*Proc. Natl. Acad. Sci. USA*
Vol. 79, pp. 6608–6611, November 1982
Genetics

# Evidence for multiple origins of the $\beta^E$-globin gene in Southeast Asia

(DNA polymorphisms/haplotypes/population genetics/hemoglobinopathies)

STYLIANOS E. ANTONARAKIS*, STUART H. ORKIN†, HAIG H. KAZAZIAN, JR.*, SABRA C. GOFF†, CORINNE D. BOEHM*, PAMELA G. WABER*, JULIANNE P. SEXTON†, HARRY OSTRER*, VIRGIL F. FAIRBANKS‡, AND ARAVINDA CHAKRAVARTI§

*Department of Pediatrics, Genetics Unit, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205; †Department of Pediatrics, Division of Hematology and Oncology, Children's Hospital Medical Center, Boston, Massachusetts 02115; ‡Department of Laboratory Medicine, Mayo Clinic, Rochester, Minnesota 55091; and §Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261

*Communicated by John W. Littlefield, July 19, 1982*

**ABSTRACT** **To investigate whether recurrent mutation has contributed to the high frequency of the $\beta^E$-globin gene in Southeast Asia, we used the haplotypes at three polymorphic restriction sites within and to the 3′ side of the $\beta$-globin gene to predict the framework of 23 $\beta^E$-globin genes. These haplotypes suggested that $\beta^E$-globin genes are present in two different $\beta$-globin gene frameworks. DNA sequence determination of one gene representing each framework demonstrated that the same mutation (GAG → AAG at codon 26) was present in both frameworks. Moreover, the frameworks differed at three nucleotide positions known to be polymorphic in Mediterraneans. These polymorphic sites are located 70 nucleotides to the 5′ side of the $\beta^E$ mutation and 382 and 1032 nucleotides to the 3′ side of it. The existence of the $\beta^E$ mutation in these two $\beta$-globin gene frameworks can be explained by (*i*) recurrent mutation giving rise to $\beta^E$-globin, (*ii*) a double crossing-over event, or (*iii*) two single crossing-over events. Mathematical analysis suggests that the first alternative, recurrent mutation of G → A at the first nucleotide of codon 26, is most likely.**

Hemoglobin E ($\alpha_2\beta_2^{26\ Glu \rightarrow Lys}$) is the most common or next-to-most common hemoglobin variant in the world. Until recently, its occurrence was generally limited to Southeast Asia, where the $\beta^E$-globin gene has a frequency approaching 0.3 in some areas and 0.5 in some isolated tribes (1). Because the $\beta^E$-globin gene is associated with reduced levels of $\beta$-globin mRNA and behaves like a mild $\beta$-thalassemia gene (2–5), a selective advantage for individuals with the hemoglobin E trait (relative resistance to malaria) has been postulated as one explanation for the high $\beta^E$-globin gene frequency (1). In this work we have examined the origin(s) of this particular mutant allele.

Since the description of the first restriction site polymorphism in the $\beta$-globin gene cluster in 1978 (6), a number of other polymorphic restriction sites and other nucleotide polymorphisms have been found (7–10). Chromosomes may be characterized by their pattern of polymorphic restriction sites, which we define as a haplotype. Several different haplotypes have been observed in chromosomes bearing the $\beta^A$- and $\beta^{thalassemia}$-globin genes (hereafter referred to as $\beta^A$ and $\beta^{thal}$ chromosomes, respectively) in Greeks and Italians (10). In those with thalassemia genes, particular haplotypes are strongly associated with specific $\beta$-thalassemia mutations (11). DNA sequence and restriction site polymorphism data have demonstrated that normal $\beta$-globin genes among Mediterraneans are of three basic types, which we call frameworks (11). Frameworks 1 and 2 differ at a single nucleotide, whereas framework 3 has the structure of framework 2 plus four additional substitutions. These frameworks can be detected in uncloned DNAs by analysis of both intragenic and extragenic restriction sites (11).

By analysis of restriction enzyme polymorphisms and by direct study of cloned $\beta^E$-globin genes, we obtained evidence for the existence of the $\beta^E$ mutation in two different gene frameworks. Mathematical analysis suggests that in all likelihood multiple origins of the $\beta^E$-globin gene have contributed to its high frequency in Southeast Asia.

