A newly discovered human alpha globin gene

Sung-Ho Goh¹, Y. Terry Lee¹, Natarajan V. Bhanu¹, Margaret C. Cam², Richard Desper³, Brian M Martin⁴, Ramy Moharram⁴, Robert B. Gherman⁵ and Jeffery L. Miller¹,*.

¹Molecular Medicine Branch, ²Microarray Core Facility, NIDDK, NIH, Bethesda, ³NCBI, NIH, ⁴Laboratory of Neurotoxicology, NIMH, NIH, ⁵National Naval Hospital, Bethesda, Maryland 20892, U.S.A.

Short Title: Human mu globin

Word count: 4,505 words

Scientific Heading: New human alpha-like globin gene

*To whom correspondence should be addressed.
Jeffery L. Miller
Molecular Medicine Branch
National Institute of Diabetes, Digestive and Kidney Disease
National Institutes of Health
10 Center Dr. Bg.10/Rm.9B17
Bethesda, MD 20892.
Phone 1-301-480-1908
FAX: 1-301-435-5148
jm7f@nih.gov
Abstract
A previously undefined transcript with significant homology to the pseudo-alpha2 region of the alpha globin locus on human chromosome 16 was detected as part of an effort to better define the transcriptional profiles of human reticulocytes. Cloning and sequencing of that transcript (GenBank: AY698022; named mu globin) revealed an insert with a 423 nucleotide open reading frame. BLASTP and ClustalW and phylogenetic analyses of the predicted protein demonstrated high level of homology with the avian alpha-D globin. In addition, the heme- and globin-binding amino acids of mu globin and avian alpha-D globin are largely conserved. Using quantitative real-time PCR, mu globin was detected at a level of approximately 0.1% that measured for alpha globin in erythroid tissues. Erythroid-specific expression was detected by Northern analysis, and maximal expression during the erythroblast terminal differentiation was also detected. Despite this highly regulated pattern of mu globin gene transcription, mu globin protein was not detected by mass spectrometry. These results suggest the human genome encodes a previously unrecognized globin member of the avian alpha-D family that is transcribed in a highly regulated pattern in erythroid cells.
Introduction

The globin genes and their products have been intensively investigated over the last fifty years. Those studies led to the description of structural and regulatory elements that are useful for the recognition and comparison of hundreds of globin gene family members. The divergence of ancestral alpha and beta globin genes is estimated to have occurred 500 million years ago. Those genes subsequently evolved and were modified by a variety of genetic processes including duplication events\textsuperscript{1}. In humans, the alpha globin gene family resides on chromosome 16p13.3, and is composed of a cluster of three genes (zeta2-alpha2-alpha1) with protein products that bind heme and assemble into hemoglobin. Transcription of the zeta2 gene is silenced during fetal life, and the two alpha genes are expressed in a balanced fashion for the remainder of ontogeny. In addition to those three genes, concerted efforts in 1980s led to the discovery of other alpha-like sequences. The downstream region of the alpha locus contains an unusual gene named theta globin that generates no detected globin protein in humans\textsuperscript{2,3}. Theta globin gene transcription is regulated, and the transcripts contain no obvious defects to explain the lack of detectable protein in erythroid tissues\textsuperscript{2-5}. Three pseudo-globin genes (pseudo-zeta\textsuperscript{6}, pseudo-alpha\textsuperscript{27}, and pseudo-alpha\textsuperscript{18}) were also identified in the alpha globin locus.

During the last two years, investigators have begun a transition toward post-genomic approaches to basic and clinical research. Hypotheses are now generated with the knowledge of whole genomic DNA sequences\textsuperscript{9}, full-length cDNA collections\textsuperscript{10} and millions of Expressed Sequence Tags (EST)\textsuperscript{11} from humans and other species. Comparisons of DNA and RNA sequences with advanced bioinformatics analyses\textsuperscript{12}
have become essential. Hematology is ideally suited for this type of genome-based research based upon the ease with which purified populations of hematopoietic cells are isolated. We hypothesized that human reticulocytes contain sufficient mRNA from the terminal stages of differentiation for the study of globin gene expression patterns in high throughput. Levels of globin mRNA detection and differences in globin gene transcription between cord and adult blood were studied to determine the potential of this approach for clinical assessments of hemoglobinopathies and hemoglobin switching. Using oligonucleotide arrays, significantly different globin transcription patterns were found in cord and adult blood samples. Evidence for transcription of the major globin genes was clearly demonstrated. Surprisingly, we also identified transcription from the genomic region previously thought to encode the pseudo-alpha2 gene. The source of that transcription is characterized in this report as a previously unrecognized globin gene.

