Volume 14 Number 4 1986 Nucleic Acids Research

A previously undetected pseudogene in the human alpha globin gene cluster

Ross C.Hardison, Ikuhisa Sawada[1], Jan-Fang Cheng, Che-Kun James Shen[2] and Carl W.Schmid[1]

Department of Molecular and Cell Biology, 206 Althouse Laboratory, Pennsylvania State University, University Park, PA 16802, [1]Department of Chemistry and [2]Department of Genetics, University of California, Davis, CA 95616, USA

Received 13 November 1985; Accepted 22 January 1986

ABSTRACT

The sequence of the DNA between two pseudogenes in the human α-like globin gene cluster has been determined. Comparison of this sequence with sequences from other α-like globin gene clusters revealed another pseudogene, ψα2, between the previously recognized pseudogenes ζ1 and ψα1. Therefore, the human α-like globin gene family is organized 5'-ζ2-ζ1-ψα2-ψα1-α2-α1-3'. The new pseudogene ψα2 is very close to ζ1, beginning only 65 base pairs 3' to the polyadenylation site of ζ1. The first exon and the first intron of ψα2 are interrupted by large inserts which are flanked by short (6 to 8 base pairs) direct repeats. The pseudogene ψα2 lacks a promoter for transcription by RNA polymerase II, the first exon is highly divergent, one splice site is mutated, and five different frameshift mutations have occurred in the coding regions. Thus ψα2 cannot encode a globin polypeptide. This pseudogene was not recognized in previous hybridization analyses of the human α-like globin gene cluster, and our discovery of it by sequence analysis suggests that divergent copies of many genes may be present in the human genome. These divergent copies of a large number of genes may comprise a substantial fraction of the slowly renaturing DNA of mammalian genomes.

INTRODUCTION

The families of genes that encode the globin polypeptides of hemoglobin are intensively studied because of their well-defined patterns of regulation during development and differentiation and because they are a rich source of information about genome evolution. The α-like and β-like globin gene families of humans are particularly well-characterized (reviewed in ref. 1 and 2). The α-like globin gene family of humans consists of an embryonic ζ-globin gene, ζ2, and a pair of almost identical α-globin genes (α2 and α1) that are expressed in the fetal liver and the adult bone marrow (3,4). These active genes are separated by pseudogenes in the gene cluster; ζ1 is very similar to ζ2 but has a premature termination codon (5), and ψα1 is defective in several aspects, including slice junction mutations and premature termination codons (6). The previously recognized genes are arranged 5'-ζ2-ζ1-ψα1-α2-α1-3'. The sequences of the intergenic regions in this gene

cluster are being determined in order to provide insights into the evolution of both single-copy and repeated sequences and to provide clues about potential regulatory sequences (7-10). In this report we present the sequence of a DNA segment that links the sequence of ζ1 (5) with the sequence of the 5' flank of ψα1 (10) and show that another pseudogene, ψα2, is present between ζ1 and ψα1.

MATERIALS AND METHODS

Materials

The 1.65 and 1.22 kilobase pair SstI fragments that span the region from ζ1 to the 5' flank of ψα1 (3) were subcloned as the plasmids pSV2-1.65 and pSV2-1.22. These plasmid DNAs were subcloned into M13 vectors to generate templates for the sequence determination.

DNA Sequencing

The sequence of the DNA was determined by the dideoxynucleotide chain termination method of Sanger et al. (11) using single-stranded DNA templates from M13 recombinant clones (12) generated from the plasmids pSV2-1.65 and pSV2-1.22.

Sequence Analysis

The DNA sequence was compared with other sequences from mammalian α-like globin gene clusters using the dot-plot program MATRIX (13). Regions that showed matches were aligned by the program NUCALN (14) or by inspection. The coding regions were compared by the algorithm of Perler et al. (15) using the program DIVERGENCE written by F. Fuller and A. Efstratiadis. All programs were run on an IBM PC XT.

