J Mol Evol (1985) 22:309–315

Journal of
Molecular Evolution

Comparison of Human and Chimpanzee ζ1 Globin Genes

Cary Willard,[1] Elsie Wong,[1] John F. Hess,[2] Che-Kun James Shen,[2] Barbara Chapman,[3]* Allan C. Wilson,[3] and Carl W. Schmid[1]

[1] Department of Chemistry, University of California, Davis, California 95616, USA
[2] Department of Genetics, University of California, Davis, California 95616, USA
[3] Department of Biochemistry, University of California, Berkeley, California 94720, USA

Summary. The DNA base sequences of the entire chimpanzee ζ1 globin gene and an additional 1 kb of DNA flanking both the human and chimpanzee genes have been determined. Whereas the human ζ1 gene contains a termination codon in the sixth position, the chimpanzee gene appears to be functional. This finding confirms Proudfoot et al.'s suggestion that the human ζ1 gene was recently inactivated. Like the corresponding human ζ1 and ζ2 genes, the first and second introns of the chimpanzee ζ1 gene are occupied largely by tandem repeats of short oligonucleotides. These tandem repeats have undergone several rearrangements since the divergence of the human and chimpanzee ζ1 genes.

Key words: Globin genes — Primate evolution — Tandem repeats

Introduction

Zeta hemoglobin is an alpha-like hemoglobin expressed in embryos. Two closely linked genes in humans, designated ζ2 and ζ1, have sequences corresponding to that of the zeta protein (Proudfoot et al. 1982). Mapping of the alpha-like globin gene cluster in apes demonstrates that the duplication of the zeta globin genes is ancient and at least predates the divergence of human and chimpanzee (Zimmer et al. 1980; B. Chapman, unpublished data). There are several interesting sequence differences between the duplicate human zeta globin genes; most notably, ζ1 appears to have recently been inactivated (Proudfoot et al. 1982). The coding regions of the two genes differ by only three point mutations. Two of these mutations result in amino acid replacements, and one mutation converts the sixth codon from that for Glu in the ζ2 to a termination codon in the ζ1 gene (Fig. 1). The absence of third-position synonymous mutations in protein-coding regions suggests that this pair of genes has undergone a very recent gene conversion event (Proudfoot et al. 1982). The inactivation of the human ζ1 gene probably occurred even more recently. Both genes are transcribed in *Xenopus* oocytes (Proudfoot et al. 1984).

The unusual introns of this pair of genes also exhibit interesting differences (Proudfoot et al. 1982). Each gene, as is typical of other globin genes, contains two introns. The first intron of the ζ2 gene is 886 bp in length, whereas the corresponding intron of the ζ1 gene is 1262 bp in length. The difference in length is due entirely to a difference in the number of copies of a 14-bp-long element (Fig. 1). This element is present at 12 tandem copies in the ζ2 gene and as 39 tandem copies in the ζ1 gene. In addition to this tandemly repeated element, the first intron of both ζ genes also contains 650 bp of unique DNA. There are only three base differences in the nonrepeated regions of the first introns of the ζ2 and ζ1 genes, indicating that the very recent conversion event described above extended also into the first intron. The second intron of the ζ genes is largely occupied by a tandem repetition of a 5-bp element (Fig. 1). There are 35 tandem copies of this 5-bp element in the second intron of the ζ2 gene, and 52 tandem copies of it in the second intron of the ζ1 gene.

Offprint requests to: C.W. Schmid
* *Present address:* Chiron Research Laboratories, Emeryville, California 94608, USA