## METHODS

**Subjects.** Our subjects were (*i*) unselected Cambodian refugees at Khao I Dang who originated from the Phnom Penh area, (*ii*) unselected Cambodians living in the Washington, D.C. area, and (*iii*) individuals homozygous for hemoglobin E who originated from Laos, Thailand, and Cambodia.

**Restriction Endonuclease Analysis and Preparation of Radioactive Probes.** DNA isolation, digestion of DNA with restriction endonucleases, electrophoresis of DNA fragments, transfer of DNA fragments, hybridization of genomic fragments with radioactive probes, washing of filters, and autoradiography were carried out as described (12–14). Restriction enzymes used in this study were *Hin*cII, *Hin*dIII, *Hin*fI, *Ava* II, *Hgi*AI, and *Bam*HI. To detect various polymorphic restriction sites, we used the following cloned DNA sequences as probes: (*i*) a 1.3-kilobase (kb) *Bam*HI–*Eco*RI genomic fragment containing $\varepsilon$-globin gene sequences, (*ii*) a 1.7-kb *Bgl* II–*Xba* I genomic fragment containing $\psi\beta_1$-globin gene sequences, (*iii*) a 1.2-kb *Mbo* II–*Hin*dIII fragment containing $\beta$-globin cDNA sequences derived from the recombinant plasmid JW 101, (*iv*) a 1.1-kb *Taq* I fragment containing $\gamma$-globin cDNA sequences derived from the recombinant plasmid JW 151, and (*v*) a 1.3-kb *Hpa* I–*Bam*HI genomic fragment containing the 5′ end of the $\beta$-globin gene. All fragments were radiolabeled with [$^{32}$P]dCTP and [$^{32}$P]dATP by the nick-translation function of *Escherichia coli* DNA polymerase I as described (15).

**Gene Cloning and DNA Sequence Analysis.** $\beta^E$-Globin genes were cloned within a 7.5-kb *Hin*dIII restriction fragment of genomic DNA in the phage Charon 28 (16) as described (17). The $\beta^E$-globin genes of positive recombinants were subcloned as 4.4-kb *Pst* I fragments in pBR322 for DNA sequence analysis. All DNA sequences were determined by the method of Maxam and Gilbert (18). Fragments were labeled at their 5′ or 3′ termini and either their strands were separated or they were subjected to secondary digests prior to sequencing. Experiments involving recombinant DNA were performed in $P_1$-$EK_1$ con-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: IVS, intervening sequence; kb, kilobase pairs.

tainment in accordance with National Institutes of Health guidelines.

## RESULTS

**DNA Polymorphism Haplotypes of $\beta^A$ Chromosomes of Cambodians.** Using nine common restriction site polymorphisms in the β-globin gene cluster and appropriate family studies to determine linkage of sites along chromosomes, we have observed nine haplotypes of $\beta^A$ chromosomes among Cambodians (Fig. 1). The majority of chromosomes had the haplotype $+-----+-\beta^A-+$ (42%) or the haplotype $+-----++\beta^A+-$ (25%) (the nomenclature is described in the legend to Fig. 1). As in our previous study of Mediterraneans, there was random association of the restriction site polymorphisms situated 5′ to the δ-globin gene with those located 3′ to it (10). The two haplotypes that are most common in Cambodians accounted for only about 30% of all $\beta^A$ chromosomes in Mediterraneans (10).

**β-Globin Gene Frameworks in Cambodians.** Precise identification of DNA frameworks requires DNA sequence analysis. However, in the sequences of 13 Mediterranean β-globin genes we have previously determined, we found a strict association between the intragenic polymorphic *Hgi*AI and *Ava* II sites, the polymorphic *Bam*HI site 3′ to the gene, and the type of gene framework (11). Framework 1 was associated with the presence of these restriction sites (Fig. 1, sites 7, 8, and 9); framework 2 was associated with the absence of the *Bam*HI site (Fig. 1, site 9); and framework 3 lacked the *Hgi*AI and *Ava* II sites (Fig. 1, sites 7 and 8). Among Cambodian DNA samples, restriction mapping revealed that frameworks 1, 2, and 3, respectively, accounted for 18, 35, and 47% of 47 $\beta^A$-globin genes studied. This distribution of gene frameworks contrasts with that found in Mediterraneans, American Blacks, and Asiatic Indians as shown in Fig. 2.