RESULTS

The sequence of a 768 base pair (bp) segment between ζ1 and ψα1 was determined by the strategy outlined in Fig. 1. Both strands were sequenced for almost all of this segment, and the cloning sites for constructing the M13 clones used as templates in the sequencing were overlapped. This sequence connects the previously determined sequence of ζ1 (previously called ψζ, ref. 5) with the sequence of the 5' flank of ψα1, and a composite sequence is shown in Fig. 2. The sequence in Fig. 2 begins with the first nucleotide past the polyadenylation site of ζ1. Nucleotide number 145 is the end of the sequence published by Proudfoot et al. (5) and nucleotide number 914 is the beginning of the sequence published by Sawada et al. (10). The sequence in Fig. 2 ends 1796 bp before the cap site of pseudogene ψα1. With these new data, the

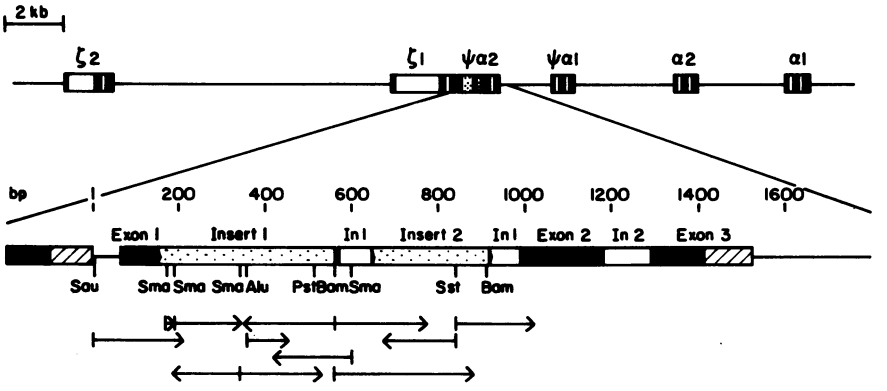

Figure 1. Structure and sequencing strategy for gene ψα2.
The top line presents the map of the human α-globin gene cluster. Filled boxes are exons, open boxes are introns, and stippled boxes are insert sequences. The lower portion shows a diagram of the pseudogene ψα2. Filled boxes are the polypeptide-coding portions of exons, hatched boxes are untranslated regions of the exons, open boxes are introns (In1 and In2), and stippled boxes are inserts in exon1 and intron 1. Restriction endonuclease cleavage sites used to generate the M13 subclones used for sequence determination are indicated below the diagram of ψα2. Arrows underneath the restriction map indicate the amount and orientation of sequence determined from each site.

sequence of a 17 kilobase pair region from the human α-like globin gene cluster has been completed. This DNA segment runs from 1000 bp 5' to ζ1 to 1500 bp 3' to α1 (5-10, 16-19).

A graphical dot-plot comparison of the sequence located between ζ1 and ψα1 with the sequences from other mammalian α-like globin gene families showed a significant match with the exons and introns of α-globin genes. These matches are shown in Fig. 2 as an alignment with the sequence of the human α2-globin gene (16). Matches were also found with exons 2 and 3 of the human ζ-globin gene, but no match was found with exon 1 or the two introns of the ζ2 gene. By using the method of Perler et al. (15), we found that the coding regions of the sequence presented in Fig. 2 match better with the human α2 gene than with the human ζ2 gene. As documented in Table I, the coding regions show a greater number of both replacement site substitutions and silent site substitutions in comparisons with the ζ2-globin gene than with the α2-globin gene. The amino acid sequence derived from the ψα2 sequence is about equally divergent from either the α-globin or the ζ-globin amino acid sequence in exon 2 and exon 3, but the ψα2 sequence is more related to α-globin in exon 1 (Table I). Therefore this sequence in the region between ζ1