Materials and Methods

Preparation of Reticulocytes RNA
Blood was collected from normal adult donors and placental umbilical cords. All cells were collected according to approved human subjects’ guidelines. The blood samples were centrifuged, and the plasma and buffycoat layers were removed. The packed red blood cells were diluted with 4 volumes of 1 x PBS and filtered thorough two consecutively linked RCXL2 High Efficiency Leukocyte Reduction Filter (PALL, NY). Platelets were removed by repeated low-speed centrifugation. RNA was isolated with
Trizol LS (Invitrogen, CA) and then treated with DNaseI to degrade residual genomic DNA in the RNeasy Mini Kit (Qiagen) according to the manufacturer’s protocol. For erythropoiesis assays, peripheral blood CD34+ cells from healthy donors were cultured for 14 consecutive days in erythropoietin-containing medium as previously described13.

**Microarray Data Analysis**

Microarray analyses were performed using 5 ug of total RNA from each samples with one cycle of complementary RNA amplification according to the Affymetrix protocol. After the hybridization and washing steps were performed, the microarray chips were scanned using MAS 5.0 software. The collected data were analyzed using Partek Pro 6.0 software (Partek, MO). The expression levels were clustered and displayed by Spotfire DecisionSite 8.0 (Spotfire, MA). A complete description of the 28 array data sets is being prepared as a separate publication.

**Cloning Full-length Coding Sequences**

Fifty nanograms of first strand cDNA made from 1 ug of adult blood reticulocyte total RNA was amplified with the forward (5’-CCA TGC TCA GCG CCC AGG AG-3’) and reverse (5’-AGC ACA GGG CTC AGC GGT ATT TTT C-3’) primers using BD Advantage-GC cDNA PCR kit (BD Clontech, CA) with the cycle conditions as follows; 94°C pre-denaturation 3 min, 94°C 30 sec and 68°C annealing and extension 3 min for 30 cycles. The amplified PCR product was purified with MinElute PCR cleanup kit (Qiagen, CA) and cloned into pcDNA3.1-V5-6His and pCR2.1 vector (Invitrogen, CA).
**Northern Blotting**

The full-length cDNA clone insert was cleaved by restriction enzyme digestion and gel extraction using MinElute GelExtraction Kit (Qiagen, CA). Five hundred nanograms of insert was labeled with alpha\(^{32}\)P [dCTP] (Amersham Phamacia, CA), and purified from unincorporated nucleotides using a G-50 column. The labeled probe was hybridized onto a nylon membrane containing 10 ug of total RNA from cord blood reticulocyte (5 pooled samples), adult blood reticulocyte (5 pooled samples), fetal liver (BD Clontech, CA) and adult bone marrow (BD Clonetech, CA) on each lane at 43.5°C. The hybridized membrane was washed and exposed on BioMax MS film (Kodak).

**Quantitative Real-time PCR**

For the quantitative PCR, the sequence specific primers and probe were designed to span the border between exons 2 and 3 of the alpha (Forward primer: 5’-GGG TGG ACC CGG TCA ACT T-3’; Reverse primer: 5’-GAG GTG GGC GGC CAG GGT; Probe: FAM-5’-AAG CTC CTA AGC CAC TGC CTG CTG-3’-TAMRA) and mu globin mRNA (Forward primer: 5’-GCG TGG ACC CAG CCA ACT T-3’; Reverse primer: 5’-CAG GTG GGA GGC CAG CAC-3’; Probe: FAM-5’-TCC GCT GCT AAT CCA GTG TTT CCA C-3-TAMRA). Copy numbers were calculated by comparison with standard curves. The specificity of each primer and probe was defined by alpha globin and mu globin cDNA templates. For each PCR reaction, 5 ng of cDNA made from pooled total RNA was mixed with 2x TaqMan Master Mix (Applied Biosystems, CA) and 10 pmoles of each primer and FAM/TAMRA-labeled probe, and amplified using the ABI 7700 Sequence Detection System (Applied Biosystems, CA). The results were analyzed by Sequence
Detector 1.7 software (Applied Biosystems, CA).