**Polymorphism Haplotypes and β-Globin Gene Frameworks of $\beta^E$-Globin Genes.** We have examined 23 $\beta^E$-globin alleles: 10 in Cambodians, 10 in Laotians, and 3 in Thais. The clinical phenotype of hemoglobin E homozygotes was similar to that previously described (1) and did not vary among ethnic groups. Among $\beta^E$-globin gene-bearing chromosomes (hereafter termed $\beta^E$ chromosomes), three haplotypes were represented (Fig. 3). Six $\beta^E$-globin genes in Cambodians from the Phnom Penh region had the haplotype (c)$-+-+++-\beta^E-+$, which predicts a framework 3 gene. The remaining $\beta^E$-globin genes in Cambodians, Laotians, and Thais had haplotypes (a)$-+-+++\beta^E+-$ and (b)$+-----++\beta^E+-$, which predict framework 2 genes. Note that the latter two haplotypes can be derived from each other by a single crossing-over event between the *Hgi*AI site and the β-globin gene and the *Hin*cII site 3′ to the $\psi\beta_1$-globin gene.

**DNA Sequence of $\beta^E$-Globin Genes.** One $\beta^E$-globin allele of haplotype (a)$-+-+++\beta^E+-$ (framework 2) and one allele of haplotype (c)$-+-+++-\beta^E-+$ (framework 3) were chosen for sequence analysis. In both cloned genes we observed the codon 26 change GAG → AAG anticipated for the $\beta^E$-globin mutation (Fig. 4). In the haplotype (a)$-+-+++\beta^E+-$ gene, we identified the intragenic polymorphism at intervening sequence 2 (IVS2) position 74 that characterizes the framework 2 gene (11, 20). In the haplotype (c)$-+-+++-\beta^E-+$ gene, we observed this polymorphism plus three others: CAC → CAT in codon 2 (the *Hgi*AI polymorphism), C → G at IVS2 position 16 (the *Ava* II polymorphism), and T → C at IVS2 position 666 (Fig. 4). The only other polymorphism found in framework 3 genes of Mediterraneans, located at IVS2 position 81, was not present in the $\beta^E$-globin gene of this type. Therefore, this β-globin gene framework appears to be intermediate between framework 2 and framework 3 genes in Mediterraneans and henceforth will be referred to as Asian framework 3. Polymorphisms in codon 2 and IVS2 positions 16 and 666 were absent in the $\beta^E$-globin gene of the haplotype (a)$-+-+++\beta^E+-$.

## DISCUSSION

Recently, DNA polymorphisms in the β-globin gene cluster have been used in an attempt to define the origin of common mutant alleles of the β-globin gene. Studies of $\beta^S$-globin gene-bearing chromosomes containing and lacking the *Hpa* I poly-