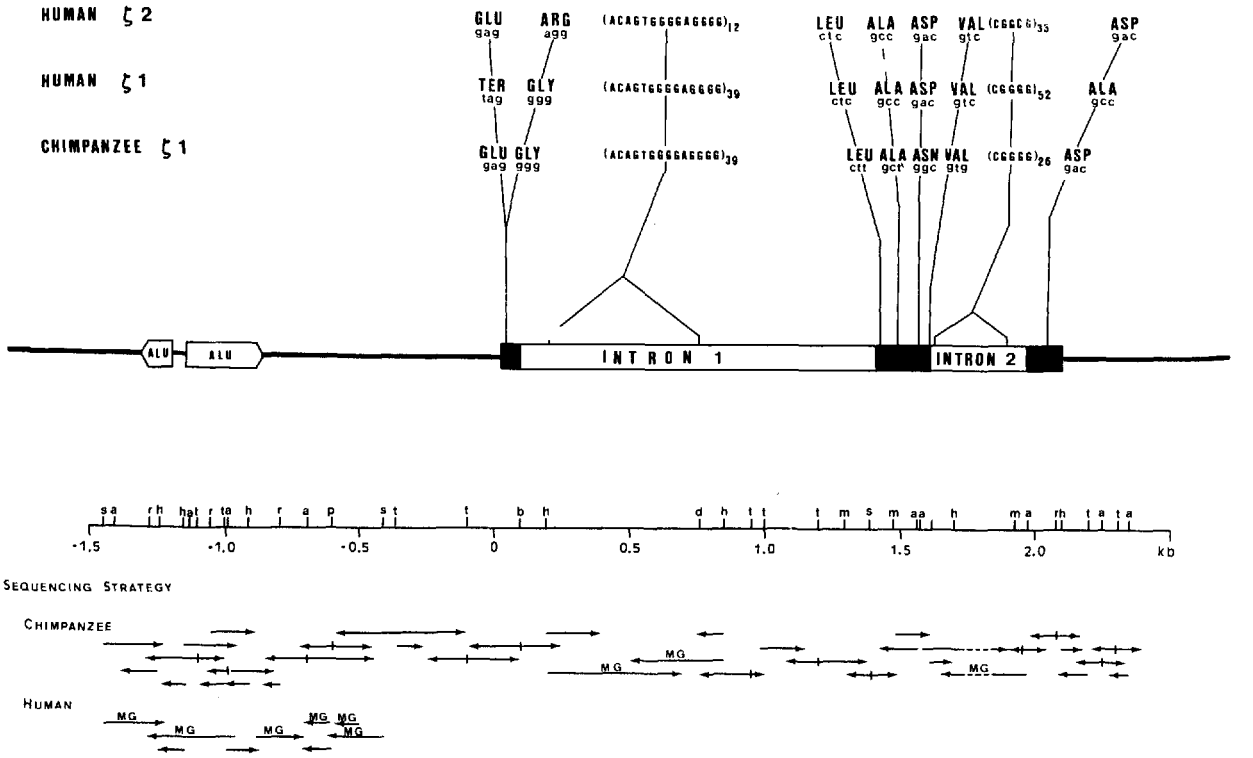


Fig. 1. Structures of ζ globin genes and their evolution. The strategy used to determine the sequence of the chimpanzee and human ζ1 globin genes is indicated by the arrows, which correspond to restriction fragments used to construct M13 subclones for dideoxy sequencing. "MG" stands for Maxam–Gilbert sequencing. Position 0 is the transcription start site. The numbering system corresponds to that for the human ζ1 globin gene (Proudfoot et al. 1982). The structure of the human/chimpanzee ζ1 gene, including two flanking Alu repeats, is schematically depicted. Codon differences between the chimpanzee ζ1, human ζ1, and human ζ2 genes as well as the lengths of tandem repeats within the introns are qualitatively indicated. The restriction sites are as follows: s, Sst I; a, Alu I; r, Rsa I; h, Hae III; t, Sau 3A; p, Hpa II; b, BG1 II; d, Dde I; m, Sma I

To obtain additional insight into the evolution and inactivation of the human ζ1 gene, we have sequenced the chimpanzee ζ1 gene. The results of this study confirm Proudfoot et al.'s suggestion that the human ζ1 gene was recently inactivated and provide additional examples of changes in the unusual introns of the ζ globin genes. In a companion paper, we report sequences of other parts of the human and chimpanzee alpha-like globin gene cluster (Sawada et al. 1985).

Materials and Methods

Restriction enzymes, T4 polynucleotide kinase, T4 polynucleotide ligase, *Escherichia coli* polymerase large (Klenow) fragment, bovine alkaline phosphatase, deoxynucleotides, and dideoxynucleotides were purchased from Bethesda Research Labs, P-L Biochemicals, or New England Biolabs. Radioactive nucleotides were purchased from Amersham or ICN. The human ζ1 globin gene was subcloned into pBR ζ from a lambda genomic clone as described by Lauer et al. (1980). The chimpanzee ζ1 globin gene was isolated as a lambda genomic clone as described by Sawada et al. (1983).