H4 ProteinChips® and calibration standard molecules for the SELDI-TOF mass spectrometer were purchased from Ciphergen Biosystems Inc. (Fremont, CA). Sinapinic acid (SA) was obtained from Sigma (St. Louis, MO). The SELDI-TOF mass spectrometer was externally calibrated using the [M + H]^+ ion peaks of somatostatin at 1637.9 m/z, insulin beta-chain at 3495.9 m/z, human recombinant insulin at 5807.6 m/z and hirudin at 7033.6 m/z. All mass spectra were recorded in the positive-ion mode using a Ciphergen PBS Ilc ProteinChip Array mass spectrometer with time-lag focusing\(^\text{14}\). Prior to SELDI-TOF MS analysis, the H4 ProteinChip was prewashed with 10% aqueous acetonitrile containing 0.1% trifluoroacetic acid (TFA). Upon drying, one microliter of sample was applied to the ProteinChip, air dried and washed with 5% aqueous acetonitrile. After drying, one microliter of matrix (saturated SA in 50% aqueous acetonitrile containing 0.1% TFA) was added to each feature of the ProteinChip array. The data were analyzed using the computer software provided by the manufacturer and are reported as mass averages\(^\text{15}\).

**Bioinformatic analyses**

The comparisons of mRNA sequences of mu globin were performed by BLAST\(^\text{16}\) and alignment to the Human Genome by BLAT\(^\text{17}\). For promoter analyses, the 200 bp upstream from the translation start site were examined using PromoterInspector\(^\text{18}\) with
default setting. Phylogenetic analyses were performed by maximum parsimony using mu globin and 291 known alpha-like globin protein sequences deposited in GenBank. A complete alignment with gap was performed using ClustalX software. Aligned sequences were input to PAUP (version 4.0b10 for UNIX), which defined the maximum parsimony tree, using the heuristic search command. The maximum parsimony (MP) tree was chosen after PAUP had not improved the score after several hours of searching. The final tree was drawn by PhyloDraw software. Maximum likelihood, NJ, BIONJ, least-squares, and balanced minimum evolution analyses were performed and demonstrated similar results.

Results

Microarray comparison of globin gene expression.

To compare the mRNA profiles in the reticulocytes circulating at the time of birth with those in adults, high-throughput arrays were generated from the blood of 28 separate donors (14 cord blood; 14 adult blood). The platelets were removed from the samples by low-speed centrifugation and nucleated cells by leukocyte reduction filtering. The samples were analyzed using Affymetrix from HG-U133 A and B chips, and the expression of 44,229 probe sets was ranked. The focus of this report is the globin genes; a description of the other probe sets will be provided in a separate manuscript. As expected, the globin gene transcripts achieved high ranking due to their abundance in reticulocytes. The rank of signal intensities for alpha2, alpha1, beta, A-gamma and G-gamma globins in adult blood reticulocytes were 1st, 2nd, 3rd, 10th and 8th, and the
rank in the cord blood reticulocytes were 1st, 2nd, 3rd, 4th and 5th, respectively. The higher ranking of A-gamma compared with G-gamma in cord blood was unexpected because it is known that G-gamma represents about 70% of the total gamma chains at birth²². In comparison with alpha, beta, and gamma transcripts, the levels of epsilon, zeta and theta globin were low in all the samples, and delta globin was reduced in the cord blood samples.

In addition to the expected globins, we identified a probe set (240336_AT) described by the Affymetrix software as having homology with a hemoglobin based blood substitute (Rhb1.1)²³. The expression rank of that probe set was 21st of 44,229 probes in adult and 11th of 44,229 probes in cord blood reticulocytes. To place the expression pattern of 240336_AT probe in the context of the other globin probe sets, the signal intensities were clustered by unsupervised hierarchical clustering (Figure 1). The clustered arrangement of the signal intensities corresponding to the nine globin gene probes and 240336_AT from 14 cord blood and 14 adult blood samples are shown. Cord blood and adult blood samples were segregated appropriately on the basis of switching patterns of intensity. The alpha1, alpha2, and beta globin genes clustered together according to their high intensity in all the samples. The gamma genes also co-clustered, and they were expressed within the same range as beta globin in cord blood. As shown, expression of the probe 240336_AT did not cluster with any other globin probe. The average signal intensity of that probe was higher than those of epsilon, theta, zeta or delta but lower than gamma, beta or alpha globin. The pattern of 240336_AT expression was variegated between donors with decreased mean expression in adult blood compared with cord blood samples.
Bioinformatic Analyses and Cloning of Reticulocyte Mu Globin.

