Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 8568–8571, October 1991
Genetics

Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variants A and A−

(human genetics/intron sequences/human evolution)

T. J. VULLIAMY*, A. OTHMAN*, M. TOWN*, A. NATHWANI*, A. G. FALUSI†, P. J. MASON*, AND L. LUZZATTO*

*Department of Haematology, Royal Postgraduate Medical School, Hammersmith Hospital, Ducane Road, London W12 OHS, Great Britain; and †Postgraduate Medical Centre, University College Hospital, Ibadan, Nigeria

Communicated by James V. Neel, June 24, 1991

ABSTRACT **The human X chromosome-linked gene encoding glucose-6-phosphate dehydrogenase (G6PD; EC 1.1.1.49) is known to be highly polymorphic from the biochemical characterization of enzyme variants. The variant A (with enzyme activity in the normal range) and the variant A− (associated with enzyme deficiency) each have a frequency of about 0.2 in several African populations. Two restriction fragment length polymorphisms have also been found in people of African descent, but not in other populations, whereas a silent mutation has been shown to be polymorphic in Mediterranean, Middle Eastern, African, and Indian populations. We report now on two additional polymorphisms that we have detected by sequence analysis, one in intron 7 and one in intron 8. The analysis of 54 African male subjects for the seven polymorphic sites, clustered within 3 kilobases of the *G6PD* gene, has revealed only 7 of the 128 possible haplotypes, indicating marked linkage disequilibrium. These data have enabled us to suggest an evolutionary pathway for the different mutations, with only a single ambiguity. The mutation underlying the A variant is the most ancient and the mutation underlying the A− variant is the most recent. Since it seems reasonable that the A− allele is subject to positive selection by malaria, whereas the other alleles are neutral, G6PD may lend itself to the analysis of the role of random genetic drift and selection in determining allele frequencies within a single genetic locus in human populations.**

The enzyme glucose-6-phosphate dehydrogenase (G6PD; EC 1.1.1.49) shows considerable variation in its biochemical properties. Based on the level of enzyme activity, the electrophoretic mobility, and the enzyme kinetics, >300 different variants have been described (1), of which 86 have been classified as polymorphic (2). Two notable examples in Africa that differ from the wild-type enzyme G6PD B are the nondeficient variant G6PD A and the deficient variant G6PD A−. Both have a gene frequency of around 0.2 in parts of that continent and in people of African ancestry.

Restriction fragment length polymorphisms (RFLPs) have been harder to find. However, a *Pst* I and a *Pvu* II RFLP have been characterized in populations of African descent (3–5). These RFLPs have recently been shown to be in marked linkage disequilibrium with the polymorphic mutations in G6PD A and A− (6, 7); as pointed out by Beutler and Kuhl (6), the results suggest that these mutations have arisen only once. To confirm this suggestion, we sought additional polymorphic sites by the comparison of G6PD intron sequences from different individuals. From this analysis, we have identified two additional DNA polymorphisms in the *G6PD* gene, again in people of African origin. Combining the data from a previous study with the current information obtained on these sites, we are able to provide additional evidence supporting a distinct pattern of evolution for the various resulting haplotypes.

MATERIALS AND METHODS

DNA Amplification. DNA was prepared by urea lysis and phenol/chloroform extraction (8) from the peripheral blood of 54 unrelated men of African origin (22 from Nigeria, 14 from Kenya, 14 from the West Indies, 2 from Ghana, and 2 from Sierra Leone), a small Nigerian family, and a number of unrelated white women. Three regions of the *G6PD* gene (see Fig. 1) were amplified from this genomic DNA using the polymerase chain reaction (PCR; ref. 9). The oligonucleotides used as primers were 5′-TGGACCCCTACACAGCCAAGTAC-3′ (oligonucleotide G.7) and 5′-GGCATGCTCCTGGGGACTGCT-3′ (oligonucleotide I) for region 1, 5′-GGAGCTAAGGCGAGCTCTGGC-3′ (oligonucleotide L) and 5′-TGCCTTGCTGGGCCTCGAAGG-3′ (oligonucleotide R) for region 2, and 5′-TGTTCTTCAACCCCGAGGAGT-3′ (oligonucleotide F) and 5′-AAGACGTCCAGGATGAGGTGATC-3′ (oligonucleotide Md) for region 3. The latter have been described previously (10). A proportion of the reaction product was digested with the appropriate restriction enzyme (see below) according to the manufacturers' instructions (New England Biolabs and Northumbria Biologicals) and then run on 1.5–2.5% agarose gels, stained with ethidium bromide, and photographed under UV light. Region 1 was digested with *Sac* I and *Sca* I together, region 2 was digested with *Bsp*HI alone, and region 3 was digested with *Bcl* I.

