The EMBO Journal vol.3 no.2 pp.439–446, 1984

Organization of the human myoglobin gene

P. Weller*, A.J. Jeffreys, V. Wilson and A. Blanchetot

Department of Genetics, University of Leicester, Leicester LE1 7RH, UK

*To whom reprint requests should be sent

Communicated by A.J. Jeffreys

Cross-hybridization of the grey seal myoglobin gene to human DNA detected a single human myoglobin gene plus an extensive family of sequences apparently related to the central exon of this gene. The functional human gene is 10.4 kb long and has a haemoglobin-like three exon/two intron structure with long non-coding regions similar to its seal homologue. At least 300 bp of 5′-flanking region are closely homologous between the two genes, with the exception of a divergent purine-rich region 68–114 bp upstream of the cap site. A diverged tandem repetitive sequence based on $(GGAT)_{165}$ is located 1100–1750 bp upstream from the gene; internal homology units within this sequence suggest sequence homogenization by gene microconversions. A second 33-bp tandem repeat element in the first intron is flanked by a 9-bp direct repeat, shares homology with other tandem repetitive elements in the human genome and may represent a novel form of transposable element.

Key words: human/myoglobin/gene structure/evolution/repeated DNA/transposition

Introduction

Vertebrate myoglobin is the principal haemoprotein of muscle where it facilitates the diffusion of oxygen from the vascular system to muscle mitochondria (Wittenberg, 1970). In mammals, myoglobin accumulates in cardiac muscle early in foetal development, whereas the appearance of myoglobin in skeletal muscle is delayed until birth, correlating with the relative activities of these two muscle types during development (Longo *et al.*, 1973; Tipler *et al.*, 1978). Elevated levels of myoglobin in skeletal and/or cardiac muscle are found in diving mammals and birds, and in some mammals adapted to hypoxic subterranean or high altitude environments (Lenfant, 1973; Nevo, 1979; Butler and Jones, 1982). In some instances, for example the sperm whale and seal, myoglobin levels are sufficiently high to serve as a significant oxygen store during diving (see Wittenberg, 1970).

Monomeric myoglobins and tetrameric haemoglobins form part of the same globin superfamily and have arisen by successive duplications of a primordial globin gene. The myoglobin/haemoglobin duplications preceded the divergence of the α- and β-globin genes ~450–500 million years ago, although the timing of this duplication, variously estimated at 500 million years (Czelusniak *et al.*, 1982) to 800 million years ago (Hunt *et al.*, 1978), is still the subject of debate (Kimura, 1981).

We have previously shown that myoglobin of the grey seal (*Halichoerus grypus*) is specified by a single gene which is transcribed to produce one of the most abundant mRNAs in juvenile seal skeletal muscle (Wood *et al.*, 1982). The seal myoglobin gene has the three exon-two intron structure found in α- and β-globin genes, indicating that this structure was established prior to the myoglobin/haemoglobin divergence (Blanchetot *et al.*, 1983). The seal myoglobin gene differs markedly from all characterized α- and β-globin genes, in having very long introns and a long 3′-non-translated mRNA sequence. In addition, the promoter region is atypical, with a highly purine-rich sequence 67–108 bp upsteam of the cap site in a region normally containing the CCAAT and –100 sequences of globin promoters (Dierks *et al.*, 1983).

To determine whether these unusual features in any way represent a recent adaptation of the seal myoglobin gene to a diving physiology, we now describe the isolation and analysis of the myoglobin gene from the terrestrial mammal, man. In addition, we discuss the properties of two tandem repetitive sequence blocks associated with the human myoglobin gene, which give insights into the origin and maintenance of these dispersed blocks.

Results and Discussion

Detection of the human myoglobin gene: evidence for an exon 2-related family

Initial cloning of myoglobin mRNA from seal muscle poly(A)$^+$ RNA generated an incomplete cDNA clone containing only part of the 3′-non-translated region of myoglobin mRNA (Wood *et al.*, 1982). Since this myoglobin cDNA failed to cross-hybridize to human DNA (data not shown), more suitable probes containing myoglobin coding sequences were prepared from clones of the seal myoglobin gene (Figure 1A). Hybridization of seal exons 1 and 3 to restriction endonuclease digests of human DNA revealed a single major hybridizing fragment plus, in the case of exon 3, one or two faint additional components (Figure 2). In marked contrast, exon 2 detected a complex set of at least 10 hybridizing fragments in all restriction endonuclease digests tested. The same complex pattern was found in the *Eco*RI-digested DNAs of six different individuals and could be simplified to a single major hybridizing component by increasing the post-hybridization washing stringency from 1 x SSC at 65°C to 0.2 x SSC at 65°C (data not shown). We conclude that human myoglobin is specified by a single gene, and that the human genome also contains an extensive family of sequences preferentially related to the central exon of the seal myoglobin gene. A more detailed anaysis of this family will be presented elsewhere.

Isolation of the human myoglobin gene

A genomic library of human DNA cloned into the bacteriophage vector λL47.1 (Loenen and Brammar, 1980) was screened by hybridization with seal myoglobin exons 1 and 3, which only detect the major gene. Ten positive plaques were isolated and shown by restriction mapping to be overlapping isolates of a single region of the human genome (Figure 1B). Comparison of the map of this region with the restriction fragments detected in human genomic DNA by seal exons

