PERSISTENCE OR RAPID GENERATION OF DNA LENGTH POLYMORPHISM AT THE ZETA-GLOBIN LOCUS OF HUMANS

BARBARA S. CHAPMAN,[1] KAREN A. VINCENT and ALLAN C. WILSON

Department of Biochemistry, University of California, Berkeley, California 94720

Manuscript received November 14, 1984
Revised copy accepted August 30, 1985

ABSTRACT

Extensive restriction mapping of 76 human genomic DNAs defines multiple sites of length and point mutation near the zeta-globin locus, which codes for an embryonic alpha-like globin chain. There are two major sites of DNA length variation: one in the intergenic region with three alleles and one in the first intron of the *zeta 1* gene with at least four alleles. Our mapping establishes that the intronic polymorphism is associated with a tandem array of short, repeated sequences. The length alleles occur in each of four human populations sampled, suggesting an ancient origin with persistence of several length alleles, or rapid regeneration of these particular variants. Four polymorphic restriction sites were also found; the frequency of polymorphic sites is comparable to that found in the human beta-globin gene region. Analysis of haplotypes indicates either that multiple recombinations have occurred near the 5′ end of the *zeta 1* gene or that this region is prone to recurrent length mutation.

THE use of recombinant DNA methods in genetic studies has led to the discovery of a new class of polymorphisms that result from rearrangements of chromosomal DNA. The origin and behavior of DNA length variants can be studied with a combination of molecular and population genetic techniques. We have applied such an approach to the length polymorphisms associated with the zeta-globin genes, located on the short arm of human chromosome *16* (McKUSICK 1982).

The zeta genes are embryonic members in a cluster of five genes encoding alpha-like subunits of hemoglobin, shown in Figure 1 (LAUER, SHEN and MANIATIS 1980). At least one allele of each gene has been cloned and sequenced together with extensive segments of flanking DNA (PROUDFOOT, GIL and MANIATIS 1982; PROUDFOOT and MANIATIS 1980; LIEBHABER, GOOSSENS and KAN 1980; MICHELSON and ORKIN 1980; SAWADA *et al.* 1983; GOODBOURN *et al.* 1983). The two zeta genes are located approximately 10 kb apart at the 5′ end of the alpha gene cluster. Their coding regions differ by three amino acid substitutions, one of which terminates translation of the *zeta 1* mRNA at codon

[1] Present address: Chiron Corporation, 4560 Horton Street, Emeryville, California 94608.

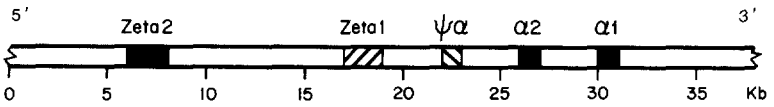

FIGURE 1.—The zeta-globin genes shown in relation to the three other genes in the human alpha-globin gene family (5′ to 3′, left to right). Two genes, shown as hatched boxes, encode nontranslatable mRNAs and have been designated pseudogenes (LAUER, SHEN and MANIATIS 1980; PROUDFOOT, GIL and MANIATIS 1982). To name the zeta genes we have adopted the convention of LAUER, SHEN and MANIATIS (1980).

6. Since there are no silent substitutions in the coding DNA, and only three substitutions in the unique sequence portions of the introns, inactivation of the *zeta 1* gene would appear to have occurred about 1 million years ago (PROUDFOOT, GIL and MANIATIS 1982; WILLARD *et al.* 1986). Although the *zeta 1* gene cannot produce a globin polypeptide, it remains transcriptionally active (PROUDFOOT, RUTHERFORD and PARTINGTON 1984).

There is extensive DNA polymorphism near the embryonic zeta genes of the cluster (CHAPMAN, VINCENT and WILSON 1981; HIGGS *et al.* 1981; CHAPMAN and WILSON 1982, 1983; GOODBOURN *et al.* 1983). These polymorphisms in restriction fragments appear to be generated by variation in the length of DNA sequences, as well as by mutation in restriction sites. Length variations detectable by genomic Southern blotting have been observed near other known loci, including the human insulin gene (BELL, KARAM and RUTTER 1981) and the *c-Ha-ras-1* gene (GOLDFARB *et al.* 1982). Nucleotide sequences of polymorphic regions in the insulin (BELL, SELBY and RUTTER 1982), *c-Ha-ras-1* (CAPON *et al.* 1983) and zeta-globin (GOODBOURN *et al.* 1983) loci have revealed tandem arrays of related oligonucleotides similar to the "minisatellite" sequences identified by JEFFREYS, WILSON and THEIN (1985).

The zeta-globin locus offers an opportunity to examine a highly polymorphic region using both molecular and population genetic approaches. Our restriction mapping of genomic DNA defines two regions of length polymorphism and four polymorphic restriction sites in the vicinity of the zeta-globin genes. We have mapped variations in the length of the *zeta 1* gene to the tandem array in intron I. To collect data on incidence in human populations, a panel of 76 human DNAs was screened by comparative Southern blotting. The stability of the length variation was examined in two small pedigrees and in cultured leukemic cells. In addition, haplotypes were constructed and used to search for recombination within the region. An unexpected outcome of this analysis is that DNA length alleles of the zeta-globin region may be ancient and stable, a possibility not previously considered. We discuss our results and those of others with respect to two models: one in which the length alleles are repeatedly generated by unequal recombination, and the other in which ancient alleles persist.