The human and chimpanzee lambda clones were digested with the restriction enzyme Sst 1. The 1.06-kb human fragment, 1.06-kb chimpanzee fragment, 1.80-kb chimpanzee fragment, and 1.2-kb chimpanzee fragment were cloned into bacteriophage M13 or plasmid pUC (Messing 1983). These subclones were further digested and recloned into appropriate sites in bacteriophage M13 strains 8, 9, 10, and 11 as indicated in Fig. 1.

The unusual sequences present in the introns proved to be refractory to the M13 method. Sequences of these regions were determined by the method of Maxam and Gilbert (1977). Reaction times were modified to enable sequencing of the two G-rich introns. All other sequences were determined using the M13 dideoxy method of Sanger et al. (1977) and Messing et al. (1981). Compression effects in the region of positions 2000 and 2100 required the use of one or the other of the base analogues inosine triphosphate and 7-deazaguanosine triphosphate. Although we cannot preclude single-base sequence errors, all human–chimpanzee sequence differences in the gene, recognizable control regions, and the Alu members were individually scrutinized. The sequence of the 5′ flanking region, positions −900 to −200, is presented for completeness, but without the accuracy required for an analysis of the divergence of human and chimpanzee DNAs.

Results

Sequence Determination and Overall Gene Structure

The organizations and restriction maps of the alpha-like gene clusters in human and chimpanzee are very similar (Lauer et al. 1980; Zimmer et al. 1980). In particular, the human ζ1 gene maps about 4 kb 5′

to the human $\psi\alpha$ gene. We identified the chimpanzee $\zeta 1$ gene by its position relative to the chimpanzee $\psi\alpha$ gene (Sawada et al. 1983). We do not show the supporting mapping data here, as the sequences flanking the human $\zeta 2$ and $\zeta 1$ genes are entirely different (Proudfoot et al. 1982). The base sequence of the chimpanzee gene reported here unambiguously identifies it as the $\zeta 1$ gene.

The DNA sequences of the chimpanzee $\zeta 1$ gene and additional DNA flanking the chimpanzee and human genes were determined by the strategy depicted in Fig. 1. The general organization of the chimpanzee $\zeta 1$ gene is the same as that of the human $\zeta 1$ gene (Proudfoot et al. 1982): It consists of three exons and two intervening sequences. In addition, a full length Alu family member and a partial Alu family repeat are present in the 5′ flanking regions of both the chimpanzee and human genes (Fig. 1). A detailed analysis of Alu family repeats is the topic of an accompanying article (Sawada et al. 1985); consequently, we note here only that corresponding Alu repeats are found at identical positions in the human and chimpanzee genes (Fig. 2). This paper analyzes those regions of the sequence that affect the expression of the gene. For this analysis we compare the sequence of the chimpanzee $\zeta 1$ gene with Proudfoot et al.'s sequences of the nonfunctional human $\zeta 1$ gene and the closely related functional human $\zeta 2$ gene (Figs. 1 and 2).

Coding Regions

The globin-coding regions of these three genes differ from each other by a small number of point mutations (Figs. 1 and 2, Table 1). The human $\zeta 1$ gene has been inactivated by a single point mutation in codon 6 that changes the Glu codon of human $\zeta 2$ to a termination codon in human $\zeta 1$ (Fig. 1, Table 1). Codon 6 of the chimpanzee $\zeta 1$ gene codes for Glu, as it does in the functional human gene. This is direct phylogenetic evidence that the inactivation of the human $\zeta 1$ gene occurred after the divergence between human and chimpanzee, in agreement with Proudfoot et al.'s conclusion that the $\zeta 1$ gene was inactivated very recently in the human lineage.

The chimpanzee gene has three synonymous mutations relative to the two human genes (Table 1). There are no synonymous mutations between the two recently converted human genes (Proudfoot et al. 1982). This small number of synonymous mutations is consistent with the known divergence time of human and chimpanzee and the rate of third-position substitutions. Each gene has one replacement mutation relative to the other two genes (Fig. 1): Codon 7 of the human $\zeta 2$ gene is different from codon 7 of the chimpanzee and human $\zeta 1$ genes, codon 75 of the chimpanzee gene is different from that of either human gene, and codon 126 of the human $\zeta 1$ gene differs from codon 126 of the human $\zeta 2$ and chimpanzee $\zeta 1$ genes (Table 1). The rate of replacement mutation in selected globin genes is estimated to be 0.1% per million years [reviewed by Lacy and Maniatis (1980)]. The low level of replacement mutation observed in the chimpanzee gene as compared with the two human genes implies that the chimpanzee and both human genes have evolved under selective pressure. By all other sequence criteria the chimpanzee gene appears to be functional (see below).