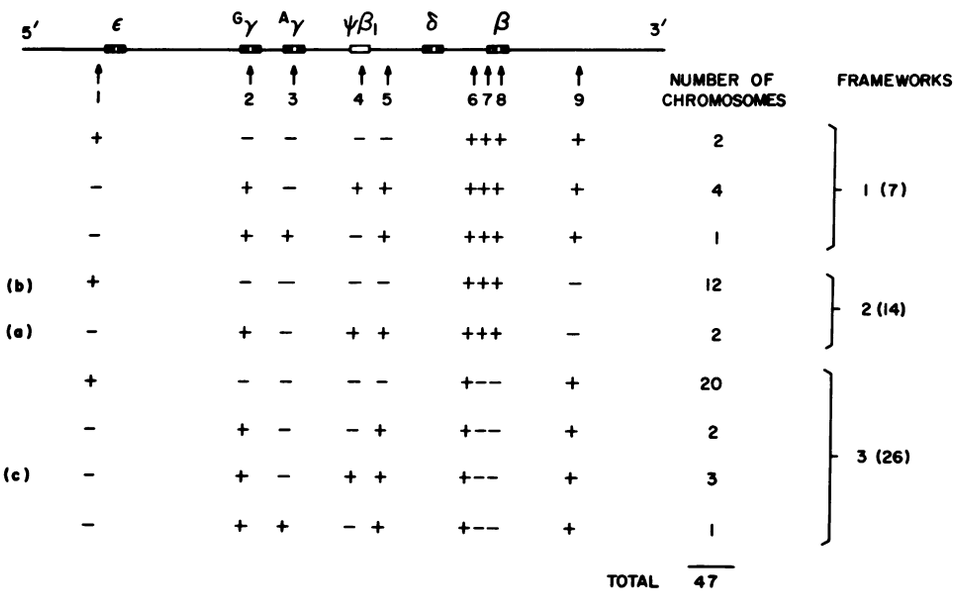


FIG. 1. Haplotypes of $\beta^A$ chromosomes in Cambodians. The + or − refers to the presence or absence of the following polymorphic restriction sites from left to right: 1, *Hin*cII 5′ to the ε gene; 2, *Hin*dIII in IVS2 of the $^G\gamma$-globin gene; 3, *Hin*dIII in IVS2 of the $^A\gamma$-globin gene; 4, *Hin*cII in the $\psi\beta_1$-globin gene; 5, *Hin*cII 3′ to the $\psi\beta_1$-globin gene; 6, *Hin*fI 5′ to the β-globin gene (19); 7, *Hgi*AI in codon 2 of the β-globin gene; 8, *Ava* II in IVS2 of the β-globin gene; and 9, *Bam*HI 3′ to the β-globin gene. The number of chromosomes with frameworks 1, 2, and 3 are shown to the right in parentheses. Haplotypes a, b, and c are those observed in $\beta^E$ chromosomes (Fig. 3).

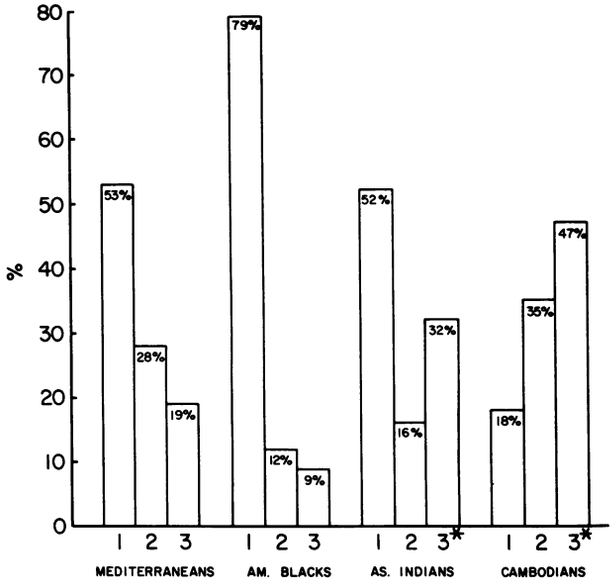


FIG. 2. β-Globin gene frameworks in various ethnic groups. Abscissa numbers 1, 2, 3, and 3* refer to frameworks 1, 2, 3, and Asian framework 3, respectively. Data were obtained on $\beta^A$-globin genes of 93 Mediterraneans, 27 American (Am.) Blacks, 32 Asiatic (As.) Indians, and 47 Cambodians. A $\chi^2$ analysis of the distribution of frameworks in Cambodians versus the distribution of frameworks in Mediterraneans, Asiatic Indians, and American Blacks gave $\chi^2$ values of 19.5, 11.7, and 26.6, respectively, each with two degrees of freedom. No significant differences were found by the $\chi^2$ analysis when the framework distributions in Mediterraneans, Asiatic Indians, and American Blacks were compared to each other.

morphic restriction site 3′ to the β-globin gene led Kan and Dozy to conclude that the $\beta^S$ mutation had at least two independent origins (21). This conclusion was based largely on the geographic differences in the $\beta^S$-globin *Hpa* I pattern association. However, on strictly mathematical grounds, the probability of a second independent $\beta^S$ mutation is not greater than that of a crossing-over event in the 5-kb region extending from codon 6 of the β-globin gene to the polymorphic *Hpa* I site (see below). Boyer *et al.* (22) used two additional sites (*Hin*dIII polymorphisms in the γ-globin genes) and analyzed the $\beta^S$ mutation in chromosomes of three common haplotypes. They argued that either multiple mutations or recombinational events could explain the findings equally well. By examination of *de novo* mutants for unstable and M hemoglobins, Nute and Stamatoyannopoulos concluded that specific mutant alleles have had multiple independent origins (23).