MATERIALS AND METHODS

DNA was prepared from blood or placenta of 75 human individuals and from an erythroleukemic cell line (K562) that can be induced to synthesize zeta globin. This

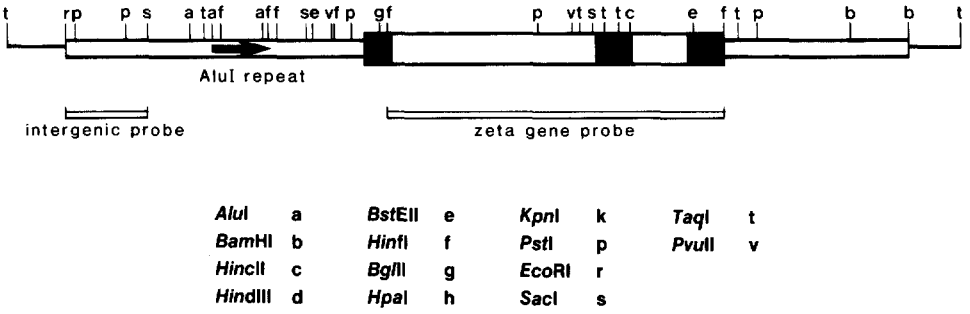

FIGURE 2.—Map of the cloned *zeta 1* gene (pBR *zeta*; LAUER, SHEN and MANIATIS 1980). The 5′ *Eco*RI site (r) is a synthetic linker. The enzymes *Alu*I (a) and *Hin*fI (f) were used to map the *Alu* family repeated sequence (arrow), and some additional *Hin*fI sites were identified in the 1.0-kb *Sac* (s) fragment in the 5′ flanking region of the *zeta 1* gene and within the gene. Other *Hin*fI and *Alu*I sites were not mapped. Black boxes indicate the exons of the *zeta 1* gene. The two probes chosen for Southern blot analysis are identified by brackets beneath. Regions t-r and b-t are sequences from the plasmid vector pBR322. Published and unpublished sequence data have confirmed the location of many sites shown in this map.

panel included DNAs identified by racial or geographic origin. Ten of these individuals are members of a Caucasian family and four are members of an Asian family. To our knowledge the remaining individuals are unrelated. The samples were digested with proteinase K in sodium dodecyl sulfate, followed by phenol extraction and ethanol precipitation as previously described (TAYLOR *et al.* 1974). The erythroleukemic cells were grown in medium RPMI 1640 with 10% fetal calf serum. DNA was prepared from K562 cultures after 20, 30, 50 and 80 passages and after induction of embryonic hemoglobin synthesis with various concentrations of hemin (RUTHERFORD, CLEGG and WEATHERALL 1979).

Human DNA was digested with restriction endonucleases (New England Biolabs) and was subjected to electrophoresis in 0.8% agarose gels. The DNA was transferred to nitrocellulose filters by a modification of the method of SOUTHERN (1975) and was hybridized to nick-translated probes (WAHL *et al.* 1979). Molecular weights of fragments were calculated from migration relative to *Hin*dIII, *Sma*I or *Bam*HI digests of lambda phage DNA. Single, double and triple digests were used to order fragments and determine distances between sites. Fragment sizes estimated from 0.8% gels vary by approximately 10% of the measured length. Errors in these estimates were reduced by averaging the sizes determined on several gels, and by the mapping procedure.

Two fragments were isolated from a cloned human *zeta 1* gene (LAUER, SHEN and MANIATIS 1980): (1) a 2.0-kb *Hin*fI fragment containing 85% of the *zeta 1* globin gene (zeta gene probe) and (2) a 0.5-kb *Eco*RI/*Sac*I fragment representing the 5′ extremity of the inserted human DNA (intergenic probe) shown in Figure 2. A probe for the human adult alpha-globin genes was prepared by *Mbo*II digestion of plasmid JW101 (WILSON *et al.* 1978). The purified fragments were labeled with ^{32}P-dCTP by nick translation (MANIATIS, JEFFREYS and KLEID 1975).

RESULTS

Restriction map of a cloned *zeta 1* gene: In order to select DNA fragments as probes in genomic blotting experiments, we mapped cleavage sites in clone pBR zeta (LAUER, SHEN and MANIATIS 1980), shown in Figure 2. A 0.5-kb *Eco*RI/*Sac*I fragment (r–s) from the 5′ end of the cloned human DNA was chosen as a probe for the intergenic sequences between *zeta 1* and *zeta 2*,

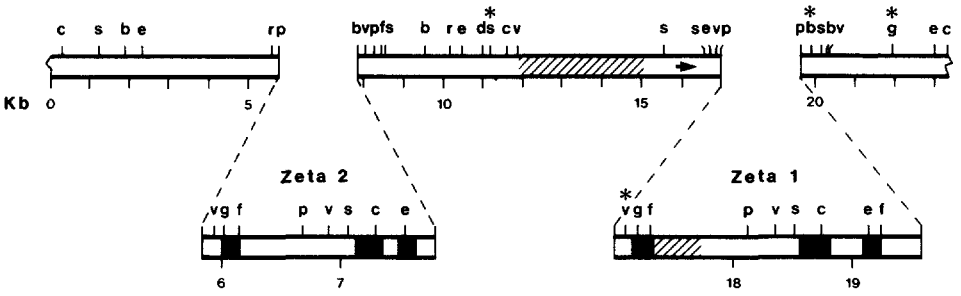

FIGURE 3.—Restriction map of the two zeta-globin genes and their flanking DNA. The one-letter code for the restriction enzymes is shown in Figure 2. Polymorphic restriction sites have asterisks. Two regions subject to variations in length are denoted by hatching; the lengths shown for these two regions apply to DNA from allele *z* at the intergenic location and allele *c* at the intron location. The *Alu* repeat is indicated by an arrow.