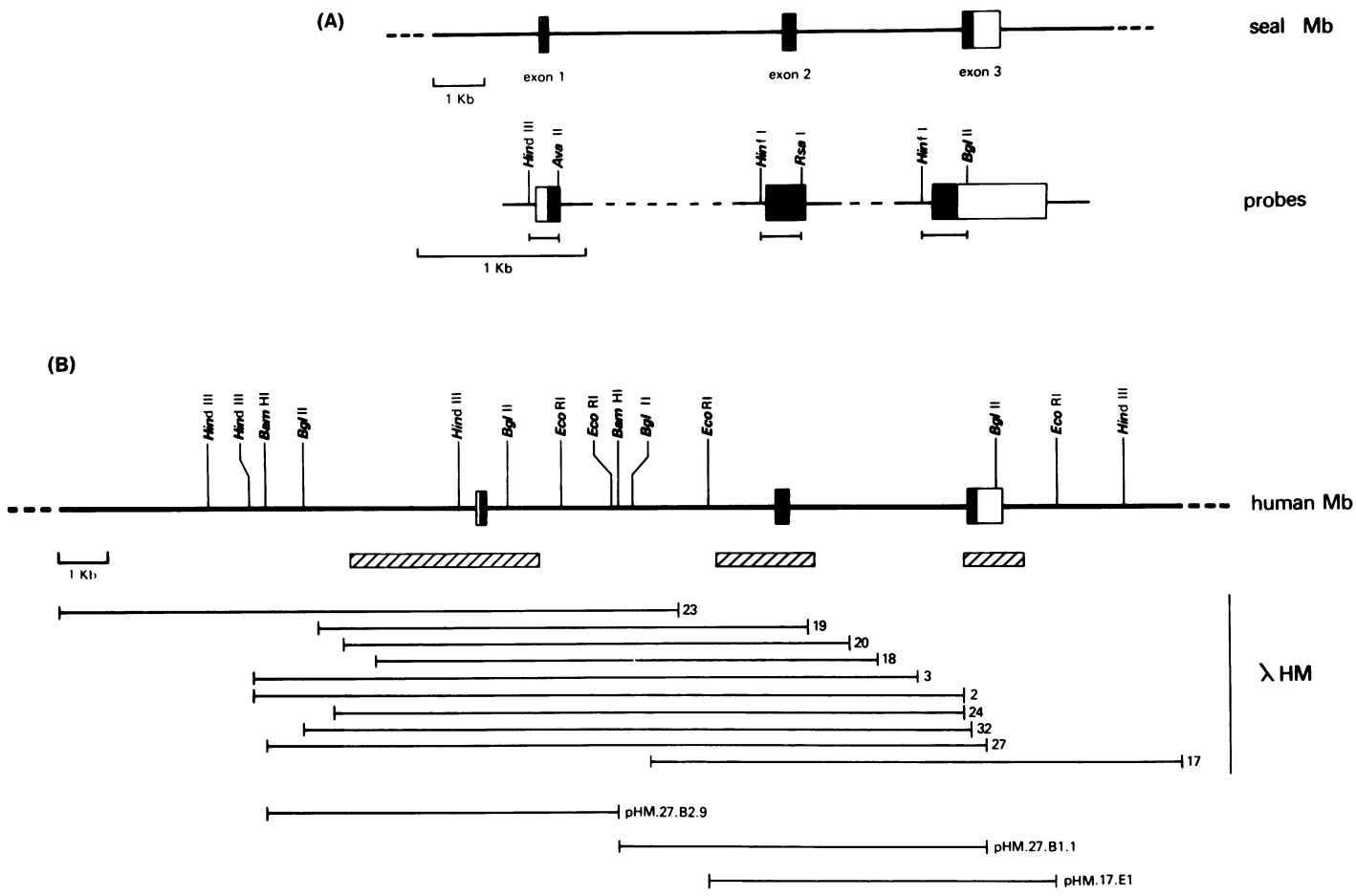


Fig. 1. Isolation and characterization of the human myoglobin gene. **(A)** The organization of the seal myoglobin gene is shown, with coding sequences indicated by filled boxes and non-translated mRNA sequences by open boxes. Exon-specific probes were purified after cleavage of seal myoglobin subclones pSM 19.5 and pSM 1.17 (Blanchetot *et al.*, 1983) with the indicated restriction endonucleases. The exon 1 fragment comprised 46 bp 5′-flanking sequence, 70 bp 5′-non-translated sequence and 84 bp coding sequence. Exon 2 was isolated in a fragment containing 25 bp of intron 1 plus 233 bp coding sequence. The exon 3 fragment consisted of 66 bp intron 2, 147 bp coding sequence and 67 bp 3′-non-translated sequence. **(B)** DNA was prepared from a human placenta (male, European) and a library of 10–20 kb *Sau*3A partials cloned into the *Bam*HI replacement vector λL47.1 (Loenen and Brammar, 1980) was prepared as described by Jeffreys *et al.* (1982). 3 x 10^6 recombinant plaques were screened by hybridization in 1 x SSC at 65°C with ^{32}P-labelled seal exon 1 and exon 3 probes. 10 strongly hybridizing plaques (λHM. 2-32) were purified and recombinant phage DNA mapped with *Bam*HI, *Bgl*II, *Eco*RI and *Hind*III. Human myoglobin exons were located by Southern blot hybridization of phage DNAs with seal myoglobin exon probes. *Bam*HI restriction fragments from λHM.27 and an *Eco*RI fragment from λHM.17 were subcloned into pAT 153 (Twigg and Sherratt, 1980) to give pHM.27.B2.9, pHM.27.B1.1 and pHM.17.E1. These plasmids were sheared by sonication and DNA fragments shotgun cloned into M13mp8 (Messing and Vieira, 1982). Phage carrying myoglobin gene sequences from the required regions were identified by hybridization and sequenced by the chain-termination procedure of Sanger *et al.* (1977). The sequences of the hatched regions were completely determined, with all regions being sequenced at least twice and 93% of the sequence being determined on both DNA strands.

1–3 (Figure 2) indicated that the cloned region contained the major myoglobin gene. Human myoglobin exons were located in the recombinant phage DNAs by hybridization with seal myoglobin exons, subcloned into pAT 153 (Twigg and Sherratt, 1980), shotgun cloned into M13mp8 (Messing and Vieira, 1982) and sequenced by the dideoxynucleotide chain termination procedure (Sanger *et al.*, 1977; Biggin *et al.*, 1983). The DNA sequence of the human myoglobin gene, excluding most of the extensive intron regions (Figure 1B), is shown in Figure 3.