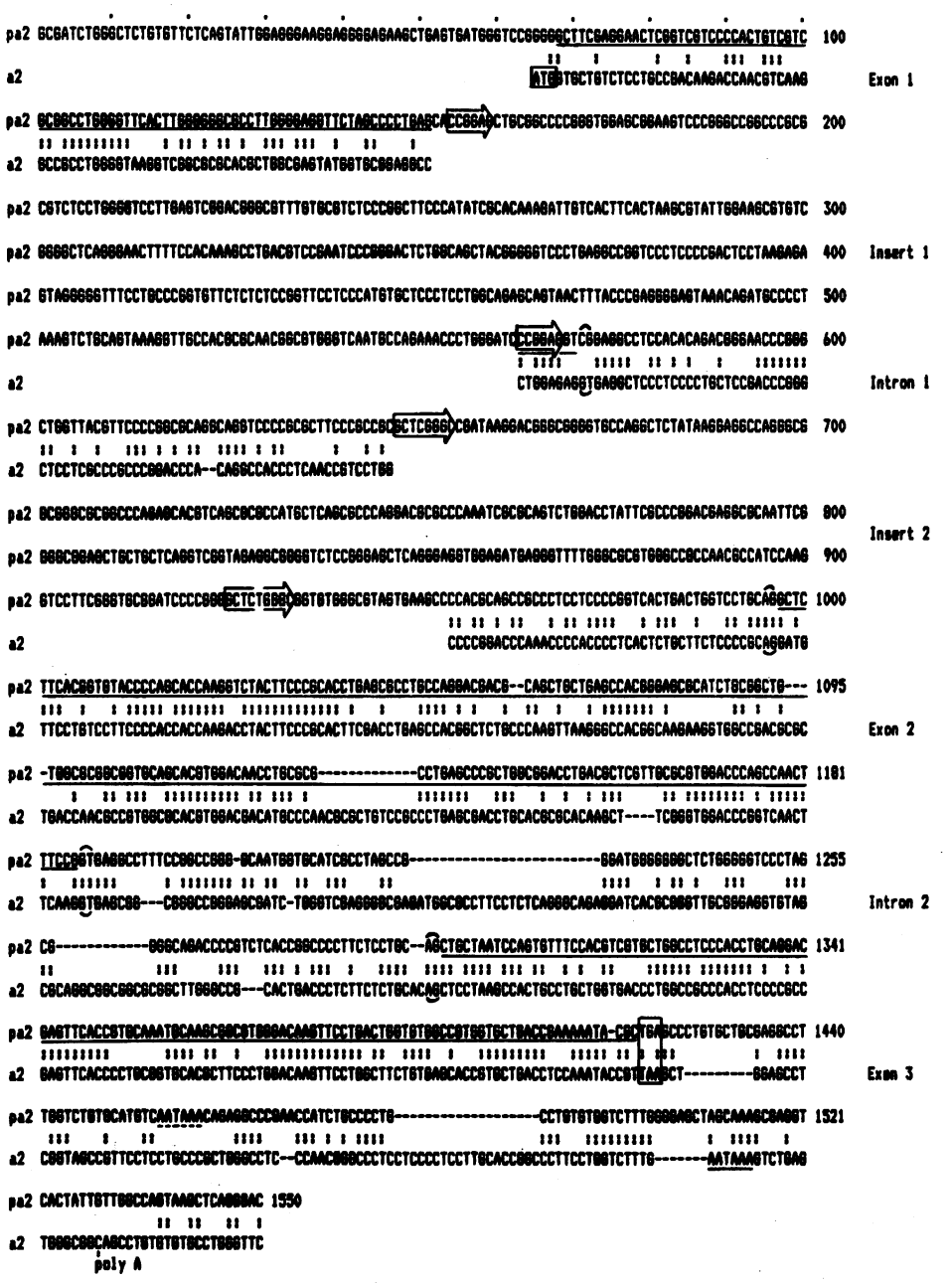

Figure 2. Sequence of pseudogene ψα2 aligned with the sequence of the α2-globin gene.

The sequence of ψα2 is presented on the top line (pa2) and the sequence

of α2 is on the lower line (a2). Spaces have been introduced optimize the alignment. The sequences in ψα2 that correspond to the polypeptide coding portions of α2 have been underlined, the GT and AG dinucleotides at the splice junctions have been overlined (ψα2) or underlined (α2), and the initiation and termination codons have been boxed. The short repeats flanking the inserts in exon 1 and intron 1 are in boxed arrows. The ψα2 sequence is numbered continuously from the first nucleotide past the polyadenylation site of ζ1.

and ψα1 appears to be that of a highly divergent gene related to the α-globin genes, and we refer to it as pseudogene ψα2.