avoiding an *Alu* family repeat unit (RUBIN *et al.* 1980) near the 5′ end of the *zeta 1* gene, which renders the region downstream of the *Sac*I site useless as a probe for genomic DNA. For a zeta gene probe, we selected a 2.0-kb *Hin*fI fragment (f–f) extending from the first zeta gene exon through the end of the third exon. This fragment hybridizes efficiently to both *zeta 1* and *zeta 2* gene sequences, with no background hybridization from repetitive DNA within the introns (under normal conditions of stringency).

Restriction mapping of chromosomal DNA: A composite restriction map shows all the sites encountered in our survey of the zeta regions of 76 individuals (Figure 3). Forty-eight positions have been deduced from fragment patterns using the zeta gene probe (for sites within and near the zeta genes) and the intergenic probe (for sites between the genes).

The sequence of nine restriction sites in and near the *zeta 2* gene (p v g f p v s c e) is identical to that of the *zeta 1* gene. By the method of NEI and TAJIMA (1983), the nucleotide sequence homology in the immediate vicinity of these two genes is 99%, in agreement with the sequencing results for a single human zeta region (PROUDFOOT, GIL and MANIATIS 1982). By contrast, there is no significant homology between the 5′ and 3′ flanking regions. The unit of homology, therefore, corresponds approximately to the presumed unit of transcription. There are no chromosomes with three zeta genes in the DNAs examined, and none were expected, because homology does not extend beyond the boundaries of the transcription unit. Unequal crossover between the two zeta genes, necessary to create the third gene, would have generated a pattern of homology outside the transcription unit had it occurred during the recent history of the region.

Site and length polymorphisms in the zeta region: Four of the sites in Figure 3 are polymorphic: an intergenic *Sac*I site (s*), a *Pvu*II site (v*) in *zeta 1*, and *Bam*HI (b*) and *Bgl*II (g*) sites downstream of *zeta 1*. These polymorphisms are probably due to point mutations, because they do not affect the lengths of fragments produced by other restriction enzymes. Two regions of DNA length variation are shown as hatched areas in Figure 3. All restriction

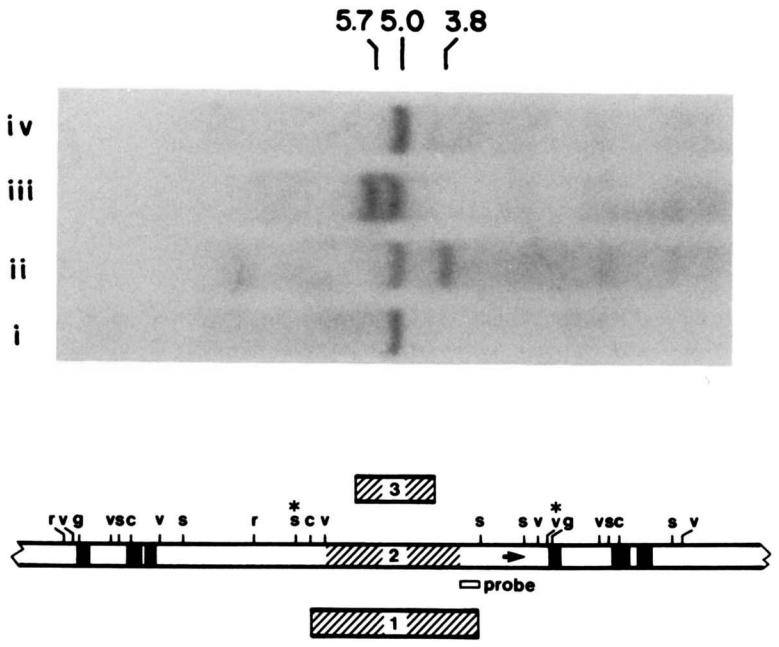

FIGURE 4.—Identification of three alleles of the zeta intergenic region. The autoradiogram shows four unrelated human DNA samples, which were digested with *Pvu*II and subjected to electrophoresis in 0.8% agarose gels, then were transferred to nitrocellulose and were hybridized with the intergenic probe. In lanes i and iv are individuals homozygous for the allele of intermediate length (*2*). Lane ii contains a heterozygote for alleles *2* and *3*, and a heterozygote for alleles *1* and *2* is shown in lane iii. The three length variants are depicted in the map below. Alleles *1*, *2* and *3* are identified by the length of the variable region (represented by a hatched box). The most common length allele (*2*) is shown in the map. The one-letter code for restriction sites is shown in Figure 2. Alleles *1*, *2* and *3* correspond to the large, medium and small alleles noted by GOODBOURN *et al.* (1983).

fragments spanning these two regions reveal the polymorphism, thus ruling out the possibility of point mutational changes in restriction sites.