Structure of the human myoglobin gene

The amino acid sequence derived from the cloned human myoglobin gene corresponds exactly with the myoglobin sequence determined by Romero-Herrera and Lehmann (1974). This establishes that the cloned gene is the functional myoglobin gene in man which specifies skeletal and cardiac muscle myoglobin. The grey seal also has a single functional myoglobin gene (Blanchetot *et al.*, 1983) and, based on protein studies, single genes are likely to be the rule in vertebrates (Romero-Herrera *et al.*, 1978). The only exception is found in certain New World monkeys (Humboldt's woolly monkey, squirrel monkey and the common marmoset), which also produce a minor myoglobin species characterised by a two residue Phe-Lys N-terminal extension (Romero-Herrera and Lehmann, 1973). It is unlikely that this minor species has evolved by premature translation in the 5′-non-translated region of a duplicated myoglobin gene (Romero-Herrera and Lehmann, 1973), since the sequence around the initiation codon

$$\overset{\text{ini}}{\text{(TGC.GCC.ATG)}}$$

could not be readily mutated to give the

$$\overset{\text{ini phe lys}}{\text{N-terminal extension (ATG.TTY.AAX).}}$$

The human myoglobin gene is interrupted by two introns, at codon 31 and between codons 105 and 106. These locations are identical to those in the seal myoglobin gene and also cor-

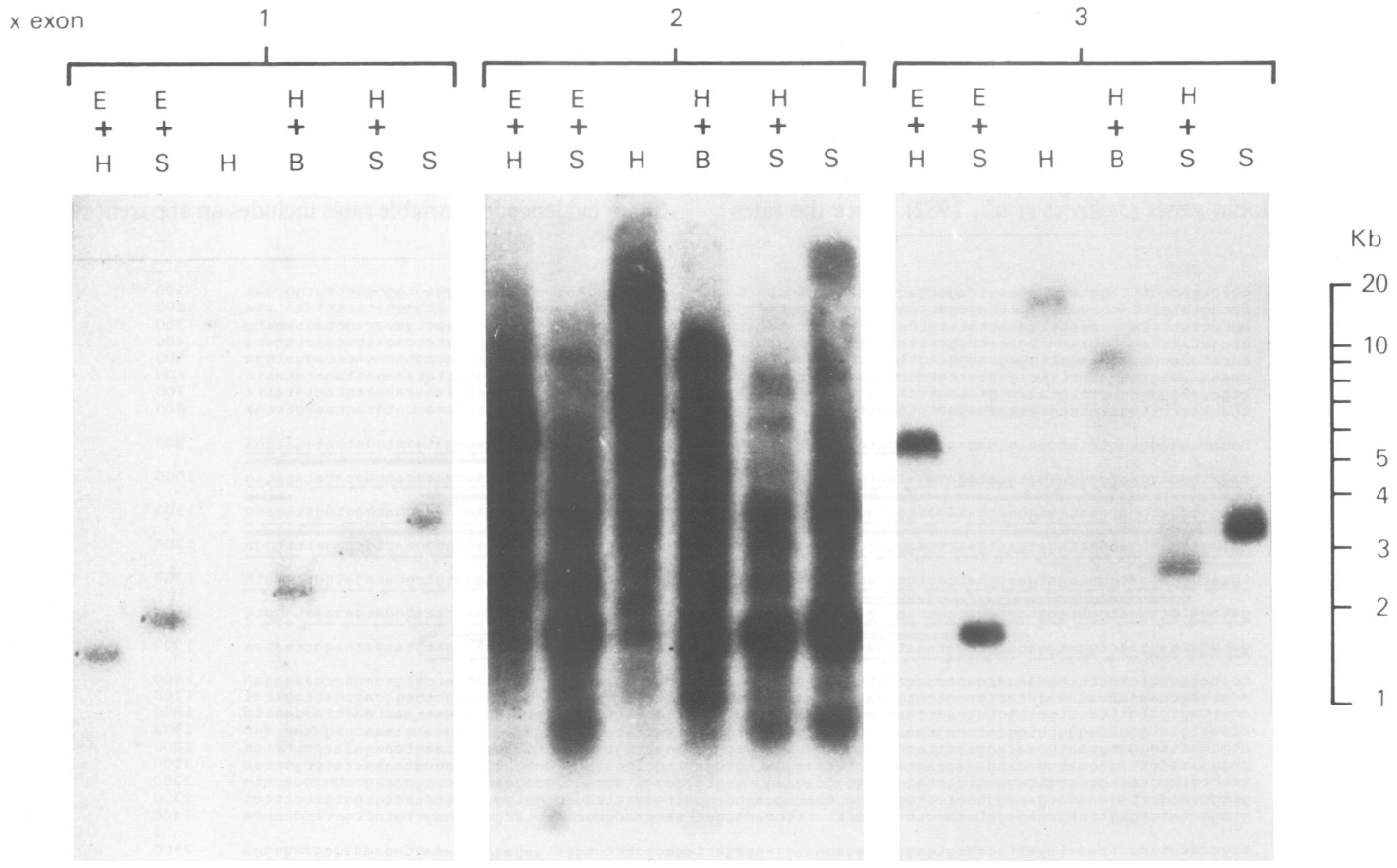


Fig. 2. Hybridization of human DNA with exons isolated from the seal myoglobin gene. 10 μg samples of human DNA were digested with combinations of *Bam*HI (B), *Eco*RI (E), *Hind*III (H) and *Sst*I (S), denatured with alkali, electrophoresed through a 0.8% agarose gel and blotted onto nitrocellulose. The filter was hybridized with ^{32}P-labelled DNA fragments containing seal myoglobin exon 1, 2 or 3 (see Figure 1) in 1 x SSC at 65°C, followed by washing in 1 x SSC at 65°C and autoradiography.

respond to the intron positions in all characterised vertebrate α- and β-globin genes (Blanchetot *et al.*, 1983). Intron 1 (5.8 kb) and intron 2 (3.6 kb) are much longer than their haemoglobin counterparts (108–886 bp and 103–906 bp, respectively; Blanchetot *et al.*, 1983; Proudfoot *et al.*, 1982) and combine to make the human myoglobin gene the longest globin gene so far described (10.4 kb long).