DNA Sequencing. All sequencing was performed on M13 phage clones using the chain-termination method (11) with either the Klenow fragment of DNA polymerase or a Sequenase kit. Introns 3, 6, and 9–12 were sequenced in both orientations during the sequencing of the exons of the G6PD variants Santiago de Cuba, G6PD Ilesha, and G6PD Mahidol using specific oligonucleotide primers and *Eco*RI subclones of genomic λ phage clones as described (12, 13). In addition to this sequence information, intron 7 and intron 8 [excluding nucleotides (nt) 270–360] were sequenced from *Pst* I fragments of the G6PD Mahidol phage clone that were subcloned into M13. For the sequence of introns 4 and 5, a 1.8-kilobase (kb) *Bgl* II fragment was purified using a Gene Clean kit from a 3.5-kb *Eco*RI subclone of the Santiago phage clone and cut briefly with DNase. A 300- to 600-base-pair (bp) fraction was

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: G6PD, glucose-6-phosphate dehydrogenase; RFLP, restriction fragment length polymorphism; IVS, intervening sequence; nt, nucleotide(s).

eluted after agarose gel electrophoresis, again by Gene Clean, and cloned into an M13 *Sma* I vector. Recombinant plaques were picked and sequenced at random. The sequencing of introns 1 and 2 was not attempted.

RESULTS

The comparison of the intron sequence obtained here (not shown) with the full sequence reported by Chen *et al.* (14) has revealed five base differences. One of these is the intervening sequence 5 (IVS5) nt 611 C → G mutation, previously identified as that responsible for a *Pvu* II polymorphism in the African population (5). The second is an A → G substitution at IVS2 nt 9722; this change does not create or destroy a restriction enzyme recognition sequence, and it has not been investigated further. The third is a T → C change at IVS11 nt 93, which creates an *Nla* III restriction site. Preliminary evidence, details of which are to be published elsewhere, shows that this site is uniformly present in people of African origin (in 28 of 28 chromosomes) but is predominantly absent in Europeans (present in 3 of 45 chromosomes analyzed).

The other two changes observed have been further investigated in this study. Both are C → T base substitutions, one at IVS7 nt 175 and the other at IVS8 nt 163. The latter destroys a *Bsp*HI recognition sequence (TCATGA) and is therefore amenable to direct detection by restriction enzyme digestion. The former does not create or destroy a restriction enzyme recognition sequence and so, for its detection, an oligonucleotide (G.7 in Fig. 1) was designed matching the sequence of intron 7 from nucleotide 152 to 174 except for a single deliberate mismatch 3 bp from the 3′ end of G.7. In this way, a *Sca* I site (AGTACT) is created after PCR amplification (using this and another oligonucleotide as primers) when there is a T at IVS7 nt 175 but not when there is a C at this position. This site is therefore shown in quotation marks (Fig. 1; see Fig. 3).