The same alternatives, multiple mutation or meiotic recombination, should be considered to explain our results. The presence of the $\beta^E$-globin gene in haplotypes of the same framework, (a)$-+-++++\beta^E+-$ and (b)$+----++\beta^E+-$, could well be explained by a crossing-over event 5′ to the β-globin gene. However, the discovery of the $\beta^E$-globin gene in two different gene frameworks argues more strongly for multiple origins of these alleles. If the mutation occurred only once, two crossing-overs would be required to generate one of these frameworks from the other.

One can estimate the probability of these recombination events and recurrent mutation. The average rate of recombination per base pair has been estimated to be $1 \times 10^{-8}$ per generation or $4 \times 10^{-10}$ per year (24). The estimated probability of recombination in the 70-nucleotide segment ($S_1$) between the *Hgi*AI site in codon 2 and codon 26, $r_1$, is $70 \times 4 \times 10^{-10}$, and that in the 382-nucleotide segment ($S_2$) between codon 26 and the *Ava* II site in IVS2, $r_2$, is $382 \times 4 \times 10^{-10}$ (Fig. 4). If we assume that the first $\beta^E$ mutation occurred $n$ years ago, then the probability of double crossing-over as a single event to place the $\beta^E$ mutation in either Asian framework 3 or framework 2 is roughly $nr_1r_2$, or $4.2 \times 10^{-15}n$. On the other hand, recombination could have occurred in two events: first in one segment ($S_1$ or $S_2$) and then in the other ($S_2$ or $S_1$). In this case, the theoretical probability of producing the two observed $\beta^E$-globin alleles, if the first $\beta^E$ mutation occurred $n$ years ago, is $n^2r_1r_2$. The probability of a second mutation at any nucleotide, given that the first occurred $n$ years ago, is estimated to be, $n\mu$ where $\mu$ is the mutation rate. The mutation rate at the first base pair of a codon has been estimated to be at least $0.7 \times 10^{-9}$ mutations per base pair per year (23, 25, 26).

Thus, the probability of a second mutation to $\beta^E$-globin compared to the probability of a single event producing a double crossing-over is $n\mu/nr_1r_2 = 0.7 \times 10^{-9}n/4.2 \times 10^{-15}n \approx 10^5$. This is the maximum probability favoring a second mutation at the same nucleotide. In contrast, the probability of a second mutation compared to the probability of two separate single crossing-over events is $\mu/nr_1r_2 = 163{,}000/n$. Because the mutation is almost completely limited to Mongoloid races, and the Mongoloids separated from the Caucasoids about 55,000 years ago (27), the minimum probability favoring a second mutation to $\beta^E$ is 163,000/55,000, or 3. Thus, the range of probability favoring a second mutation giving rise to $\beta^E$-globin gene has a minimum of 3 and a maximum of $10^5$.

Further evidence argues against the possibility of recombination leading to association of $\beta^E$-globin genes with different haplotypes and frameworks. First, no crossing-over intermediates have been observed in the 23 $\beta^E$ chromosomes we have examined. There has been no discordance of the *Hgi*AI and *Ava* II polymorphisms. Second, the framework 3 gene of the Asian variety containing the $\beta^E$ mutation has been found only in Cambodians and not in other Southeast Asian peoples, suggesting that it has an independent origin.

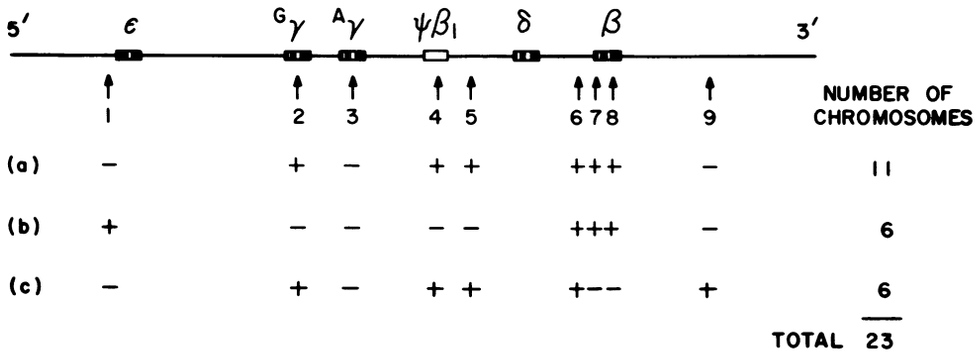


FIG. 3. Haplotypes of $\beta^E$ chromosomes. See legend to Fig. 1 for listing of polymorphic restriction sites analyzed.