The region of DNA containing the intergenic length variation lies on a single *Pvu*II fragment that can be detected in genomic DNA using the intergenic probe (Figure 4). *Pvu*II sites (v) define a segment of intergenic DNA, the length of which is 3.8, 5.0 or 5.7 kb. The three length alleles observed in human DNAs are shown in samples i–iv (Figure 4). We use the term "allele" following the usage of previous authors (BELL, KARAM and RUTTER, 1981; GOLDFARB *et al.* 1982; GOODBOURN *et al.* 1983; JEFFREYS, WILSON and THEIN 1985). Variants differing by <100 bp probably would not be distinguished.

From the map in Figure 2, it can be seen that *Hin*fI (f) and *Bst*EII (e) may be used to identify length alleles in the *zeta 1* gene. By digesting with either of these enzymes and hybridizing with the zeta gene probe, we can identify four alleles (*a–d*) within the *zeta 1* gene (Figure 5). The upper, nonpolymorphic bands generated by both enzymes carry the *zeta 2* genes. The lower, polymorphic bands carry the *zeta 1* genes of the five individuals. We have observed

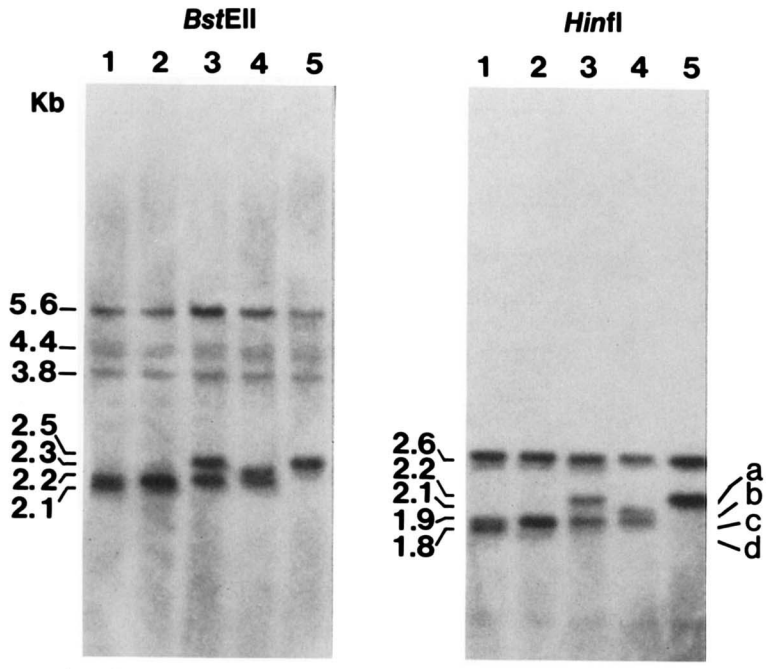

FIGURE 5.—Identification of four length alleles in intron I of the *zeta 1* gene. Autoradiograms of DNA samples from five unrelated humans are shown in the left panel digested with *Bst*EII and in the right panel digested with *Hin*fI. In the left panel, fragments carrying intron I of the *zeta 1* gene range in size from 2.1 to 2.5 kb. The 5.6-, 3.8- and 4.4-kb *Bst*EII fragments carry, respectively, the 5′ and 3′ ends of the *zeta 2* gene and the 3′ end of the *zeta 1* gene. In the right-hand panel, the variable *zeta 1* gene appears as a series of *Hin*fI fragments ranging from 2.2 to 1.8 kb and labeled *a*, *b*, *c* and *d*. The hybridizing portion of the *zeta 2* gene migrates as a 2.6-kb *Hin*fI band.

only four alleles (*a–d*), which differ by approximately 50 bp (*c* vs. *d*) to 350 bp (*a* vs. *d*). The published zeta-globin sequence probably corresponds to allele *b*. The *c* and *d* alleles can just be resolved by this method, but alleles differing by fewer than 30 bp would be difficult to distinguish.

Additional restriction enzymes *Dde*I (i), *Pst*I (p) and *Bst*NI (n) were used to assign this length variation among the four alleles to the 5′ half of intron I, shown schematically in Figure 6. Sequence data for one allele show that the 5′ part of intron I contains 39 tandem copies of a 14-bp sequence, while the 3′ part is composed of unique sequence (PROUDFOOT, GIL and MANIATIS 1982). *Dde*I and *Bst*NI cleave in intron I near the boundary between unique and repeated sequences. Intron length variation among alleles is found in fragments spanning the 5′ half of the intron and, therefore, must be associated with the tandem array.

Frequency of alleles in major human populations: The frequencies of the alleles at three polymorphic restriction sites are shown in Table 1. (The number of examined individuals varies from table to table and from site to site because we did not score all individuals at every position.) The most striking differences seen between populations in this limited survey is the low frequency

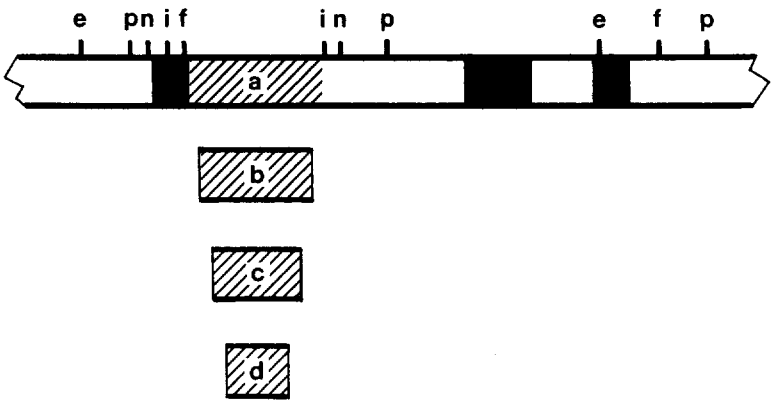

FIGURE 6.—Map of restriction sites used to define four length variants (*a*, *b*, *c* and *d*) of the *zeta 1* gene. The single-letter abbreviation used is *Bst*EII (e), *Hin*fI (f), *Pst*I (p), *Bst*NI (n), *Dde*I (i) and *Pvu*II (v). The exons are solid black, and the introns are hatched to indicate the length of the variable portion of the intron in each allele.