In 7 of 7 English women, amplified DNA that includes nucleotide 163 of intron 8 (region 2) was always cut with the enzyme *Bsp*HI, demonstrating the presence of a C residue at this position in 14 of 14 chromosomes. In the same series, *Sca* I failed to digest amplified DNA that includes IVS7 nt 175 (region 1), again demonstrating a C at this position in all cases. However, in 54 men of African descent the sites were found to be polymorphic: in 23 (43%) a T was found at IVS8 nt 163 and in 16 (30%) a T was found at IVS7 nt 175. The segregation of these two polymorphisms in a small Nigerian family is shown in Fig. 2. In addition, we have investigated the frequency of a silent mutation in exon 11 (nt 1311 C → T) in this series: 13 (24%) were found to have this mutation by the cleavage of amplified DNA from region 3 with *Bcl* I (not shown) as described (10). This frequency is in good agreement with data published elsewhere (15).

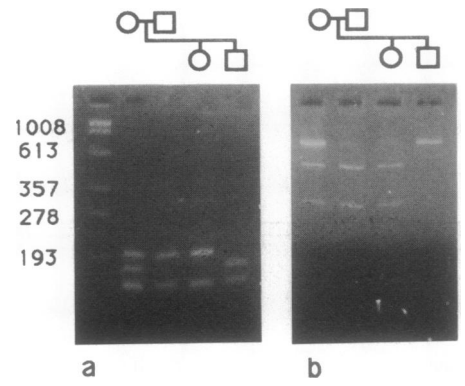

FIG. 2. Inheritance of two G6PD DNA polymorphisms in a small family. The family tree is drawn above the lanes of gels in which the amplified DNA from that member of the family has been loaded after digestion with the appropriate restriction enzyme. (*a*) Amplified DNA from region 1 digested with *Sac* I and *Sca* I, to yield fragments of 189 bp, 168 bp, and/or 153 bp. (*b*) Amplified DNA from region 2 digested with *Bsp*HI, to yield fragments of 794 bp and/or 293 bp and 501 bp. In both cases the mother is heterozygous, passing one allele to her daughter and the other to her son. Numbers indicate the sizes (bp) of the fragments generated by *Taq* I/*Pvu* II digestion of EMBL4 plasmid DNA, used as a standard marker and loaded in the left-hand lane of the gel in *a*.

Because this group of 54 men has already been studied with respect to the other polymorphic sites of the *G6PD* gene (7), we are able to combine the previous and the present data to determine the haplotype of seven polymorphic sites for each individual and therefore establish the haplotype frequencies in the group as a whole (Fig. 3). A C at IVS8 nt 163 [*Bsp*HI (+)] is found in all individuals with G6PD B, in 1 of 10 with G6PD A, and in none with G6PD A−. A T at IVS7 nt 175 ["*Sca* I" (+)] is present in all individuals with G6PD A−, in 2 of 10 with G6PD A, and in none with G6PD B; this latter distribution perfectly matches that of the *Pvu* II site. The silent mutation in exon 11 [nt 1311 T, "*Bcl* I" (+)] is found only in individuals with G6PD B.

DISCUSSION

DNA polymorphisms, identified as RFLPs, have proved to be invaluable tools in the study of the human genome, not least in providing markers for gene mapping and linkage analysis. An alternative approach to the identification of DNA polymorphism is through the direct comparison of DNA sequences from individual to individual. This approach has rarely been used in a deliberate way because it is much more laborious. However, as repetitive sequencing of the same gene from different people is carried out in the search for mutations that affect protein function, more information on the rate of silent polymorphic mutation will be obtained.