The cap site and polyadenylation site were located by homology with sites in the seal myoglobin gene. The lengths of the myoglobin mRNA 5′-non-translated region (70 nucleotides) and 3′-non-translated region (531 nucleotides) predict a myoglobin mRNA 1066 nucleotides long [excluding the poly(A) tail], similar to the 1083-nucleotide seal myoglobin mRNA (Blanchetot *et al.*, 1983). Northern blot analysis of seal muscle and human gastrocnemius muscle poly(A)$^+$ RNA probed with seal myoglobin exon 2 confirmed that the seal and human myoglobin mRNAs are indistinguishable in size (data not shown).

The 5′-flanking region of the human myoglobin gene contains only one of the three conserved elements frequently found near globin genes and required for promotion of transcription by RNA polymerase II (Efstratiadis *et al.*, 1980; Grosveld *et al.*, 1982; Dierks *et al.*, 1983). A normal TATA box is located 33 bp before the cap site. The CCAAT box generally located 70–90 bp upstream is absent, though a CCATT sequence exists at position −67 but is not conserved between the seal and human myoglobin genes (Figure 4). The dimerised CACCC element in the −100 region of β-globin genes (Dierks *et al.*, 1983) is also absent. Instead, the region 68–114 bp upstream from the cap site is occupied by an unusually purine-rich sequnce (90% AG) also found in the seal myoglobin gene.

Comparison of the human and seal myoglobin genes

The overall organization of the human and seal genes is very similar (Figure 1), the principal difference being an increase in the length of intron 1 from 4.8 kb in seal to 5.8 kb in man. Clearly, the long non-coding regions of the seal myoglobin gene compared with α- and β-globin genes are not peculiar to this species but instead are likely to represent a stable organisation of this gene, at least in the mammals. Evolutionary stability of intron length in mammalian α- and β-globin genes has also been noted (Van Den Berg *et al.*, 1978; Blanchetot *et al.*, 1983) although it is not yet clear whether this stability reflects functional conservation, or whether instead the frequency of deletions/insertions is insufficient to alter substantially the length of these regions during the time following the mammalian radiation, 80 million years ago (Romero-Herrera *et al.*, 1973).

Sequence divergence between the human and seal myoglobin genes is shown in Figure 4. The overall level of sequence substitution in the non-coding regions is fairly uniform and shows no significant difference from the mean level of silent site divergence in the coding sequences (35%). Since seal and man last shared a common ancestor 80 million years ago at the time of the mammalian radiation (Romero-

Herrera *et al.*, 1978) this gives a mean rate of non-coding DNA evolution of 2.2 x 10^{-9} nucleotide substitutions per site per year. This rate is substantially lower than the apparently constant intron and silent site rates (4−6 x 10^{-9}/year) reported for a wide variety of mammalian genes (Hayashida and Miyata, 1983) and is more in line with the rate of 2 x 10^{-9}/year derived from comparisons of prosimian and human β-globin genes (Jeffreys *et al.*, 1982). Since the rates of Hayashida and Miyata (1983) were largely determined from human/rodent and human/lagomorph comparisons, we suggest that the rate discrepancies reflect different substitution rates in different mammalian lineages. In particular, it suggests a low rate of substitution in non-coding DNA of 2 x 10^{-9}/year in primates and carnivores, and an accelerated rate of 6−8 x 10^{-9}/year in the rodent lineage. Other evidence for variable rates includes an apparent further

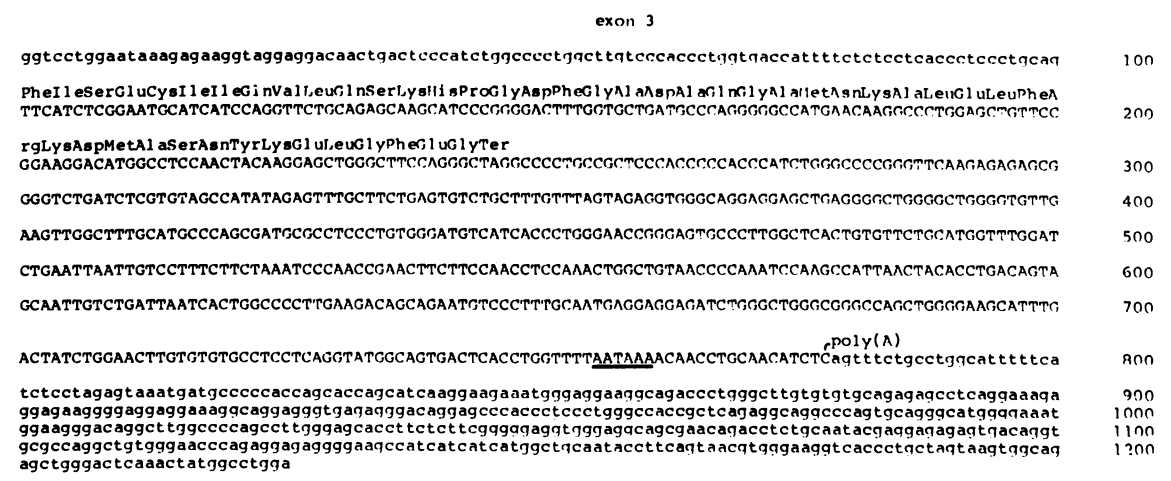


Fig. 3. DNA sequence of the human myoglobin gene. Sequences present in mature myoglobin mRNA are shown in capital letters. Positions of the cap site and poly(A) addition site were deduced from homologous sequence locations in the seal myoglobin gene (Blanchetot *et al.*, 1983). The 5′ polypurine and TATA boxes and the AATAAA polyadenylation sequence are underlined. Tandem repetitive sequence elements are underlined with arrows.

rate reduction to 1.2 x 10^{-9}/year in the great ape/man lineage (Maeda *et al.*, 1983) and lineage-dependent cross-hybridization efficiencies of genes between different mammalian orders (Wilson *et al.*, 1983), and may reflect a neutral mutation rate geared to generation time rather than absolute time (Kohne, 1970).