Pseudogene ψα2 has suffered many deleterious mutations. No obvious 5' end of the gene can be assigned because the first part of exon 1 is so highly divergent. However, based on the matches with the α2 gene that start at nucleotide 101 in Fig. 2, the sequence GGG at nucleotides 65-67 corresponds to the ATG initiation codon of α2. Thus the polyadenylation site of gene ζ1 is juxtaposed with the 5' end of ψα2, and presumably the promoter region of ψα2 has been deleted. Three deletions and one short insertion were found in exon 2 of ψα2; all of these cause frameshifts in the α-globin translational reading frame. A single base pair deletion occurred in exon 3. The 5' splice site for intron 1 has mutated to the dinucleotide CG, and therefore it should not be functional. The matches with the α2 sequence fall off after the termination codon, and the sequence that corresponds in position to the AATAAA polyadenylation signal of α2 is highly mutated in ψα2, although the sequence AATAAA is present in ψα2 35 bp past the TGA termination codon. Any one of these mutations would prevent productive expression of gene ψα2 as a functional globin polypeptide, and thus we conclude that it is a pseudogene.

A DNA segment of 404 bp has inserted into exon 1 of ψα2 and a segment of 271 bp has inserted into intron 1. Both of these inserts are flanked by short direct repeats (boxed in Fig. 2) that are characteristic of transposable elements. Neither insert corresponds to the human repetitive elements that appear to be transposable elements -- Alu repeats (20), KpnI or L1 repeats (21) and O repeats (22). Part of the second insert matches for 22 of 40 nucleotides with part of the Alu repeat and thus it could be derived from a distantly related sequence. Alternatively, the inserts could be derived from DNA elements with a low repetition frequency. For example, they could be portions of processed pseudogenes (23). Insert 1 contains an open reading frame of 115 codons (in the opposite orientation to ψα2), but neither insert contains an open reading frame for its entire length.

The previously described α-globin pseudogene, ψα1, is much more closely related to gene α2 than is ψα2. As shown in Table I, ψα1 shows 0.22

Table I. Comparison of ψα2 with the coding regions of the α2- and ζ2-globin genes

Genes compared	Substitutions per site		% Mismatch of amino acids			
	Replacement	Silent	Exon 1	Exon 2	Exon 3	Total
ψα2 vs α2	0.54	0.66	69	60	48	58
ψα2 vs ζ2	0.61	0.94	91	57	48	62
α2 vs ζ2	0.38	0.60	66	37	33	42
ψα1 vs α2	0.22	0.48	38	47	50	46

The substitutions per site for replacement and silent sites were calculated by the method of Perler et al. (15); this includes a correction for multiple substitutions at a single site. These calculations involving ψα2 did not include the four deletions and one insertion in the coding region; thus these values are underestimates of the amount of divergence of ψα2. The percentage mismatch in the amino acid sequences was calculated by inspection of aligned sequences of the coding regions. These latter values have not been corrected for multiple substitutions at single sites, and each codon in a gap region was counted as a separate mismatch.

substitutions per replacement site when compared with α2, whereas ψα2 shows over twice as much divergence (0.52 substitutions per replacement site). Pseudogene ψα1 has suffered fewer insertions and deletions than has ψα2, and exon 1 of ψα1 matches quite well with α2 (6), in sharp contrast to the extensive divergence in exon1 of ψα2. The promoter and 5' untranslated region of ψα1 is still easily recognizable, and the ψα1 promoter retains some activity both *in vitro* and in transfected cells (24). Comparisons between ψα1 and ψα2 show even less similarity than that seen in comparisons with the functional α-globin genes (data not shown). These data imply that ψα2 has been diverging separately from ψα1 (i.e. there is no indication of gene conversion between them) and that ψα2 is an older pseudogene than ψα1.