TABLE 1

Frequency of alleles at three polymorphic restriction sites

Restriction site[a]	Allelic frequency			
	Europe	Africa[b]	Asia	Australia
*Sac*I (s*)				
Frequency	0.78	0.50	0.30	0.29
No. of individuals	18	3	5	12
*Bam*HI (b*)				
Frequency	0.08	0.71	0.75	0.54
No. of individuals	18	7	4	12
*Bgl*II (g*)				
Frequency	0.02	0	0	0.04
No. of individuals	22	14	11	12

[a] See Figure 3 for the locations of these sites.
[b] Includes American Black samples.

of the *Bam*HI site (b*) in Europeans and an elevation of the frequency of the intergenic *Sac*I site (s*). The *Bgl*II (g*) site is common in Sardinians (M. PIRASTU and Y. W. KAN, personal communication), but rare in the people we have examined. The fourth polymorphic site (PvuII, v*) could not be scored independently of the intron length polymorphism, but the site appears to be present at high frequency in association with the *c* and *d* alleles of the intron.

Table 2 gives the frequencies of the length alleles at the intergenic location in 58 unrelated humans. There is geographic variation in the frequencies of these alleles. All four populations are polymorphic for alleles *1*, *2* and *3*; the frequency of allele *3*, for example, varies from 0.02 to 0.42. In general, allele *2* is the most common, except in people of Asian origin, where allele *3* may be the most common. The intron length alleles (*a–d*) also are widely distributed in all four populations (Table 3). The allele frequencies vary considerably

TABLE 2

Frequency of length alleles at the intergenic location

Continent of origin	No. of individuals[a]	Allelic frequency			Heterozygosity	
		1	*2*	*3*	Observed[b]	Expected[c]
Europe	21	0.31	0.67	0.02	0.35	0.45
Africa[d]	14	0.25	0.61	0.14	0.31	0.56
Asia	12	0.25	0.33	0.42	0.69	0.65
Australia	11	0.09	0.82	0.09	0.50	0.31
All	58	0.25	0.61	0.15	0.43	0.54

[a] The number of individuals is the number of (diploid) DNA samples examined by Southern blotting.
[b] A single band representing a chromosomal region was assumed to indicate homozygosity, and two bands were scored as heterozygosity.
[c] Heterozygosity was calculated by the method of NEI (1975).
[d] Includes American Blacks.

TABLE 3

Frequency of length alleles in intron I of the *zeta 1* gene

Continent of origin	No. of individuals[a]	Allelic frequency				Heterozygosity	
		a	*b*	*c*	*d*	Observed[b]	Expected[c]
Europe	20	0.50	0.42	0.08	0	0.77	0.57
Africa[d]	6	0.08	0.17	0.42	0.33	0.17	0.68
Asia	4	0.25	0.38	0.38	0	0.75	0.65
Australia	12	0.33	0.17	0.46	0.04	0.83	0.65
All	42	0.37	0.31	0.26	0.06	0.69	0.70

[a] The number of individuals is the number of (diploid) DNA samples examined by Southern blotting.
[b] For purposes of determining observed heterozygosity, a single band representing a chromosomal region was assumed to indicate homozygosity, and two bands were scored as heterozygosity.
[c] Heterozygosity was calculated by the method of NEI (1975).
[d] Includes American Black samples.

among these populations. Allele *b* is common in populations of Europe and Asia, and allele *c* is common in Australia and Africa. The shortest allele, *d*, is rare except in individuals of African origin, where it is found at a frequency of 0.33.

Haplotypes: We were able to determine haplotypes for four of the six defined polymorphisms: the two sites of length variation and two restriction sites (*Sac*I and *Bam*HI), which span a 12-kb segment around the *zeta 1* gene. Haplotypes were obtained for members of two small pedigrees, and it was possible to assign alleles to one chromosome in unrelated individuals homozygous at three of the four sites. (From unrelated individuals, only the most common haplotypes in a population can be ascertained.) Examination of 30 chromosomes (Table 4) revealed 12 of the possible 48 haplotypes. This seems a large number of haplotypes for such a small sample; for example, eight haplotypes are represented among six European individuals.