There is evidence in primate β-globin genes and the *Drosophila melanogaster* alcohol dehydrogenase gene that 3′-non-translated regions of mRNAs may have been subjected to strong purifying selection (Martin *et al.*, 1981; Kreitman, 1983). This appears not to be the case with myoglobin genes, in which most of this region has diverged at the 'neutral' rate and has accumulated numerous microdeletions/insertions (Figure 4). The only significantly conserved region, other than coding sequences, extends at least 300 bp upstream from the cap site (Figure 4). Similar conserved 5′-flanking regions have been noted for β-globin genes (Moschonas *et al.*, 1982). Curiously, substitutions in this region of the myoglobin gene are not randomly distributed but tend to be clustered, particularly into the divergent purine-rich region located 68 – 114 bp upstream from the cap site. Since these sequences show evidence of internal repetition [for example, the seal region is based on $(GGA)_7.(GA)_6$], the divergence may have arisen by slippage events during DNA replication rather than by base substitution. It is not yet known whether any of these changes in the 5′-flanking regions are implicated in the different levels of myoglobin production in human and seal muscle.

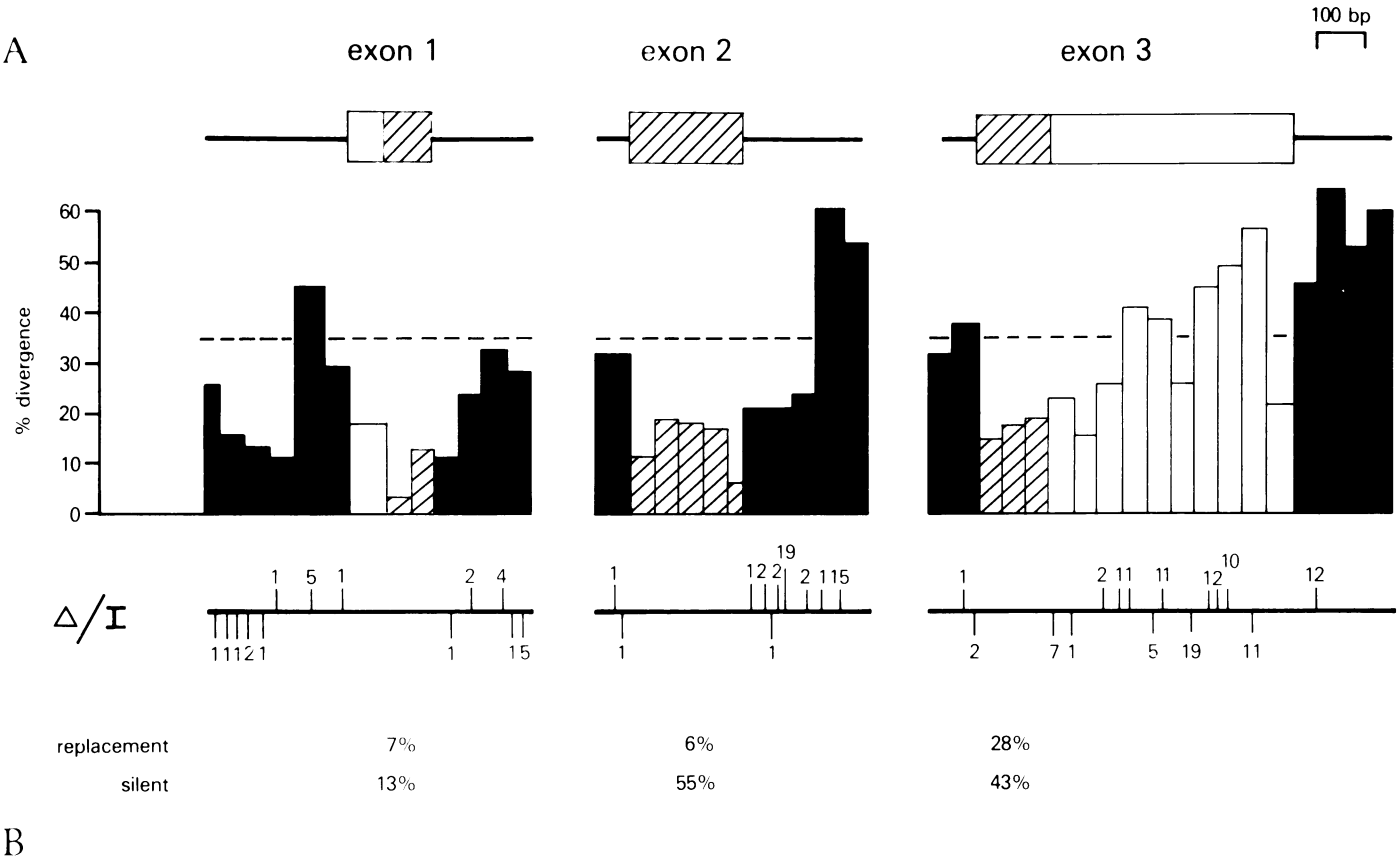
A simple repetitive region upstream from the human myoglobin gene: evidence for microconversion

During initial characterization of the myoglobin gene, an unusual region 1100 – 1750 bp upstream from the cap site was noted; this region was devoid of any restriction endonuclease cleavage sites, and restriction fragments covering this region showed pronounced strand separation after alkali denaturation and electrophoresis on neutral agarose gels, indicating a strong strand asymmetry in base composition (data not shown). DNA sequencing of this region (Figure 3) revealed a 652-bp tandem repetitive sequence based on $(GGAT)_{165}$ but substantially diverged (22%) from this hypothetical ancestral sequence, suggesting that this repeated sequence is ancient. Dot matrix analysis also revealed two prominent pairs of longer homology blocks in this region (Figure 3). One 116-bp block is repeated 12 bp downstream by a second 116-bp block differing in only 3 bp from the first. Similarly, another 58-bp duplication exists, with duplicates differing by 3 bp and separated by 59 bp. Since neither pair of homology blocks exists as a tandem duplication, it is most unlikely that either of these has arisen by unequal crossing over between different elements of the $(GGAT)_{165}$ array. Instead, these dispersed homology blocks are more reminiscent of the patch homologies commonly seen between members of gene families (see Dover, 1982; Jeffreys and Harris, 1982) which are thought to have arisen by gene conversion of misaligned duplicate genes.