Public sequence data describing the human genome, full-length cDNA, and Expressed Sequence Tags (EST) provided a clear path for further investigation of the 240336_AT probe. The probe set was designed using over twenty-five million EST sequences deposited on dbEST. The template EST sequence (GenBank ID: BE244453) that aligns with the 240336_AT probe was identified by using the reference sequence for a BLAT search. Surprisingly, 240336_AT aligned in the same region as the pseudo-alpha2 gene rather than other gene regions (Figure 2A). Unlike the pseudo-alpha2 globin nucleotide sequences, the 240336_AT EST sequences aligned to generate a gene structure familiar to the other globin genes. Based upon this bioinformatic comparison, a full-length cDNA was cloned from reticulocyte RNA to generate a 506 bp transcript encoding the gene probed by 240336_AT. An additional four bases were identified by 5’ RACE. The 510 bp reticulocyte-derived sequence was originally deposited in the GenBank in July 2004 (AY698022, NM_001003938) and named mu (µ) due to the smaller size of the predicted globin product (141 amino acids compared with 142 for the other human alpha globin genes).

Additional bioinformatic analyses of the mu globin sequence are shown in Figure 2. The mu gene aligned within the 3’ region of pseudo-alpha2 on chromosome 16p13. The gene contained a Kozak sequence (24 bp downstream of the transcriptional start site) indicating the predicted translation initiation site of the mu globin gene. While that Kozak sequence (CGCCATATG) was not found in the other human globin genes, it is encoded in approximately 5-10% of vertebrate genes including the duck alpha-D
globin (gil62724l). The mu globin sequence also possesses a poly(A) signal sequence (AATAAA) located 35 base pairs downstream of the translational stop codon in the 3'UTR. The alpha globin 3'UTR is thought to increase the mRNA stability by forming a specialized secondary structure that is complexed with alpha globin poly(C) binding protein (alphaCP)\textsuperscript{25}. The predicted secondary structure of the mu globin 3'UTR suggests that mu globin mRNA may be less stable due to decreased resistance against degradation by endo- or exo-nucleases (data not shown).

The mu globin (HBM) sequence was aligned to the alpha globin mRNA using ClustalW\textsuperscript{20} and its overall similarity was 59% (302/510 bases). The predicted protein sequence from the open reading frame was identified in GenBank as NP_001003938. We aligned this sequence to the human alpha globin protein using ClustalW. It showed incomplete conservation of heme binding and alpha-beta globin contact sites. As shown in Figure 2B, the mu globin promoter region (upstream 200 bp) was also examined using PromoterInspector\textsuperscript{18}. Unlike the alpha and zeta promoters, mu globin promoter did not contain a CAAT motif. A muscle TATA (TATAGA) core sequence was identified 60 bp upstream of the ATG. Erythroid Kruppel like factor (EKLF) and GATA binding factor 1 (GATA1) binding sites were identified at -70 and -82 bp, respectively. The mu globin gene also contained a hypoxia inducible factor (HIF) binding site commonly associated with higher-affinity hemoglobins\textsuperscript{26}, but not identified in other human globin gene promoters.

**Mu Globin Gene Expression**

In addition to array-based assays, mu globin gene expression was examined by
Northern analysis (Figure 3). Hybridizations containing ten microgram of total RNA for each four erythroid tissues (cord blood reticulocytes, adult blood reticulocytes, fetal liver, and bone marrow) were performed. As expected, alpha globin expression was detected at similar levels in both of adult blood reticulocytes and cord blood reticulocytes. Equivalent bands were also detected in fetal liver and adult bone marrow (Figure 3A). Compared with bone marrow, no detectable mu globin signal was identified on non-erythroid tissues (Figure 3C). The expression of mu globin in cord blood reticulocytes was approximately five times higher than that of adult blood reticulocytes. The mu globin expression on fetal liver was also higher than adult bone marrow (Figure 3B). Consistent with the array data, these data suggest that the levels of mu globin mRNA in erythroid tissues decrease during the postnatal development.