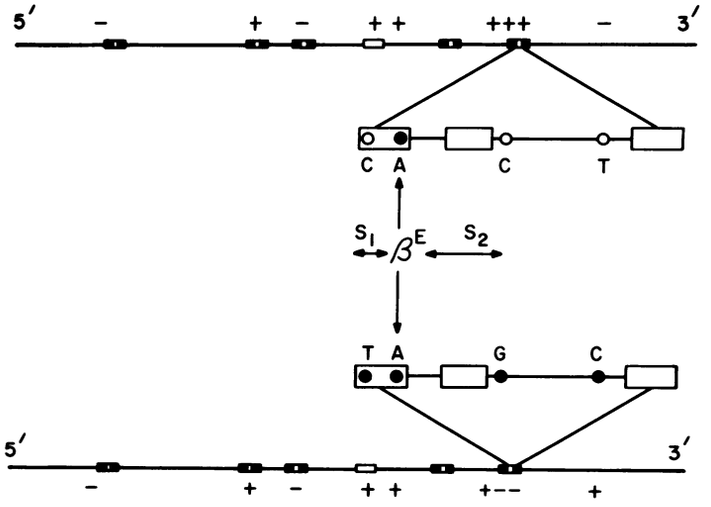


FIG. 4. $\beta^E$ mutation in two $\beta$-globin gene frameworks. Haplotypes of $\beta^E$-globin gene-bearing chromosomes selected for gene sequencing are shown. Nucleotide differences between the $\beta^E$-globin genes of framework 2 and Asian framework 3 (3*) are depicted. The mutation (GAG → AAG at codon 26 in both frameworks) is shown by arrows. $S_1$ and $S_2$ represent the segments between the mutation and the closest nucleotide differences in the frameworks.

There are reports of rare $\beta^E$-globin alleles in populations outside of Southeast Asia (1). We predict that these $\beta^E$-globin genes will be associated with different haplotypes from those observed in Asians and may represent other independent mutations. Another conclusion from our data is that the mutations leading to the $\beta^E$-globin gene, like those producing $\beta$-thalassemia (11), are more recent than those that have produced the highly polymorphic restriction sites.

The Asian framework 3 gene differs at a single nucleotide from the similar framework in Mediterraneans and appears to represent an intermediate between frameworks 2 and 3 of Mediterraneans. This new framework seems to be quite common among Asians, as it has been observed in a $\beta$-thalassemia allele of an Asiatic Indian (unpublished data). The nucleotide polymorphism at IVS2 position 81 that distinguishes the Asian and Mediterranean framework 3 genes may be a relatively recent event.

Studies of individuals with the $\beta^E$-globin gene have demonstrated reduced accumulation of $\beta^E$-globin mRNA in erythroid cells (4, 5). Because reduced $\beta$-globin mRNA is a hallmark of $\beta$-thalassemia, $\beta^E$-globin genes might be expected to reveal two mutations: one within codon 26 and another elsewhere causing reduction in mRNA production or stability. Additional DNA sequence determinations of the $\beta^E$-globin gene of the framework 3 variety demonstrated only a single mutation, raising the possibility that the reduced accumulation of mRNA results from the codon 26 mutation itself (unpublished data). Studies of $\beta^E$-globin genes in a transient expression system indicate a mechanism for this association.

**Note Added in Proof.** By DNA sequence determination, we recently have found the thalassemia mutation to a nonsense codon at codon 39 of the $\beta$-globin gene in $\beta$-globin gene frameworks 1 and 3 in Italians. This is a second example of the same mutation in two different $\beta$-globin gene frameworks.