These gene conversion units are generally much longer (900 – 1800 bp; see Proudfoot *et al.*, 1982) than the 58-bp and 116-bp units near the myoglobin gene. Nevertheless, it is likely that both gene conversion and these microconversion events proceed by a similar mechanism; the short length of the microconversion units might be due either to divergent DNA flanking these units which prevents branch migration spreading far from the initial site of recombination, or to resolution of a non-isomerized recombination complex shortly after displacement, uptake and assimilation of a single DNA strand at a short heteroduplex junction (Radding, 1978). In any event, it suggests that sequence homology may be maintained in these simple tandem-repeated elements by conversions which do not alter the copy number of repeats in the array, and may also be implicated in microhomogenization events in other gene families. There is evidence that microconversions occur between genes in the major histocompatibility complex and may be significant in generating polymorphism at this locus (Weiss *et al.*, 1983).

A tandem repeated element in intron 1: evidence for transposition

A second tandem repetitive element was discovered towards the 3′ end of intron 1 in the human myoglobin gene (Figure 3). This element consists of four repeats of a 33-bp sequence which differ from each other at only two nucleotide sites. Several lines of evidence suggest that this element may have arisen in the myoglobin gene by transposition. First, the element is flanked by a 9-bp direct repeat (8/9 bases identical) characteristic of the target site duplication seen at the insertion site of numerous classes of transposable elements. Second, the 33-bp repeat element shows significant homology to



B

```
man     cctttctgtcggct-atgagagtccagacagt-gcccaacctcctccc-cttctttccacacgcacaaccaccccac--ccccctgtggcctgagctgtcc
seal    cccttctgctggcttctgagactccagacagtagcccaacctcctccctgctcttgccacatacacaaacacccccacgcccctgtggcctgagttgtcc
man     tgcctcgccac-aatggcacctgccctaaaatagcttcccatgtgagggctagagaaaaggaaaagattagaccctccctggatgagagagagaaagtgaa
seal    tacctcccccccaaatggcacctgccccaaaatagcttcc-atgtgagagctacagaaaggaaaagattagaccccgcatgaaaaggagggaggaggagga
                                                                                                        CAP
man     ggagggcagggggaggggacagcgagccattgagcgatctttgtcaagcatcccagaaggtataaaaacgccttgggaccaggcagcctcaaACCCCAG
seal    gg-----aggagagagagtgagcaagctgtcaagtgatctgtgtcaagcttctgggaaagtataaaatc-cctctgggggcaggcgatctcaaACCCCAG
                                                                                ini Gly Leu Ser Asp Gly Glu Trp Gln L
man     CTGTTGGGGCCAGGACACCCAGTGAGCCCATACTTGCTCTTTTTGTCTTCTTCAGACTGCGCCATG.GGG.CTC.AGC.GAC.GGG.GAA.TGG.CAG.T
seal    CTGTCAGAGCCAGGACACCGAGTCAGCCGGGACTTGCTCTTCTTGTCTTCCTCAGACTGCACCATG.GGG.CTC.AGC.GAC.GGG.GAA.TGG.CAC.T
                                                                                ini Gly Leu Ser Asp Gly Glu Trp His L

        eu Val Leu Asn Val Trp Gly Lys Val Glu Ala Asp Ile Pro Gly His Gly Gln Glu Val Leu Ile Ar(q)
man     TG.GTG.CTG.AAC.GTC.TGG.GGG.AAG.GTG.GAG.GCT.GAC.ATC.CCA.GGC.CAT.GGG.CAG.GAA.GTC.CTC.ATC.AGgtaaaaggaaag
seal    TG.GTG.CTG.AAC.GTC.TGG.GGG.AAG.GTA.GAG.ACT.GAC.CTC.GCG.GGC.CAT.GGG.CAG.GAG.GTC.CTC.ATC.AGgtaaaaggaaag
        eu Val Leu Asn Val Trp Gly Lys VAl Glu Thr Asp Leu Ala Gly His Gly Gln Glu Val Leu Ile Ar(q)
```

Fig. 4. DNA sequence divergence between the human and seal myoglobin genes. Homologous sequences were identified and aligned by dot matrix analysis of the two genes (Konkel *et al.*, 1979). **(A)** Percentage base substitution divergence between the two genes was calculated over 50-bp intervals and corrected for multiple substitutions at single sites (Kimura and Ohta, 1972). Replacement and silent site divergences were calculated for each exon by the method of Perler *et al.* (1980). Δ/I: distribution and sizes of microdeletions/insertions between the human and seal genes; sequences present in human but not seal are shown above the line, and those specific to seal below the line. **(B)** Detailed alignment of the 5′-flanking region and exon 1 of the human and seal genes. Substitution as shown by asterisks, and the TATA and polypurine regions are indicated by boxes.

a 36-bp repeat sequence located between the human ζ- and ψζ-globin genes in man (Goodbourn *et al.*, 1983) which in turn shows similiarites to 14-bp repeat elements in the first intron of the ζ-globin gene (Proudfoot *et al.*, 1982) and in the 5′-flanking region of the human insulin gene (Bell *et al.*, 1982) (Figure 5). Third, low stringency (3 x SSC at 60°C) hybridization of a 169-bp *Hinf*I fragment, containing the entire human myoglobin repeat region, to Southern blots of human DNA detected several hybridizing components in addition to the myoglobin gene sequence (data not shown).

The mechanism of transposition of such a simple repeated element is not clear, although a perfect 12-bp palindrome just 3′ to the repeat block in the myoglobin gene (Figure 3) might be implicated, either by destabilising the neighbouring DNA helix and opening up an insertion site, or by serving as a recognition sequence for the endonucleolytic activity needed to generate the double-stranded cleavage required for insertion.