The distribution of length alleles *1–3* and *a–d* among these haplotypes im-

TABLE 4

Haplotypes of the zeta region

Haplotype	Allelic state				No. of chromosomes	Distribution
	*Sac*I site	Intergenic allele	Intron allele	*Bam*HI site		
I	+	*1*	*a*	−	4	Europe/Australia
II	+	*1*	*b*	−	1	Europe
III	+	*2*	*a*	−	5	Europe/Australia
IV	+	*2*	*b*	−	4	Europe/Africa
V	+	*2*	*c*	+	2	Europe/Africa
VI	−	*2*	*a*	−	2	Europe
VII	−	*2*	*b*	−	2	Europe/Africa
VIII	−	*2*	*b*	+	1	Australia
IX	−	*2*	*c*	+	5	Africa/Australia
X	−	*2*	*d*	+	1	Australia
XI	−	*3*	*a*	−	1	Europe
XII	−	*3*	*b*	+	2	Asia

plies that associations between these sites are random. By contrast, alleles *1* and *3* are always associated with the plus and minus alleles, respectively, at the *Sac*I site (s* in Figure 3). Likewise, the *Bam*HI site (b* in Figure 3) is always plus when associated with alleles *c* or *d*, but is minus when associated with the longest allele *a*. The *2* and *b* alleles occur in association with both states of the s* and b* sites, suggesting that *2b* is the ancestral state. To explain the haplotype diversity in Table 4, without appealing to parallel or back mutations, it is necessary to postulate a minimum of seven new mutations and four recombination events. These four hypothetical recombinations could all have taken place between the two sites of length variation. Alternatively, the haplotype diversity could be explained without recombination, by postulating three additional length mutations and one additional site change.

Stability of the polymorphism: DNA rearrangements associated with the *zeta 1* gene appear to be stable in the germ line and in somatic cell lineages. Examination of ten individuals in one human family and four in another showed that length alleles were inherited in a Mendelian fashion. Furthermore, the same DNA arrangements are seen in samples, whether derived from blood or placenta. In addition, the restriction map of DNA from erythroleukemic cells did not change after 60 passages in culture or after induction of embryonic hemoglobin synthesis (data not shown).

DISCUSSION

Polymorphism in the zeta-globin gene region: Within the region that has been mapped most intensively, we have defined four polymorphic restriction sites and two locations of length polymorphism. The frequency of point mutational polymorphism in this region of human chromosome *16* is the same as in the human beta-globin locus on chromosome *11* (ORKIN, ANTONARAKIS and KAZAZIAN 1983). No length mutational polymorphism was found near the *zeta*

2 globin gene. The high incidence of length polymorphism has no counterpart in the beta-globin locus. There are fewer length alleles at the two zeta region loci than reported for the human insulin locus, which has more than 20 (BELL, KARAM and RUTTER 1981), or at the human *c-Ha-ras-1* locus, which has more than six (GOLDFARB *et al.* 1982).

A limited haplotype analysis reveals the possibility of multiple recombinations within a 3–4-kb segment bearing exon I of the *zeta 1* gene and bounded by the two variable length sequences. The data could also be explained by a high rate of mutation or conversion of one allele to another. A "hot spot" for recombination has been reported in the beta family of globin genes, between the major adult gene and the embryonic genes (ANTONARAKIS *et al.* 1982), together with evidence for recurrent mutation or gene conversion in the adult beta-globin gene (ANTONARAKIS *et al.* 1984). Multiple recombination events leading to deletion and haplotype diversity have been documented for alpha thalassemia chromosomes in the Thai population (WINICHAGOON *et al.* 1984). Our data and previously published results are consistent with a high rate of recombination near the *zeta 1* gene. In addition, they support the idea that DNA sequences in the sites of length polymorphism may be recombinogenic (JEFFREYS, WILSON and THEIN 1985). Recombination in the *zeta 1* region could produce potentially advantageous combinations of alleles at the embryonic zeta locus and the adult alpha-globin locus 4 kb downstream. A high rate of recombination in this region could explain the linkage equilibrium noted between alleles at the intergenic location and alpha thalassemia alleles (WINICHAGOON *et al.* 1984).

Persistence or regeneration of length variation: There are two conventional ways of accounting for the observation that each of the four human groups represented in our samples contains the allelic length variants. Either most of the alleles predate the time of divergence of these populations (approximately 100,000 yr ago, NEI and ROYCHOUDHURY 1982) or the same length variants have been generated in each population.

By comparing restriction maps or base sequences, one can estimate the possible time since two alleles diverged. Both comparisons can be made for alleles *1* and *2* at the intergenic location. These two alleles differ in base sequence of the repeats, as well as in the number of repeated elements (GOODBOURN *et al.* 1983). DNA sequences obtained for repeats 27–32 of alleles *1* and *2* show a point mutational divergence of 5% (10 substitutions in 216 bp). Differences in the restriction maps of the two cloned alleles, when converted to base sequence differences by the method of NEI and TAJIMA (1983), also indicate that the alleles differ throughout their length by at least 5%. In comparing the restriction maps, we have taken into account the fact that the allelic elements have undergone rearrangement in addition to point mutation. We recognize that there is uncertainty about alignment of such alleles and that incorrect alignments can give inflated estimates of point-mutational divergence.

In cases where the complete sequence of a single allelic array has been determined, the time elapsed since its formation or homogenization can be estimated from the number of variant types relative to a consensus sequence.

This estimate is independent of the size of the array and position of variants. For the zeta intergenic allele *2*, there are 23 different types of variants out of 32 repeat units. These variant types differ from the 36-bp consensus sequence by 65 mutations (2.8 mutations per variant), which is a 5.6% sequence divergence. Thus, the tandem array has existed long enough to have given rise to alleles differing by 5%.