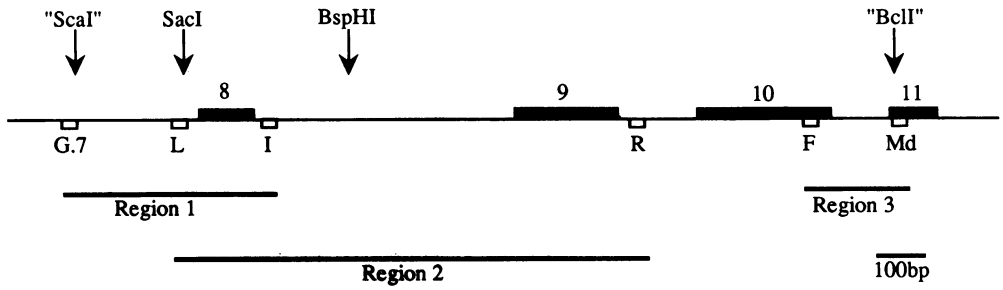

FIG. 1. Amplification of three regions of the *G6PD* gene. A sketch of a part of the *G6PD* gene shows the locations of the oligonucleotides (open boxes: G.7, L, I, R, F, and Md) that are used as primers for PCR. Exons 8–11 are shown as numbered black boxes and the locations of the relevant restriction sites are shown above. The restriction sites *Sca* I and *Bcl* I are shown in quotation marks as they are only generated by using mismatch-containing primers in a PCR. The extent of the three regions amplified is shown below.

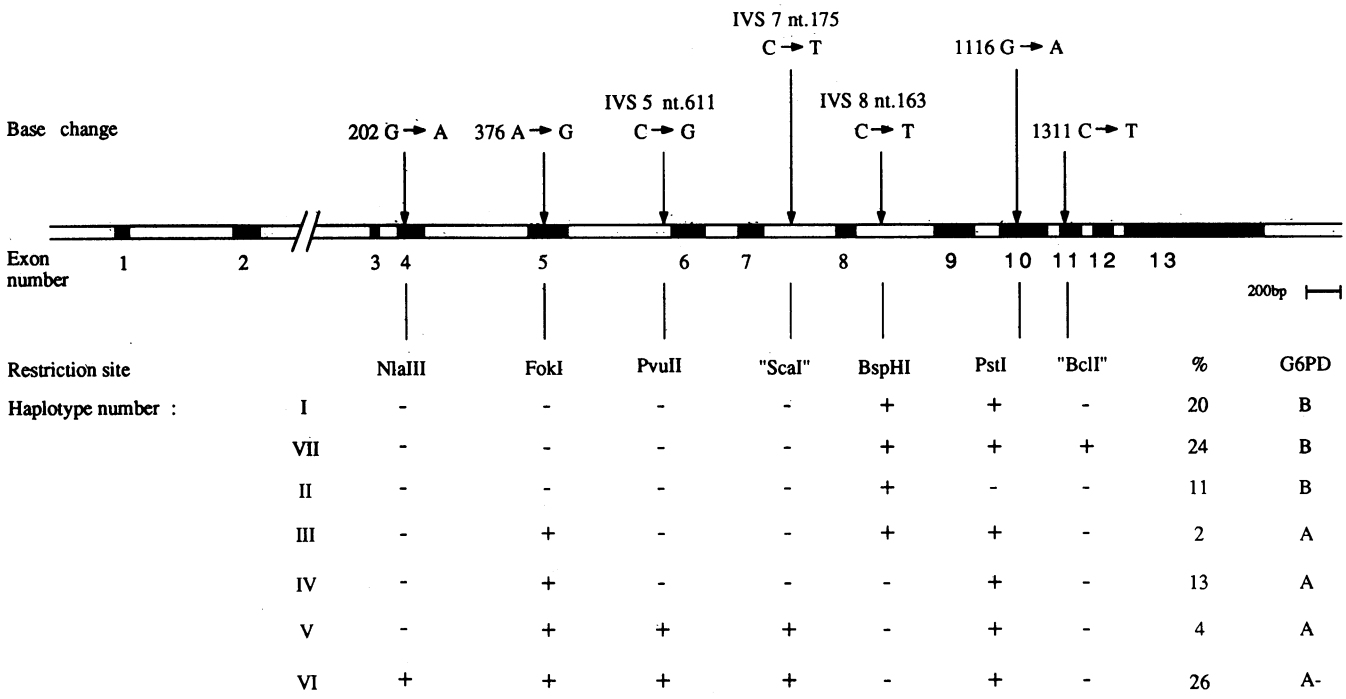

FIG. 3. Haplotypes of the *G6PD* gene in people of African origin. A sketch of the *G6PD* gene shows the locations of seven different polymorphic sites. The presence or absence of these sites is indicated below in the seven different haplotypes identified and designated by the Roman numerals to the left. The relative frequency found in this study is shown to the right.