Due to concerns that the hybridization signals detected by array and Northern analyses may be non-specific, quantitative PCR was performed (Figure 4). Sequence specific primers and a probe spanning the mu globin exon2 and exon3 boundary were designed to avoid the amplification of unprocessed RNA, genomic DNA or other alpha globin transcripts. In confirmation of the array and Northern data, mu globin mRNA levels were significantly higher in cord blood reticulocytes than adult blood reticulocytes (1.71x10(5) ± 9.51x10(4) copies/ng cDNA in cord blood and 2.17x10(4) ± 6.84x10(3) copies/ng cDNA in adult blood; p<0.0002). This pattern of decreased adult expression was also noted by the comparison of fetal liver and adult bone marrow (3.04x10(4) ± 1.68x10(3) copies/ng cDNA in fetal liver and 1.15x10(4) ± 1.19x10(3) copies/ng cDNA in bone marrow; p<0.0002) (Figure 4A). Alpha globin amplification was performed for comparison. While the expression of alpha globin was two to three orders of magnitude
higher than mu globin, the levels of alpha globin mRNA were equivalent in fetal and adult tissues (Figure 4B). Therefore, at the transcriptional level the mu globin expression is only approximately 0.1% of the normal adult alpha globin.

Mu globin expression during the erythropoiesis was examined and compared with alpha globin using cultures of adult CD34+ cells. Mu globin was not detected above background levels until day 4 when the large, immature erythroblasts began to appear in culture. After day 4, a rapid rise in mu globin was detected until day 10 followed by an equally rapid loss as the cells underwent terminal maturation. This pattern was similar to that identified for alpha globin, but the alpha globin peak occurred later during the culture period on day 12. The overall expression level of mu globin was 100 fold less than that of alpha globin throughout the culture period (compare scales on Figures 4C, 4D).

**SELDI-TOF Mass-Spectrometry of Red Blood Cell Lysate**

Based upon the highly regulated transcription of the mu globin gene, assays were developed to determine whether significant quantities of mu globin proteins are expressed in circulating erythrocytes. Importantly, the literature provided little evidence for the existence of this protein in humans or other mammals. Efforts to raise mu-specific antibodies using peptide sequences of mu globin have not been successful to date (data not shown). Therefore, cord blood and adult blood lysates were directly examined using mass spectroscopy using the SELDI-TOF MS technology. This method is one of the more sensitive proteomic detection tools and may have some advantages over matrix-assisted laser desorption/ionization (MALDI). A detection sensitivity of
20 ng per sample was determined using serially diluted hemoglobin standards. In parallel assays, cation exchange HPLC peaks\textsuperscript{13} were not detected at levels below 300 ng per sample. SELDI analysis of cell lysates from three cord and adult bloods revealed the relative amount of alpha, beta, and gamma protein in each sample over the 15kDa to 16.5kDa mass range (Figure 5). We predicted the molecular weight of globin protein without start the methionine residue due to the adjacent valine residue\textsuperscript{28}. The alpha globin chain (15,126 Dalton) was identified as the major peak in adult and cord blood samples. The beta (15,867 Dalton) and gamma (15,996 Dalton) globin peaks were detected in cord blood samples, but the gamma peak was not detected in the adult lysates. No peak was seen at the expected size of mu globin (15,487 Dalton) in any samples, i.e. no significant peaks with sufficient signal to noise ratio were observed. This suggested the absence of measurable quantities of mu globin in the lysates. The identities of the other peaks demonstrated by this method of globin analysis are being studied separately (data not shown).

**Homology Comparisons and Phylogenetic Analyses**

At the nucleotide level, several mammalian orthologues were identified by EST alignments, but no significant homologies were noted with avian or reptilian mRNA examined to date (data not shown). The predicted protein sequence of mu globin was also searched against GenBank nr database using BLASTP default parameters. The search demonstrated that the predicted mu globin protein most closely aligned with the avian alpha-D globin chain of bar-headed goose (GenBank ID gi|70296) with a similarity of 55% (78/141 aa). A lower similarity of mu globin with the human alpha globin chain
was found (64/141 aa). Alignments of the heme binding sites, alpha1-beta1 and alpha1-beta2 contact sites were also studied. The heme binding homologies were equivalent between mu, the human alpha chain and the avian alpha-D chain (84% [16/19]). However, the alpha1-beta1 and alpha1beta2 contact sites demonstrated considerably more homology between mu globin and the avian alpha-D globin. Interestingly, mu globin and the avian alpha-D globin chain have the same length of 141 amino acid residues.