Assuming a substitution rate in noncoding DNA of 0.4% per million years (ZIMMER 1980), such alleles would appear to have been diverging for more than 10 million years, implying that the length polymorphism may have persisted for at least 10 million years. If this time estimate were correct, length polymorphism might be expected at corresponding locations in the zeta region of apes. Multiple alleles of similar lengths to those of humans are observed at both locations (*i.e.*, between the zeta genes and within intron I of *zeta 1*) (CHAPMAN, VINCENT and WILSON 1981; CHAPMAN and WILSON 1982; CHAPMAN and WILSON, unpublished results). Indeed, a recently sequenced allele of intron I from a chimpanzee is identical in length to the human allele of known sequence, although differing in base sequence (WILLARD *et al.* 1986). The possibility of persistence of the same alleles deserves further attention.

The persistence of the same set of alleles at a locus for 10 million years would be unexpected, both from the standpoint of standard population genetic theory and on empirical grounds. One can calculate how long a two-allele polymorphism is likely to last in the absence of selection. Assuming values of 10^4 for the long-term effective population size (N_e) and 15 yr for the mean length of a generation (g) in the lineage leading from apes to humans (NEI and GRAUR 1984), the persistence time for such a polymorphism should be less than $4N_eg$ yr; that is, less than 600,000 yr. For a three-allele polymorphism, the persistence time should be even shorter (JEFFREYS, WILSON and THEIN 1985). Consistent with this expectation is the empirical observation that there are no shared polymorphisms at 44 protein loci between humans and chimpanzees, which diverged about 5 million years ago (KING and WILSON 1975).

The lifetime of a polymorphism could be enhanced if balancing selection favored maintenance of all the alleles, or if the effective population size were large and partially subdivided as regards recurrent exposure to certain selective pressures (*e.g.*, malaria). A set of alleles could persist if selective advantage for each allele were inversely related to its frequency (*cf.* NEI and GRAUR 1984; MAY 1984), or if biased allelic conversion (SZOSTAK *et al.* 1983) were related inversely to allele frequency.

Alternatively, the length alleles may be continuously generated. Our mapping study has established that intron length variants *a–d* map within the repeating structure of the *zeta 1* large intron. Similarly, the intergenic variants *1–3* occur in a region of repeated DNA (GOODBOURN *et al.* 1983). It is attractive to suppose that the tandem arrays could produce length variants by unequal alignments and genetic exchange. JEFFREYS, WILSON and THEIN (1985) have argued that such a process can be very fast, the mutation rate possibly being as high as 10^{-4} per kb per generation. To explain the limited set of observed length alleles, which differ by multiples of the length of a repeat, the

process must use only specific alignments, or there must be strong selection against products of intermediate length. In addition, a high rate of point mutation is required to compensate for the tendency of rapid unequal exchange to result in homogenization of repeat units (JEFFREYS, WILSON and THEIN 1985). Extensive sequencing of length alleles from several human populations and from apes will provide a means to evaluate the age of length alleles and to investigate the mechanism by which variation is generated or persists at this locus.

We thank W. S. DAVIDSON, R. L. CANN, L. NICOLAISEN, K. LEE, Y. W. KAN, J. F. GUSELLA, D. HOUSMAN, S. GOLUB, T. MANIATIS and J. LAUER for human DNA samples and cloned DNA. We appreciate the diligent tissue culture work of T. MULLENBACH, and we thank L. HAUGEN, M. GARLIN, K. RINE and Design Enterprises for preparation of the figures. For critical reading of the manuscript, we are indebted to E. M. PRAGER, R. L. CANN, M. STONEKING and H. OCHMAN. This work was supported by the American Cancer Society (senior fellowship to B.S.C.), the National Science Foundation (A.C.W.) and the National Institutes of Health (NIEHS Center grant ES01896).

LITERATURE CITED

ANTONARAKIS, S. E., C. D. BOEHM, P. V. J. GIARDINA and H. H. KAZAZIAN, JR., 1982 Nonrandom association of polymorphic restriction sites in the beta-globin gene cluster. Proc. Natl. Acad. Sci. USA **79:** 137–141.

ANTONARAKIS, S. E., C. D. BOEHM, G. R. SERGEANT, C. E. THEISEN, G. J. DOVER and H. H. KAZAZIAN, JR., 1984 Origin of the beta S-globin gene in Blacks: the contribution of recurrent mutation or gene conversion or both. Proc. Natl. Acad. Sci. USA **81:** 853–856.

BELL, G. I., J. H. KARAM and W. J. RUTTER, 1981 Polymorphic DNA region adjacent to the 5′ end of the human insulin gene. Proc. Natl. Acad. Sci. USA **78:** 5759–5763.

BELL, G. I., M. J. SELBY and W. J. RUTTER, 1982 The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. Nature **295:** 31–35.

CAPON, D. J., E. Y. CHEN, A. D. LEVINSON, P. H. SEEBURG and D. V. GOEDDEL, 1983 Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue. Nature **302:** 33–37.

CHAPMAN, B. S., K. A. VINCENT and A. C. WILSON, 1981 Extensive polymorphism and evolution in zeta globin genes. J. Cell. Biochem. **5** (Suppl): 400.

CHAPMAN, B. S. and A. C. WILSON, 1982 Variable structure of IVS I in human and ape zeta globin genes. J. Cell. Biochem. **6** (Suppl): 257.

CHAPMAN, B. S. and A. C. WILSON, 1983 Tempo and mode of gene correction in the zeta globin region. J. Cell. Biochem. **7B** (Suppl): 155.

GOLDFARB, M., K. SHIMIZU, M. PERUCHO and M. WIGLER, 1982 Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells. Nature **296:** 404–409.

