 1–5. 1294
- Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F., 2000. 1295 Vertebrate pseudogenes. FEBS Lett. 468, 109–114. 1296
- 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.
  1300
- 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 1305 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. 1000. Interpreted repeats and other mammates of transposols.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable
 elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.
 1316

- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W:1317improving the sensitivity of progressive multiple sequence alignment<br/>through sequence weighting, position-specific gap penalties and weight<br/>matrix choice. Nucleic Acids Res. 22, 4673–4680.13181320
- 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. 1322
  Chem. 263, 6791–6796. 1325
- 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.
   1329
- 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.

1301

1302

1303

1304

1307

1308

1309

1310

1311

1312

1313

1321

1326

1327

1339

- 1340
- 1341
- 1342

1343