Based upon the genetic and predicted protein homologies between mu globin and avian alpha-D globin, a more complete phylogenetic comparison was performed. A total of 291 alpha-like globin protein sequences were collected from GenBank. These amino acid sequences were used for the construction of phylogenetic tree using maximum parsimony algorithm. The constructed tree demonstrated several clustered globin families (Figure 6). As expected, the human alpha chain was clustered with other mammalian alpha chains, and human zeta globin clustered with the mammalian zeta globin group. The theta globin group clustered most closely to the alpha family. In comparison, human mu globin clustered with the avian and reptile alpha-D chains at the greatest distance from the alpha globin cluster (Figure 6). No other mammalian globin within the group of 291 was placed within the alpha-D family.

Discussion

In this report, the availability of a fully-sequenced genome and high-throughput expression profiles led to a re-examination of the region of the alpha globin locus identified as pseudo-alpha27. A novel globin transcript was identified and named mu
globin. Mu globin is not a pseudogene because it is transcribed from a 510 nt genomic sequence that contains two introns, and it possesses an open reading frame encoding 141 amino acids without disruption\textsuperscript{29}. The mu globin gene also contains a promoter region with erythroid transcription factor binding sites, a 24 nt mRNA leader sequence, a Kozak sequence\textsuperscript{24}, and a functional polyadenylation signal (Figure 2A). In contrast, the originally described pseudo-alpha2 pseudogene\textsuperscript{7} was reported to contain no promoter due to the close proximity of its first exon with the zeta1 globin gene located just upstream. Pseudo-alpha2 was also reported to contain a mutated 5’ splice site for intron 1, several frameshift deletions and an insert in the second and third exons, as well as significant mutations in the polyadenylation signal region when compared to the alpha2 gene. When aligned with current maps, the originally reported genomic sequence for pseudo-alpha2 was found to contain several unmatched nucleotides or gaps (data not shown). Therefore, the pseudogene annotation may have resulted from DNA sequencing limitations that existed 20 years ago.

In the context of the 44,299 probes examined by microarray, relatively high-level expression from the mu globin probe set (ID: 240336_AT) was detected in erythroid cells \textit{in vivo} (ranked among the top 0.2% of reticulocyte transcripts). Among differentiating primary erythroblasts, mu globin gene expression was highly regulated with a pattern nearly identical to that of alpha globin. However, the level of mu globin mRNA represented only a small percentage of the amount of alpha globin mRNA in fetal and adult erythroid tissues. The differences in the levels of mu and alpha globin gene transcripts may be due to differences in their promoters or the stability of their mRNA. The delayed peak level of alpha globin transcripts compared to mu globin (culture day
12 versus 10 respectively) is consistent with increased stability of the alpha globin transcripts.

Since the expression pattern of the beta globin genes in humans during ontogeny generally follows their gene order in the cluster\textsuperscript{30}, the location of the mu globin gene between the embryonic and adult genes in the alpha globin cluster also suggests that the gene might be developmentally regulated. Northern analyses of other tissues revealed no detectable mu globin gene expression among non-erythroid tissues suggesting tissue specificity. Microarray, Northern, and quantitative PCR analyses consistently demonstrated significantly higher levels of mu globin in the fetal tissues compared to that found in the adult tissues. Hence, mu globin demonstrates erythroid-specific expression with a pattern during ontogeny similar to that described for the gamma globin genes in the beta cluster.