GOODBOURN, S. E. Y., D. R. HIGGS, J. B. CLEGG and D. J. WEATHERALL, 1983 Molecular basis of length polymorphism in the human zeta globin gene complex. Proc. Natl. Acad. Sci. USA **80:** 5022–5026.

HIGGS, D. R., S. E. Y. GOODBOURN, J. S. WAINSCOAT, J. B. CLEGG and D. J. WEATHERALL, 1981 Highly variable regions of DNA flank the human alpha globin genes. Nucleic Acids Res. **9:** 4213–4224.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985 Hypervariable 'minisatellite' regions in human DNA. Nature **314:** 67–73.

KING, M.-C. and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107–116.

LAUER, J., C-K. J. SHEN and T. MANIATIS, 1980 The chromosomal arrangement of human alpha-like globin genes: sequence homology and alpha globin gene deletions. Cell **20:** 119–130.

LIEBHABER, S. A., M. J. GOOSSENS and Y. W. KAN, 1980 Cloning and complete nucleotide sequence of a human 5′ alpha globin gene. Proc. Natl. Acad. Sci. USA **77:** 7054–7058.

MANIATIS, T., A. JEFFREYS and D. G. KLEID, 1975 Nucleotide sequence of the rightward operator of phage lambda. Proc. Natl. Acad. Sci. USA **72:** 1184–1188.

MAY, R. M., 1984 Mathematical modeling: the cubic map in theory and practice. Nature **311:** 13–14.

MCKUSICK, V. A., 1982 The human gene map. Genet. Maps **2:** 327–350.

MICHELSON, A. M. and S. H. ORKIN, 1980 The 3′ untranslated regions of the duplicated human alpha globin genes are unexpectedly divergent. Cell **22:** 371–377.

NEI, M. 1975 *Molecular Population Genetics and Evolution.* North-Holland, New York.

NEI, M. and D. GRAUR, 1984 Extent of protein polymorphism and the neutral mutation. Evol. Biol. **17:** 57–87.

NEI, M. and R. ROYCHOUDHURY, 1982 Origins of human races. Evol. Biol. **14:** 1–59.

NEI, M. and F. TAJIMA, 1983 Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105:** 207–217.

ORKIN, S. H., S. E. ANTONARAKIS and H. H. KAZAZIAN, 1983 Polymorphism and molecular pathology of the human beta globin gene. Prog. Hematol. **13:** 372–374.

PROUDFOOT, N. J., A. GIL and T. MANIATIS, 1982 The structure of the human zeta globin gene and a closely linked, nearly identical pseudogene. Cell **31:** 553–563.

PROUDFOOT, N. J. and T. MANIATIS, 1980 The structure of a human alpha globin pseudogene and its relationship to alpha globin gene duplication. Cell **21:** 537–544.

PROUDFOOT, N. J., T. R. RUTHERFORD and G. A. PARTINGTON, 1984 Transcriptional analysis of human zeta globin genes. EMBO J. **3:** 1533–1540.

RUBIN, C. M., C. M. HOUCK, P. L. DEININGER, T. FRIEDMANN and C. SCHMID, 1980 Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences. Nature **284:** 372–374.

RUTHERFORD, T. R., J. B. CLEGG and D. J. WEATHERALL, 1979 K562 human leukemic cells synthesize embryonic hemoglobin in response to hemin. Nature **280:** 164–165.

SAWADA, I., M. P. BEAL, C-K. J. SHEN, B. CHAPMAN, A. C. WILSON and C. SCHMID, 1983 Intergenic DNA sequences flanking the pseudoalpha globin genes of human and chimpanzee. Nucleic Acids Res. **11:** 8087–8101.

SOUTHERN, E. M., 1975 Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:** 503–517.

SZOSTAK, J. W., T. L. ORR-WEAVER, R. J. ROTHSTEIN and F. W. STAHL, 1983 The double-strand-break repair model for recombination. Cell **33:** 25–35.

TAYLOR, J. M., A. DOZY, Y. W. KAN, H. E. VARMUS, L. E. LIE-INJO, J. GANESAN and D. TODD, 1974 Genetic lesion in homozygous alpha thalassemia (hydrops fetalis). Nature **251:** 392–393.

WAHL, G. M., M. STERN and G. R. STARK, 1979 Efficient transfer of large DNA fragments from agarose gels to diazobenzyloxymethyl-paper and rapid hybridization by using dextran sulfate. Proc. Natl. Acad. Sci. USA **76:** 3683–3687.

WILLARD, C., E. WONG, J. F. HESS, C-K. J. SHEN, B. S. CHAPMAN, A. C. WILSON and C. W. SCHMID, 1968 Comparison of human and chimpanzee *zeta 1* globin genes. J. Mol. Evol. In press.

WILSON, J. T., L. B. WILSON, J. K. DERIEL, L. VILLA-KOMAROFF, A. EFSTRATIADIS, B. G. FORGET and S. M. WEISSMAN, 1978 Insertion of synthetic copies of human globin genes into bacterial plasmids. Nucleic Acids Res. **5:** 563–581.

WINICHAGOON, P., D. R. HIGGS, S. E. Y. GOODBOURN, J. B. CLEGG, D. J. WEATHERALL and P. WASI, 1984 The molecular basis of alpha thalassemia in Thailand. EMBO J. **3:** 1813–1818.

ZIMMER, E. A., 1980 Evolution of primate globin genes. Ph.D. Thesis, University of California, Berkeley.

Communicating editor: M. NEI