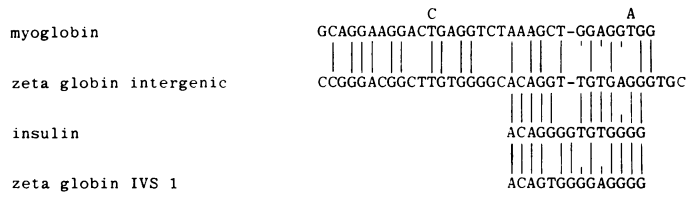


Fig. 5. Sequence homologies between dispersed tandem-repetitive DNA sequences in the human genome. Canonical repeat units are shown for the 33-bp tandem repeat in the human myoglobin intron 1, the 36-bp repetitive region between the human ζ- and $\psi\zeta$-globin genes (Goodbourn *et al.*, 1983), the 14-bp repeat element in the first intron of the human ζ-globin gene (Proudfoot *et al.*, 1982) and the 14-bp repeat in the 5′-flanking region of the human insulin gene (Bell *et al.*, 1982). Similarities between the ζ-globin and insulin repeats have been noted by Goodbourn *et al.* (1983) and Proudfoot *et al.* (1982). The optimal alignment between the ζ-globin intergenic repeats was determined by computer, without permitting any micro-deletions/insertions. The significance of the optimum match (18/33 bp) was determined as follows. Either completely random 33-bp sequences, or random permutations of the 33-bp myoglobin repeat, were optimally aligned with all cyclic permutations of the ζ-intergenic repeat or its complement. The proportion of sequences which gave an optimal alignment of $\geq$18/33 bp was scored. By these tests, the myoglobin - ζ intergenic alignment is significant (p = 0.006 for a random sequence, 10 000 trials; p = 0.04 for permutations of the myoglobin sequence, 5000 trials), and is only partially dependent on the G-rich base composition of these sequences.

There are precedents for the existence of dispersed tandem-repetitive sequences in the human genome, for example, the 319-bp repeat of the *Hinf* family (Shimizu *et al.*, 1983). If the 33-bp repeat in the human myoglobin gene is transposable, then it might provide a probe for tandem repetitive regions of the human genome which are frequently associated with multiallelic polymorphism due to repeat number variation (see Goodbourn *et al.*, 1983). We are currently isolating other members of the 33-bp repeat 'family' to test the feasibility of this approach.

Materials and methods

General methods

High mol. wt. human DNA was prepared from fresh placenta by the method of Jeffreys (1979). Poly(A)$^+$ RNA was isolated from adult human gastrocnemius muscle removed during amputation (Wood *et al.*, 1982). Southern blot hybridisations were performed in the presence of dextran sulphate (Jeffreys *et al.*, 1980), Northern blots by the method of Thomas (1980), and a human genomic DNA library constructed by the method of Jeffreys *et al.* (1982). Restriction fragments used for hybridization probes were purified by agarose gel electrophoresis onto DE81 paper (Dretzen *et al.*, 1981). ^{32}P-Labelling of DNA fragments was performed by a modification of the nick-translation protocol of Jeffreys *et al.* (1980); 80 ng DNA in 5 μl water were heated to 100°C for 3 min, chilled on ice and labelled with ^{32}P in a conventional nick-translation mix. >50% incorporation of [α-^{32}P]dCTP was observed, even with impure substrates which failed to label in a normal nick-translation reaction. ^{32}P-Labelled DNA (structure unknown, but probably a partially hairpinned single-stranded DNA) behaved indistinguishably from normal ^{32}P-labelled nick-translated DNA of comparable specific activity in filter hybridizations.

DNA sequencing

Human myoglobin gene subclones in pAT 153 were sheared by sonication (Deininger, 1983), end-repaired using the Klenow fragment of DNA polymerase I and fragments 400–900 bp long recovered by agarose gel electrophoresis onto DE81 paper. Fragments were blunt-end ligated into the *Sma*I site of M13mp8 RF DNA and transfected into *Escherichia coli* JM103 (Messing and Vieira, 1982). White plaques were toothpicked and grown for 5 h on *E. coli* JM103, and recombinant phage from 1.5 ml culture were collected by PEG precipitation into 0.1 ml 10 mM Tris-HCl, 0.1 mM EDTA (pH 8.0). Clones were screened for the required inserts by spotting 1 μl phage plus 1 μl 3 M NaCl, 0.2 M NaOH, 0.1% SDS, 0.1% bromophenol blue onto a gridded nitrocellulose filter (Schleicher and Schull). After 15 min drying at room temperature, the filter was rinsed in 3 x SSC, baked at 80°C for 4 h and hybridized with ^{32}P-labelled probe DNA (Jeffreys *et al.*, 1980). DNA was purified from positive phage and DNA sequences determined by the dideoxynucleotide chain-termination method of Sanger *et al.* (1977) as modified by Biggin *et al.* (1983), using the 17-mer primer (Duckworth *et al.*, 1981) generously provided by Dr. P. Meacock (ICI, Leicester). 6% polyacrylamide, 8 M urea buffer gradient sequencing gels 40 cm x 0.25 mm were run at 1500 V for 3 h (mean length of readable sequence, 240 bases) or 6 h (360 bases) before fixation, drying and autoradiography (Biggin *et al.*, 1983). Sequencing data were assembled with the aid of a Digital PDP 11/44 computer using programs developed by Staden (1980).

Acknowledgements

We are grateful to Drs. Alister Hawkins and Jeff Almond (Leicester) for much helpful advice on M13 cloning and sequencing, and to Simon Walker (Sheffield) for human tissue specimens. A.J.J. is a Lister Institute Research Fellow, and this work was supported by a grant from the MRC and a Fellowship in the European Science Exchange Programme, Royal Society/CNRS (to A.B.).