DISCUSSION

Analysis of the DNA sequence between genes ζ1 and ψα1 in the human α-like globin gene family has revealed the presence of another pseudogene, ψα2. Thus this gene family is arranged 5'-ζ2-ζ1-ψα2-ψα1-α2-α1-3'. The three pseudogenes in the human α-like globin gene cluster are located within a region of 6.5 kilobase pairs. This is unusually close for linked mammalian genes. Pseudogene ψα2 is only 65 bp downstream from ζ1, but ψα2 lacks any semblance of a normal 5' end. It is possible that ζ1 and ψα2 were fused together by a

deletion of the intergenic sequences that previously separated them, and this fusion could have removed the 5' end of ψα2.

The divergence of ψα2 from the active α-globin genes has not occurred evenly throughout the gene. Exon 3 is the most similar to α2, followed by exon 2, and exon 1 is the most divergent. This is similar to the gradient of divergence recognized by Hess et al. (9) in the α2-α1 duplication unit in this same gene cluster, except that the 5' end of the duplication unit is the most highly conserved. Hess et al. (9) attribute this gradient of divergence to more frequent gene corrections that initiate at the 5' end of the duplication unit. A similar explanation could be invoked in the case of ψα2, although this gene clearly is not correcting against an active α-globin gene as frequently as α1 and α2 are correcting against each other. The two large inserts, one into exon 1 and another into intron 1, could contribute to the more extensive divergence in the 5' end of ψα2, possibly by blocking gene correction events.

Pseudogene ψα2 has been nonfunctional for a very long time. Proudfoot and Maniatis (6) estimate that the divergence of ψα1 from the functional α-globin genes occurred about 60 million years ago; this could represent the time of the last gene correction event. The more extensive divergence of ψα2 shows that this gene has been diverging from the functional α-globin genes for an even longer time. Thus ψα2 could have been a pseudogene prior to the mammalian radiation (about 60 to 85 million years ago, ref. 25), and therefore it may still be present in the α-like globin gene clusters from other orders of mammals.

Pseudogene ψα2 was not detected in previous mapping and blot-hybridization analysis of cloned DNA from the human α-like globin gene cluster, but it is quite apparent in computer-generated comparisons of nucleotide sequences. This situation is reminiscent of the discovery of pseudogene βh2 in the mouse β-like globin gene cluster (26). This gene also was not detected by blot-hybridization analysis but was found in a nucleotide sequencing project. These observations raise the possibility that many other genes may have highly divergent copies in the genome that have so far escaped detection. Although the examples cited here (human ψα2 and mouse βh2) are members of multi-gene families that are evolving by gene duplication and divergence, it is possible that many of the genes that are thought to be "single copy" also have some number of duplicated copies in the genome. Analysis of the thermal denaturation profiles of slowly renaturing DNA from primates (27,28), mouse (29) and chicken (30) showed two components, and it

has been proposed that the low-melting component contains multiple, divergent copies of the "single copy" DNA. Our discovery of ψα2 in the human α-like globin gene cluster supports this hypothesis, and emphasizes the possibility that very few genes are actually present in only one copy in the haploid genome. If in fact many genes have several divergent copies (presumably pseudogenes), then a substantial portion of the genome may be occupied by pseudogenes. This apparent accumulation of pseudogenes supports the hypothesis of Hutchinson et al. (31) that pseudogenes persist in the genome because gene duplications occur more frequently than do deletions of defective pseudogenes, thereby resulting in an increase in genome size.

ACKNOWLEDGEMENTS

We thank R.D. Porter for writing the version of DIVERGENCE that runs on an IBM PC. This work was supported by USPHS grants AM 27635 and AM 31961 to R.C.H., AM 29800 to C.-K.J.S. and GM 21346 to C.W.S.