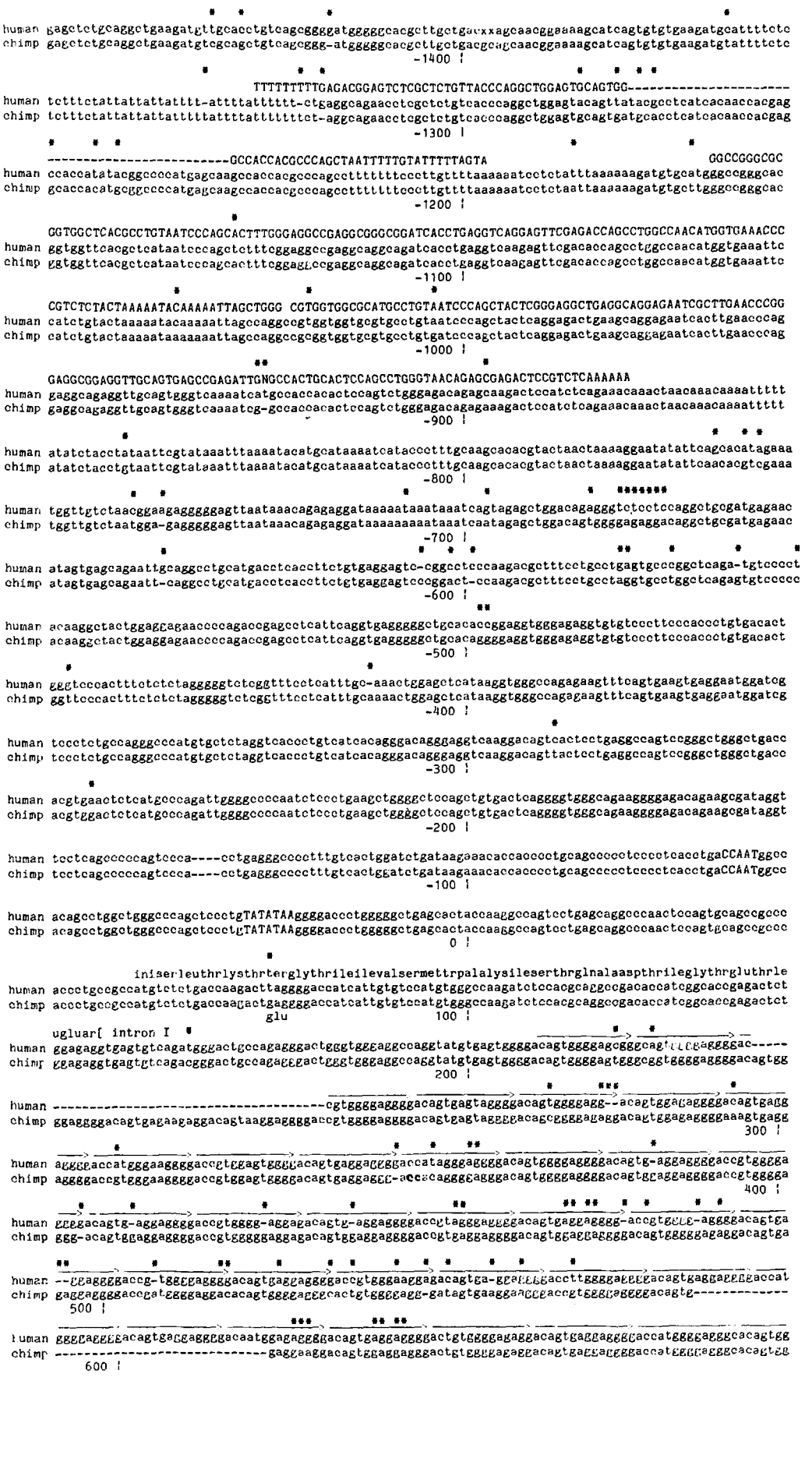
Table 1. Differences in the exons of the human $\zeta 2$ (H$\zeta 2$), human $\zeta 1$ (H$\zeta 1$), and chimpanzee $\zeta 1$ (C$\zeta 1$) genes [see Fig. 2 and Proudfoot et al. (1982)]