We thank John Phillips for helpful discussions, Joseph Lee for technical assistance, and Sun Sitana and So Den for assistance in collecting blood samples. H.H.K. and S.H.O. are each supported by grants from the National Institutes of Health and the National Science Foundation. S.H.O. is supported also by a National Institutes of Health Research Career Development Award.

1. Flatz, G. (1966) *Humangenetik* **3**, 189–234.
2. Fairbanks, V. F., Oliveros, R., Brandabur, J. H., Willis, R. R. & Fiester, R. F. (1980) *Am. J. Hematol.* **8**, 109–121.
3. Fairbanks, V. F., Gilchrist, G. S., Brimhall, B., Jereb, J. A. & Goldston, E. C. (1979) *Blood* **53**, 109–115.
4. Traeger, J., Wood, W. G., Clegg, J. B., Weatherall, D. J. & Wasi, P. (1980) *Nature (London)* **288**, 497–499.
5. Benz, E. J., Berman, B. W., Tonkonow, B. L., Coupal, E., Coates, T., Boxer, L. A., Altman, A. & Adams, J. G. (1981) *J. Clin. Invest.* **68**, 118–126.
6. Kan, J. W. & Dozy, A. M. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5631–5635.
7. Jeffreys, A. J. (1979) *Cell* **18**, 1–10.
8. Tuan, D., Biro, P. A., deRiel, J. K., Lazarus, H. & Forget, B. G. (1979) *Nucleic Acids Res.* **6**, 2519–2544.
9. Kan, J. W., Lee, K. Y., Furbetta, M., Anguis, A. & Cao, A. (1980) *N. Engl. J. Med.* **302**, 185–188.
10. Antonarakis, S. E., Boehm, C. D., Giardina, P. V. J. & Kazazian, H. H., Jr. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 137–141.
11. Orkin, S. H., Kazazian, H. H., Jr., Antonarakis, S. E., Goff, S. C., Boehm, C. D., Sexton, J. P., Waber, P. G. & Giardina, P. V. J. (1982) *Nature (London)* **296**, 627–631.
12. Kunkel, L. M., Smith, K. D., Boyer, S. H., Borgaonkar, D. S., Wachtel, S. S., Miller, O. J., Breg, W. R., Jones, H. W. & Rary, J. M. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 1245–1249.
13. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 506–517.
14. George, D. L., Phillips, J. A., Franke, U. & Seeburg, P. H. (1981) *Hum. Genet.* **57**, 138–141.
15. Maniatis, T., Kee, G. S., Efstratiadis, A. & Kafatos, F. C. (1976) *Cell* **8**, 163–182.
16. Rimm, D. L., Hogness, D., Kucera, J. & Blattner, F. R. (1980) *Gene* **12**, 301–309.
17. Orkin, S. H. & Goff, S. C. (1981) *J. Biol. Chem.* **256**, 9782–9784.
18. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
19. Moschonas, N., deBoer, E. & Flavell, R. A. (1982) *Nucleic Acids Res.* **10**, 2109–2120.
20. Lawn, R. M., Efstratiadis, A., O'Connell, C. & Maniatis, T. (1980) *Cell* **21**, 647–651.
21. Kan, Y. W. & Dozy, A. M. (1980) *Science* **209**, 388–391.
22. Boyer, S. H., Panny, S. R., Smith, K. D. & Dover, G. J. (1981) in *Population and Biological Aspects of Human Mutations*, eds. Hook, E. B. & Porter, I. H. (Academic, New York), pp. 35–48.
23. Nute, P. E. & Stamatoyannopoulos, G. (1981) in *Population and Biological Aspects of Human Mutations*, eds. Hook, E. B. & Porter, I. H. (Academic, New York), pp. 337–350.
24. Kurnit, D. M. & Hoehn, H. (1979) *Annu. Rev. Genet.* **13**, 235–258.
25. Li, W.-H., Gojobori, T. & Nei, M. (1981) *Nature (London)* **292**, 237–239.
26. Boyer, S. H. & Smith, K. D. (1981) in *Population and Biological Aspects of Human Mutations*, eds. Hook, E. B. & Porter, I. H. (Academic, New York), pp. 49–64.
27. Nei, M. & Roychoudhury, A. K. (1974) *Am. J. Hum. Genet.* **26**, 421–443.