To determine the similarity of the predicted mu globin protein with other known alpha-like proteins, homology analyses were performed. The ORF-predicted protein from human mu globin was compared with 291 known alpha-like globin protein sequences deposited in GenBank using a maximum parsimony algorithm. BLASTP and ClustalW alignments were also performed including focused comparisons of heme, alpha1beta1, and alpha1beta2 binding. In each case, the predicted human mu globin was most closely related to the avian and reptilian alpha-D globins. Initial analyses of primate, bovine and porcine genomes or associated ESTs suggest alpha-D encoded open reading frames will soon be identified in a variety of mammals using comparative genomics. Hence, the mu globin gene may represent an expressed homologue of an ancient globin gene\textsuperscript{1}. The alpha-D globin family was first identified as an alpha chain of
hemoglobin M in chicken embryos. Alpha-D proteins assemble into high oxygen affinity hemoglobins among avian and reptilian species. In both species, hemoglobins containing alpha-D chain are expressed at all stages of ontogeny. Embryonic expression may be advantageous as an embryonic adaptation to hypoxia. Adult expression may also provide a survival advantage associated with the ability to respond to the hypoxic conditions of high-altitude flight or that associated with prolonged submersion. Thus, it was postulated that an evolutionary advantage for alpha-D globin expression may have arisen from hypoxic or anoxic conditions. The higher level of mu globin gene expression in cord blood is consistent with the utility of high oxygen affinity hemoglobins during fetal life. Unfortunately, the absence of detectable mu globin protein in human erythroid tissues makes it difficult to extrapolate the avian and reptilian functional data to humans.

Interestingly, most of the proteins predicted by recent genome mapping efforts have not yet been detected in nature. This may be due to low-level expression or low sensitivity of protein assays. In this context, our inability to detect mu globin is not unusual. However, it is extremely curious that mu globin is not the only gene in the human alpha globin cluster that lacks a detectable hemoglobin product. The theta globin gene is also transcribed, but no protein or hemoglobin product has been detected. Both genes are well conserved at the genomic level with appropriate splicing junctions and maintenance of open reading frames. Like mu globin, theta-globin also possesses a highly regulated pattern of transcription in erythroid cells. In addition, both genes demonstrate only fractional levels of transcription compared with the dominant alpha genes, and their deletion in humans has no reported effects upon the clinical
phenotype\textsuperscript{42, 43}. Therefore, the evolutionary conservation of the mu and theta globin genes in the absence of a hemoglobin product represents a biological paradox. This is especially puzzling when considering the low levels of gene products compared with the amount of globin that is required in humans for the transport of oxygen. In the case of mu globin, the protein homology between the human, avian and reptilian species in the absence of any significant genetic homology suggests a selective pressure to sustain the open reading frame. Therefore, it is not certain that this gene is evolving toward being a pseudogene. Instead, the possibility exists that this newly discovered, but ancient globin possesses a function for which high level protein expression is not required.

**Acknowledgements** We thank Drs. Douglas R. Higgs and Alan S. Schechter for helpful discussions. The ongoing cell processing support of the NIH Department of Transfusion Medicine is greatly appreciated.

**References**


2002;99:3005-3013.
14. Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser
desorption/ionization-time of flight-mass spectrometry. Electrophoresis
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search
18. Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in
large genomic sequences by PromoterInspector: a novel context analysis approach.
19. Chenna R, Sugawara H, Koike T, Lopez et al. Multiple sequence alignment with the
20. Rogers JS, Swofford DL. Multiple local maxima for likelihoods of phylogenetic trees:
22. Steinberg MH, Forget BG, Higgs DR, Nagel RL eds. Disorders of Hemoglobin:
Genetics, Pathophysiology, and Clinical Management. Cambridge University Press.
23. Kroeger KS, Kundrot CE. Structures of a hemoglobin-based blood substitute:

21


33. Hiebl I, Schneeganss D, Braunitzer G. High-altitude respiration of birds. The primary structures of the alpha D-chains of the Bar-headed Goose (Anser indicus), the Greylag Goose (Anser anser) and the Canada Goose (Branta canadensis). Biol.


2003;300:258-260.


Figure Legends

Figure 1. Unsupervised hierarchical clustering analysis. The microarray data for 14 cord blood and 14 adult blood reticulocytes samples probed by nine globin genes and the 240336_AT (mu globin) probe set. The data were clustered (unsupervised), and the signal scale is based upon the raw intensities for each probe set.