Amino acid position	Codon	Base (X)			Amino acid		
		H$\zeta 2$	H$\zeta 1$	C$\zeta 1$	H$\zeta 2$	H$\zeta 1$	C$\zeta 1$
6	XAG	G	T	G	Glu	Ter	Glu
7	XGG	A	G	G	Arg	Gly	Gly
32	CTX	C	C	T	Leu	Leu	Leu
64	GCX	C	C	T	Ala	Ala	Ala
75	XAC	G	G	A	Asp	Asp	Asn
96	GTX	C	C	G	Val	Val	Val
126	GXC	A	C	A	Asp	Ala	Asp

Intervening Regions

An unusual feature of the intervening regions of the zeta globin genes is the presence of long arrays of tandemly repeated oligonucleotide sequences. The first intron of all three genes consists of a long unique sequence of about 650 bp (Figs. 1 and 2). This unique region is preceded by a long tandem array of a 14-nucleotide-long sequence. The consensus sequence of this 14-nucleotide-long element is the same in the chimpanzee and human genes: ACAGTGGGGAGGGG. The length of the tandem array is 12 copies in the human $\zeta 2$ gene, 39 copies in the human $\zeta 1$ gene, and 39 copies in the chimpanzee $\zeta 1$ gene.

Although the lengths of these tandem repeats are identical in chimpanzee and human, they are not necessarily identical sequences. There is 18% divergence between the human and chimpanzee tandem repeats when they are simply aligned in phase. The divergence can be reduced to 11% by allowing for a deletion and corresponding insertion of three repeat units between the human and chimpanzee tandem arrays (Fig. 2). The divergence could be further reduced by additional rearrangement of these repeat units. This implies that the length of this tandem repeat has not been static since divergence between human and chimpanzee, and that there may have been numerous expansions, contractions, or


human
chimp
-1400
-1300
-1200
-1100
-1000
-900
-800
-700
-600
-500
-400
-300
-200
-100
0
100
200
300
400
500
600
ugluar[intron I