REFERENCES

1. Maniatis, T., Fritsch, E. F., Lauer, J. & Lawn, R. M. (1980) Annu Rev. Genet. 14, 145-178.
2. Collins, F. S. & Weissman, S. M. (1984) Progress in Nucleic Acids Res. and Mol. Biol. 31, 315-462.
3. Lauer, J., Shen, C.-K. J. & Maniatis, T. (1980) Cell 20, 119-130.
4. Peschle, C., Mavilio, F., Care, A., Migliaccio, G., Migliaccio, A. R., Salvo, G., Samoggia, P., Petti, S., Guerriero, R., Marinucci, M., Lazzaro, D., Russo, G. & Mastroberardino, G. (1985) Nature (London) 313, 235-238.
5. Proudfoot, N. J., Gil, A. & Maniatis, T. (1982) Cell 31, 553-563.
6. Proudfoot, J. N. & Maniatis, T. (1980) Cell 21, 537-544.
7. Michelson, A. M. & Orkin, S. H. (1983) J. Biol. Chem. 258, 15245-15254.
8. Hess, J. F., Fox, M., Schmid, C. W. & Shen, C.-K. J. (1983) Proc. Natl. Acad. Sci. U. S. A. 80, 5970-5974.
9. Hess, J. F., Schmid, C. W. & Shen, C.-K. J. (1984) Science 226, 67-70.
10. Sawada, I., Beal, M. P., Shen, C.-K. J., Chapman, B., Wilson, A. C. & Schmid, C. W. (1983) Nucleic Acids Res. 11, 8087-8101.
11. Sanger, F., Nicklen, S. & Coulson, A. (1977) Proc. Natl. Acad. Sci. U. S. A. 74, 5463-5467.
12. Messing, J. (1983) Methods Enzymol. 101, 20-78.
13. Zweig, S. E. (1984) Nucleic Acids Res. 12, 767-776.
14. Wilbur, W. J. & Lipman, D. J. (1983) Proc. Natl. Acad. Sci. U. S. A. 80, 726-730.
15. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) Cell 20, 555-565.
16. Liebhaber, S. A., Goossens, M. J. & Kan, Y. W. (1980) Proc. Natl. Acad. Sci. U. S. A. 77, 7054-7058.
17. Michelson, A. M. & Orkin, S. H. (1980) Cell 22, 371-377.
18. Willard, C., Wong, E., Hess, J. F., Shen, C.-K. J., Chapman, B., Wilson, A. C. & Schmid, C. W. (1986) J. Mol. Evol. 22, in press.
19. Hardison, R. C. & Gelinas, R. (1986) Mol. Biol. Evol., submitted.
20. Schmid, C. W. & Jelinek, W. R. (1982) Science 216, 1065-1070.
21. Singer, M. F. & Skowronski, J. (1985) Trends Biochem. Sci. 10, 119-122.

22. Paulson, K. E., Deka, N., Schmid, C. W., Misra, R., Schindler, C. W., Rush, M. G., Kadyk, L. & Leinwand, L. (1985) Nature (London) 316, 359-361.
23. Hollis, G., Heiter, P., McBride, O. W., Swan, D. & Leder, P. (1982) Nature (London) 296, 321-325.
24. Whitelaw, E. & Proudfoot, N. J. (1983) Nucleic Acids Res. 11, 7717-7733.
25. Romero-Herrera, A. E., Lehmann, H., Joysey, K. A. & Friday, A. E. (1973) Nature (London) 246, 389-395.
26. Phillips, S. J., Hardies, S. C., Jahn, C. L., Edgell, M. H. & Hutchison, C. A. III. (1984) J. Biol. Chem. 259, 7947-7954.
27. Deininger, P. L. & Schmid, C. W. (1979) J. Mol. Biol. 127, 437-460.
28. Fox, G. M. & Schmid, C. W. (1980) Biochim. Biophys. Acta 609, 349-363.
29. Ivanov, I. G. & Markov, G. G. (1974) FEBS Lett. 47, 323-326.
30. Burr, H. E. & Schimke, R. P. (1980) J. Mol. Evol. 15, 291-307.
31. Hutchinson, C. A. III, Hardies, S. C., Padgett, R. W., Weaver, S. & Edgell, M. H. (1984) J. Biol. Chem. 259, 12881-12887.