Figure 2. Bioinformatic analyses of mu globin. (A) Alignment of the mu globin mRNA and ORF-predicted protein with alpha globin cluster on chromosome 16p13. The locations (arrowheads) of the Kozak (CGCCAUGC), stop (UGA), and polyA signal (pA) relative to the three mu globin exons are shown. The protein is annotated according to globin helical domains44, and the residues involved in heme binding and alpha-beta globin contact are indicated below (conserved residues between human alpha globin and mu globin are denoted with colored boxes1). (B) Transcription factor binding site comparison of 200 bp promoter regions of the selected globin genes. The locations of erythroid transcription factor binding motifs are indicated in red, hypoxia-inducible factor motif in green, general transcription factor binding motifs in blue, and other binding motifs are shown as open boxes. (AP2: activator protein 2; EKLF: erythroid Krueppel like factor; GATA1: GATA binding factor 1; HIF: hypoxia-inducible factor; Muscle-INI: muscle initiation factor; MZF1: Myeloid zinc finger protein 1; NRSE: Neural restrictive silencer element; SRF: serum response factor)

Figure 3. Northern blot analyses of mu and alpha globin expression. The
hybridizations of erythroid tissues with alpha globin (A) and mu globin (B) probes are shown with molecular weight positions on the right. (CB-cord blood, AB-adult blood, FL-fetal liver, BM-adult bone marrow). (C) Multi-tissue Northern blot hybridized with alpha globin (upper) versus mu globin (lower) probes.

**Figure 4. Quantitative PCR analyses.** Total RNAs from erythroid tissues (CB-cord blood; AB-adult blood; FL-fetal liver; BM-adult bone marrow) were amplified for detection of mu globin (A) and alpha globin (B). One million copies of DNA encoding alpha globin (Alpha clone) and mu globin (Mu clone) were included as negative controls to demonstrate specificity. Also shown is a comparison of mu globin (C) and alpha globin (D) expression levels in differentiating erythroblasts during 14 days culture of CD34+ cells in erythropoietin (culture day on x-axis). All studies were performed in triplicate. Copy numbers were calculated by standard curve comparison. Mean values and standard deviation bars are shown. Asterisks indicate a significant change (t-test \( p<0.0002 \)) versus no significant change (N.S.) between fetal and adult erythroid tissues.

**Figure 5. SELDI-TOF-MS analyses of red blood cell lysates.** Globin proteins detected in erythrocytes from three individual cord blood (A) and adult blood (B) samples are displayed in the 14,000-17,000 Dalton mass range (top of each panel). Arrows indicate the alpha, beta and gamma globin proteins. The dotted lines indicate the expected location of mu if it were present.

**Figure 6. Phylogenetic tree constructed by maximum parsimony algorithm**
**method.** 291 alpha-like globin proteins and the predicted mu globin protein were clustered as described in the text. The general locations of the human alpha-like globin chains (open boxes) as well as a general description of the predominant globin type within in each cluster are provided.
Goh et al Figure 2

A

Chr 16p13

Mu-globin mRNA 510 nt

Predicted Protein 141 a.a

Heme binding

α1β1 contact

α1β2 contact

CGCCAUGC

Exon1

Exon2

Exon3

UGA

pA

0 kb

10 kb

20 kb

25 kb

B

Human Zeta

Human alpha1

Human alpha2

Human Mu

Human theta

Chicken alphaD

GATA1

CAAT

TATA

EKLF

EKLF

AP2

CAAT

TATA

AP2

AP2

AP2

AP2

NRSE

SRF

Muscle-IN

Muscle-IN

NRSE

GATA1
Goh et al. Figure 4

A

B

C

D

Copies per ng cDNA x 10(5)

Copies per ng cDNA x 10(6)

Copies per ng cDNA x 10(5)

Copies per ng cDNA x 10(7)

CB AB FL BM

Alpha clone

CB AB FL BM Mu clone

0 2 4 6 8 10 12 14

Days

Days

N.S.

N.S.

0 0.5 1.0 1.5 2.0

0 0.5 1.0 1.5 2.0

0 1.0 2.0 2.5

0 0.5 1.0 1.5 2.0
Goh et al. Figure 5

A  Cord Blood

B  Adult Blood

Relative Intensity

Mass / Charge (m/z)
Goh et al. Figure 6
A newly discovered human alpha globin gene

Sung-Ho Goh, Y T Lee, Natarajan V Bhanu, Margaret C Cam, Richard Desper, Brian M Martin, Ramy Moharram, Robert B Gherman and Jeffery L Miller

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include digital object identifier (DOIs) and date of initial publication.