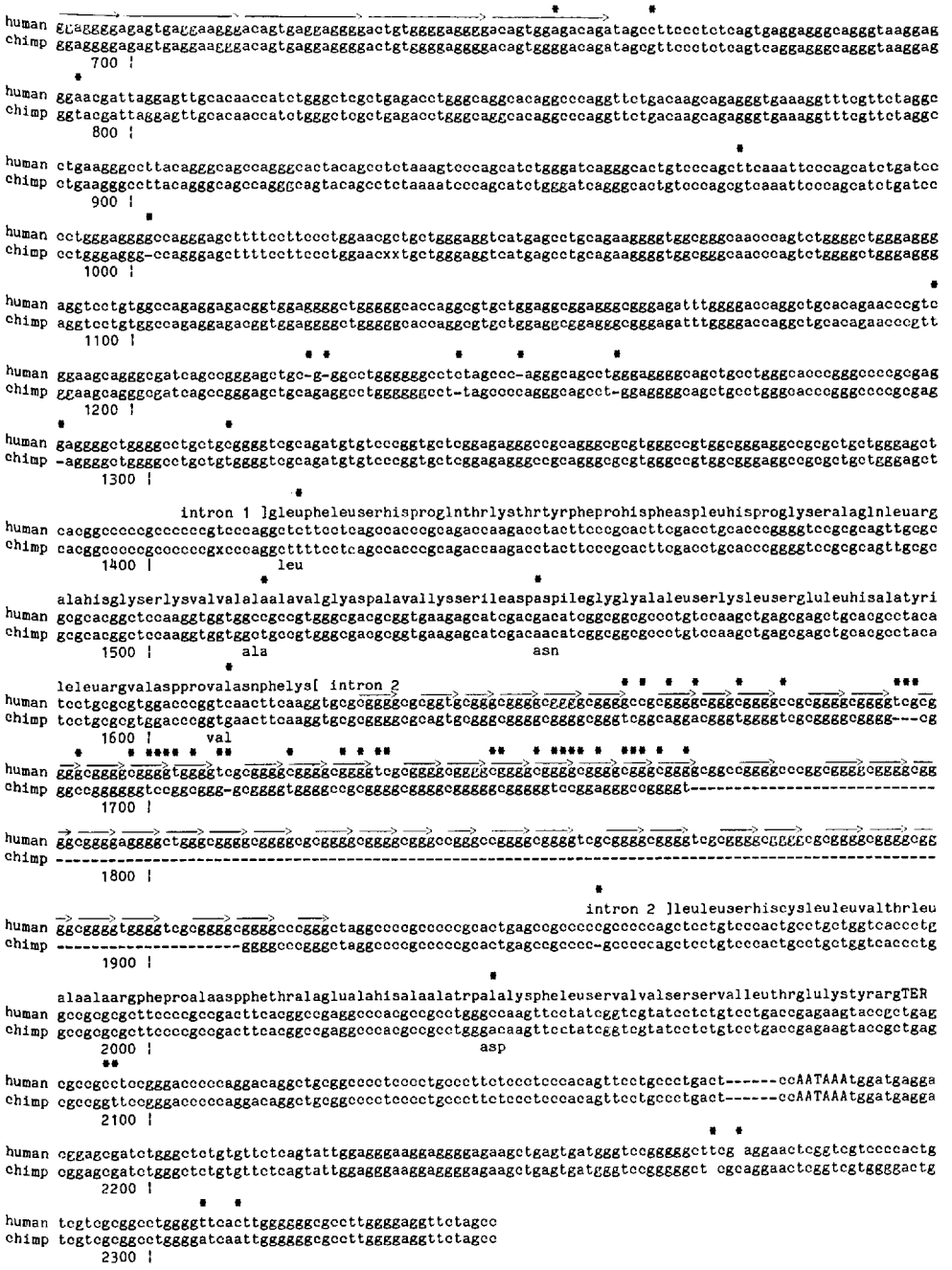


Fig. 2. DNA sequences of the human and chimpanzee ζ1 genes, compared according to the numbering system of Proudfoot et al. (1982). Differences are marked by asterisks (*). The full-length and partial Alu repeats (ca. bases −900 to −1300) are compared with the genomic consensus, indicated by superimposed capital letters (Schmid and Shen 1985). Promoter and polyadenylation signals (ca. positions −50 and 2180) are highlighted with capital letters. Intron–exon boundaries are indicated with brackets and the polypeptides corresponding to exons are shown. The 14-base and 5-base tandem repeat elements in the first intron (ca. intron positions 200–1000) and second intron (ca. positions 1500–1900) are highlighted with arrows

other rearrangements in this tandem array. Jeffreys et al. (1985) found that "minisatellite" regions such as the tandem repeats in this intron are hypervariable in human DNA.

There are 12 differences between the 650-bp-long unique sequences present in the first introns of the human and chimpanzee ζ1 genes (Fig. 2). This value approximates the expected divergence (1.5–2%) for nonselected human and chimpanzee DNAs (Zimmer et al. 1980; Chang and Slightom 1984; Sibley and Ahlquist 1984). For comparison, there are two differences in this same region between the human ζ2 and ζ1 genes (Proudfoot et al. 1982). This reemphasizes Proudfoot et al.'s (1982) conclusion that the gene conversion of the human ζ2 and ζ1 genes was an extremely recent event, one that certainly occurred after the divergence between human and chimpanzee.

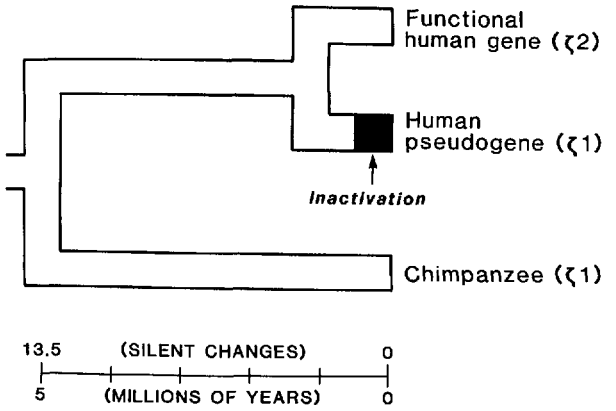


Fig. 3. Tree and approximate time scale relating sequences of three zeta globin genes. In the single-copy regions compared, the human ζ1 and ζ2 genes differ by three silent substitutions, whereas the chimpanzee ζ1 gene differs from them by an average of 13.5 silent substitutions. Based on the assumption that the chimpanzee and human lineages diverged 5million years ago, one estimates that the human ζ1 and ζ2 genes diverged from each other roughly 1.1 million years ago