References

Bell,G.I., Selby,M.J. and Rutter,W.J. (1982) *Nature,* **295**, 31-35.
Biggin,M.D., Gibson,T.J. and Hong,H.F. (1983) *Proc. Natl. Acad. Sci. USA,* **80**, 3963-3965.
Blanchetot,A., Wilson,V., Wood,D. and Jeffreys,A.J. (1983) *Nature,* **301**, 732-734.
Butler,P.J. and Jones,D.R. (1982) *Adv. Comp. Physiol. Biochem.,* **8**, 179-364.
Czelusniak,J., Goodman,M., Hewett-Emmett,D., Weiss,M.L., Venta,P.J. and Tashian,R.E. (1982) *Nature,* **298**, 297-300.
Deininger,P. (1983) *Anal. Biochem.,* **129**, 216-223.
Dierks,P., Van Ooyen,A., Cochran,M.D., Dobkin,C., Reiser,J. and Weissmann,C. (1983) *Cell,* **32**, 695-706.
Dover,G. (1982) *Nature,* **299**, 111-117.
Dretzen,G., Bellard,M., Sassone-Corri,P. and Chambon,P. (1981) *Anal. Biochem.,* **112**, 295-298.
Duckworth,M.L., Gait,M.J., Goelet,P., Hong,G.F., Singh,M. and Titmas, R.C. (1981) *Nucleic Acids Res.,* **9**, 1691-1706.
Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz,R.A., DeRiel,J.K., Forget,B.G., Weissman,S.M., Slightom,J.L., Blechl,A.E., Smithies,O., Baralle,F.E., Shoulders,C.C. and Proudfoot, N.J. (1980) *Cell,* **21**, 653-668.
Goodbourn,S.E.Y., Higgs,D.R., Clegg,J.B. and Weatherall,D.J. (1983) *Proc. Natl. Acad. Sci. USA,* **80**, 5022-5026.
Grosveld,G.C., de Boer,E., Shewmaker,C.K. and Flavell,R.A. (1982) *Nature,* **295**, 120-126.
Hayashida,H. and Miyata,T. (1983) *Proc. Natl. Acad. Sci. USA,* **80**, 2671-2675.
Hunt,T.L., Hurst-Calderone,S. and Dayhoff,M.O. (1978) in Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure,* National Biomedical Research Foundation, Washington, DC, pp. 229-251.
Jeffreys,A.J. (1979) *Cell,* **18**, 1-10.
Jeffreys,A.J. and Harris,S. (1982) *Nature,* **296**, 9-10.
Jeffreys,A.J., Wilson,V., Wood,D., Simons,J.P., Kay,R.M. and Williams, J.G. (1980) *Cell,* **21**, 555-564.
Jeffreys,A.J., Barrie,P.A., Harris,S., Fawcett,D.H., Nugent,Z.J. and Boyd, A.C. (1982) *J. Mol. Biol.,* **156**, 487-503.
Kimura,M. (1981) *J. Mol. Evol.,* **17**, 110-113.
Kimura,M. and Ohta,T. (1972) *J. Mol. Evol.,* **2**, 87-90.
Kohne,D.E. (1970) *Q. Rev. Biophys.,* **3**, 327-375.
Konkel,D.A., Maizel,J.V. and Leder,P. (1979) *Cell,* **18**, 865-873.
Kreitman,M. (1983) *Nature,* **304**, 412-417.
Lenfant,C. (1973) *Am. Zool.,* **13**, 447-456.
Loenen,W.A.M. and Brammar,W.J. (1980) *Gene,* **20**, 249-259.
Longo,L.D., Koos,B.J. and Power,G.G. (1973) *Am. J. Physiol.,* **224**, 1032-1036.
Maeda,N., Bliska,J.B. and Smithies,O. (1983) *Proc. Natl. Acad. Sci. USA,* **80**, 5012-5016.
Martin,S.L., Zimmer,E.A., Davidson,W.S., Wilson,A.C. and Kan,Y.W. (1981) *Cell,* **25**, 737-741.
Messing,J. and Vieira,J. (1982) *Gene,* **19**, 269-276.
Moschonas,N., de Boer,E. and Flavell,R.A. (1982) *Nucleic Acids Res.,* **10**, 2109-2120.
Nevo,E. (1979) *Annu. Rev. Ecol. Syst,* **10**, 269-308.
Perler,F., Efstratiadis,A., Lomedico,P., Gilbert,W., Kolodner,R. and Dodgson,J. (1980) *Cell,* **20**, 555-566.
Proudfoot,N.J., Gil,A. and Maniatis,T. (1982) *Cell,* **31**, 553-563.

Radding,C.M. (1978) *Annu. Rev. Biochem.*, **47**, 847-880.
Romero-Herrera,A.E. and Lehmann,H. (1973) *FEBS Lett.*, **31**, 175-180.
Romero-Herrera,A.E. and Lehmann,H. (1974) *Proc. R. Soc. Ser. B.*, **186**, 249-279.
Romero-Herrera,A.E., Lehmann,H., Joysey,K.A. and Friday,A.E. (1973) *Nature*, **246**, 389-395.
Romero-Herrera,A.E., Lehmann,H., Joysey,K.A. and Friday,A.E. (1978) *Proc. R. Soc. Ser. B.*, **283**, 61-163.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Shimizu,Y., Yoshida,K., Reu,C.-S., Fujinaga,K., Rajagopalan,S. and Chinnadurai,G. (1983) *Nature*, **302**, 587-590.
Staden,R. (1980) *Nucleic Acids Res.*, **8**, 3673-3694.
Thomas,P.S. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5201-5205.
Tipler,T.D., Edwards,Y.H. and Hopkinson,D.A. (1978) *Ann. Hum. Genet.*, **41**, 409-418.
Twigg,A.J. and Sherratt,D. (1980) *Nature*, **283**, 216-218.
Van Den Berg,J., Van Ooyen,A., Mantei,N., Schamböck,A., Grosveld,G., Flavell,R.A. and Weissmann,C. (1978) *Nature*, **276**, 37-44.
Weiss,E.H., Mellor,A., Golden,L., Fahrner,K., Simpson,E., Hurst,J. and Flavell,R.A. (1983) *Nature*, **301**, 671-674.
Wilson,V., Jeffreys,A.J., Barrie,P.A., Boseley,P.G., Slocombe,P.M., Easton,A. and Burke,D.C. (1983) *J. Mol. Biol.*, **166**, 457-475.
Wittenberg,J.B. (1970) *Physiol. Rev.*, **50**, 559-636.
Wood,D., Blanchetot,A. and Jeffreys,A.J. (1982) *Nucleic Acids Res.*, **10**, 7133-7144.

Received on 14 November 1983