Most of the second intron of the zeta globin genes consists of an imperfect tandem repetition of a pentanucleotide, CGGGG. The length of this tandem repeat can be quite variable. There are 26 tandem copies of the pentanucleotide in the chimpanzee ζ1 gene, 52 copies the human ζ1 gene, and 35 copies in the human ζ2 gene. It is not possible by simple alignment of the chimpanzee and human intron 2 sequences to account for the shortening of the chimpanzee intron by a single or small number of events. This difference in length may thus be the result of multiple steps, suggesting that the size of the intron may have expanded and contracted many times since the divergence between human and chimpanzee.

5′ and 3′ Flanking Regions

The 5′ flanking sequences of the two human zeta globin genes and the chimpanzee ζ1 gene are nearly identical for ~200 bp upstream from the initiation codon. Included within this flanking region are recognizable promoter elements (the CCAAT and TATATAA boxes), mRNA cap sites, and the initiation codon (Fig. 2). The high level of sequence conservation between the human and chimpanzee ζ1 genes persists for about 500 bp upstream from the cap site (ca. position −500). The 3′ flanking region of the ζ1 gene is also conserved, showing 96.6% homology between human and chimpanzee for 240 bp following the termination codon. This region includes the polyadenylation signal at about position 2170 and the polyadenylation site (Fig. 2). These homologous sequences in the flanking regions of the chimpanzee and human ζ1 genes are not included in the conversion unit of the human ζ2 and ζ1 genes (Proudfoot et al. 1982).

Discussion

The introns of the zeta globin genes are unusually long compared with those of other globin genes. They are also unusual in that their lengths are so plastic. Blot hybridization studies indicate that the length of zeta globin introns is very polymorphic within the human population (Chapman et al. 1985). This length variability is due entirely to copy-number variation in the tandem arrays contained within the introns. The single-copy regions of the introns are well conserved when one compares either human with chimpanzee genes or the two human genes with each other. We believe that length variability is an inherent biochemical property of any tandemly arranged simple sequence and that the zeta globin introns are not unusual in this regard (Bell et al. 1982; Heilig et al. 1982; Kominami et al. 1983a,b; Maroteaux et al. 1983; Sawada et al. 1983, 1985; Jeffreys et al. 1985). We therefore expected to find such differences in the lengths of the chimpanzee and human ζ1 introns as were found in the second intron. Our present finding that the length of the first intron is unchanged since the divergence between human and chimpanzee is surprising. We do not know if this intron has any biological function. The finding that the first intron of an immunoglobulin gene has an enhancer activity (Banerji et al. 1983; Gillies et al. 1983) demonstrates one possible function of introns.

Proudfoot et al. (1982) reported that the human ζ1 gene is a very recently inactivated pseudogene. The data reported here for the chimpanzee ζ1 gene confirm and extend their findings. The synonymous mutations in the chimpanzee gene relative to both human genes demonstrate that the conversion of the two human genes must have followed the divergence between human and chimpanzee. Based on comparison of the number of silent mutations between the two human genes and the chimpanzee gene, we conclude that the divergence between the two human genes occurred approximately one million years ago (Fig. 3).

The significance of pseudogenes is unknown. As suggested by Proudfoot and Maniatis (1980) and Lacy and Maniatis (1980) pseudogenes may be a natural consequence of the existence of multigene families. The inactivation of one of a pair of duplicate genes would not normally be lethal. The thus inactivated pseudogene could subsequently either be rescued by correction or evolve as nonselected DNA.

There is a complete spectrum of known pseudoglobin genes that illustrates this progression. The pseudo α globin gene is an ancient pseudogene that has been multiply inactivated (Proudfoot and Maniatis 1980). Mutations are so widely distributed

throughout the sequence of the pseudo α globin gene that its correction would require the replacement of the entire sequence. For all practical purposes the pseudo α gene can be regarded as irreparable. The pseudo β globin gene is an ancient pseudogene in a similar state of advanced decomposition (Chang and Slightom 1984; Harris et al. 1984). δ-Hemoglobin is a minor β-like hemoglobin in humans. However, in Old World monkeys (e.g., rhesus and colobus), where the δ gene became inactive relatively recently, the number of mutations inactivating the gene's expression is moderate (Martin et al. 1983). The δ globin gene is thought to result from correction of a preexisting pseudo β globin gene, in which case the corrected gene was subsequently maintained in the human lineage but reinactivated in monkeys.

Relative to the time of inactivation of the pseudo δ globin gene in monkeys, the human ζ1 globin gene was inactivated very recently (Fig. 3). This gene was converted by the functional ζ2 gene sequence in recent times, and presumably future conversion and correction of the ζ1 gene can be anticipated. Perhaps we are observing the earliest steps in a process that decides whether a duplicate gene will continue to function or will ultimately escape correction long enough to result in a virtually irreparable pseudogene. Alternatively, the ζ1 gene could be selected for under some circumstances and selected against in others. In this case we predict that the ζ1 genes will be polymorphic with respect to structure and function in human and chimpanzee populations.

Acknowledgment. This investigation was supported by USPHS grants GM 21346 to C.W.S. and AM 29800 to C.-K.J.S. and NSF grant BSR-84-15867 to A.C.W.

References

Bangerji J, Olson L, Schaffner W (1983) Cell 33:729–740
Bell GI, Selby MJ, Rutter WJ (1982) Nature 295:31–35
Chang L-Y, Slightom JL (1984) J Mol Biol 180:767–784
Chapman BS, Vincent KA, Wilson AC (1985) Genetics, in press
Gillies SD, Morrison SL, Oi VT, Tonegawa S (1983) Cell 33: 717–728
Harris S, Barrie PA, Weiss ML, Jeffreys AJ (1984) J Mol Biol 180:785–801
Heilig R, Muraskowsky R, Mandel J-L (1982) J Mol Biol 156: 1–19
Jeffreys AJ, Wilson V, Thein SL (1985) Nature 314:67–73
Kominami R, Muramatsu M, Moriwaki K (1983a) Nature 301: 87–89
Kominami R, Urano Y, Mishima Y, Muramatsu M, Moriwaki K, Yoshikura H (1983b) J Mol Biol 165:209–228
Lacy E, Maniatis T (1980) Cell 21:545–553
Lauer I, Shen C-KJ, Maniatis T (1980) Cell 20:119–130
Maroteaux L, Helig R, Dupret D, Mandel JL (1983) Nucleic Acids Res 11:1227–1245
Martin SA, Vincent KA, Wilson AC (1983) J Mol Biol 164: 513–528
Maxam AM, Gilbert W (1977) Proc Natl Acad Sci USA 74: 560–564
Messing J (1983) In: Wu R, Grossman L, Moldave K (eds) Methods in enzymology, vol 101 (part C): recombinant DNA. Academic Press, New york, pp 20–78
Messing J, Crea R, Seeburg PH (1981) Nucleic Acids Res 9: 309–321
Proudfoot NJ, Maniatis T (1980) Cell 21:537–544
Proudfoot NJ, Gil A, Maniatis T (1982) Cell 31:553–563
Proudfoot NJ, Rutherford TR, Partington GA (1984) EMBO J 3: 1533–1540
Sanger F, Nicklen S, Coulson AR (1977) Proc Natl Acad Sci USA 74:5463–5467
Sawada I, Beal MP, Shen C-KJ, Chapman B, Wilson AC, Schmid CW (1983) Nucleic Acids Res 11:8087–8101
Sawada I, Willard C, Shen C-K, Chapman B, Wilson A, Schmid CW (1985) Evolution of Alu family repeats since the divergence of human and chimpanzee. J Mol Evol 22:316–322
Schmid C, Shen C-KJ (1985) In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum, New York, pp 323–358
Sibley CG, Ahlquist JE (1984) J Mol Evol 20:2–15
Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC (1980) Proc Natl Acad Sci 77:2158-2162

Received July 9, 1985/Revised and accepted September 18, 1985