To date, the "framework" polymorphisms of the β-globin gene are probably still the best examples of variation discovered in this way (16).

The comparative analysis of intron sequences of the *G6PD* gene that we originally made on an African, an Indian, and a Cuban variant (12) did not yield any such polymorphism. However, since the intron sequence of an Italian variant with the same coding sequence as G6PD A− (G6PD "Matera") became available (14), we have established the presence of two additional DNA polymorphisms in the *G6PD* gene of people of African origin. Therefore, together with the *Pst* I and *Pvu* II RFLPs reported previously, the silent mutation in exon 11, and the two mutations identified in G6PD A and A−, seven polymorphic sites are now known within 3 kb of the *G6PD* gene in African people. The use of these sites as X chromosome-linked genetic markers is largely limited to the black population, but the tightly linked haplotype data that these polymorphisms provide are informative in terms of the origin and spread of different mutations in the *G6PD* gene.

As expected, due to the fact that they are so close together, there is marked linkage disequilibrium between the different sites (Table 1): only 7 of the 128 possible different haplotypes have been observed so far; others that may exist must be too rare to be picked up in this series. From statistical analysis, it appears that the *Pst* I (−) allele is the only one approaching equilibrium, suggesting that it is quite ancient. The fact that the A mutation is found in the context of four different haplotypes suggests that it too is relatively ancient. By contrast, the two mutations in G6PD A− are only found in the same context with respect to the other polymorphic sites [*Pvu* II (+), "*Sca* I" (+), *Bsp*HI (−), *Pst* I (+), "*Bcl* I" (−)], indicating strongly that the combination has arisen only once. This conclusion is based on the most economical pattern of evolution for the different polymorphisms, with each of the mutations having arisen only once and with the A− allele as the most recent (Fig. 4). G6PD B is taken as the starting point for this evolutionary tree because it is by far the most common and also because the G6PD of the chimpanzee is

Table 1. Linkage disequilibrium within the *G6PD* gene

Restriction site	*Nla* III		*Fok* I		*Pvu* II		"*Sca* I"		*Bsp*HI		*Pst* I	
	E	O	E	O	E	O	E	O	E	O	E	O
"*Bcl* I" (+)	0.06	0	0.10	0*	0.07	0†	0.07	0†	0.10	0*	0.03	0
Pst I (−)	0.03	0	0.05	0	0.03	0	0.03	0	0.05	0		
*Bsp*HI (−)	0.11	0.26‡	0.19	0.43‡	0.13	0.30‡	0.13	0.30‡				
"*Sca* I" (+)	0.06	0.26‡	0.14	0.30‡	0.09	0.30‡						
Pvu II (+)	0.08	0.26‡	0.14	0.30‡								
Fok I (+)	0.12	0.26‡										

For each polymorphic site the frequency of the rarer allele has been calculated from the data shown in Fig. 3. The values are as follows: *Bcl* I (+) = 0.24, *Pst* I (−) = 0.11, *Bsp*HI (−) = 0.43, *Sca* I (+) = 0.30, *Pvu* II (+) = 0.30, *Fok* I (+) = 0.45, and *Nla* III (+) = 0.26. The expected frequency (E) of the simultaneous occurrence in the same DNA sample of the rarer allele at two restriction sites is calculated by multiplying the respective individual frequencies by each other. For instance, the expected frequency of the simultaneous occurrence of *Bsp*HI (−) and *Fok* I (+) is 0.43 × 0.45 = 0.19, and this is entered as the E value at the intersection of the *Bsp*HI row with the *Fok* I column. The observed frequency (O) of the simultaneous occurrence of *Bsp*HI (−) and *Fok* I (+) is 0.43 and this is entered as the O value. The difference E − O (= 0.24 in this example) is a measure of linkage disequilibrium. For each pair of sites the statistical significance of the linkage disequilibrium has been assessed by a χ^2 test (with Yates correction). *$P < 0.01$; †$P < 0.05$; ‡$P < 0.0001$.

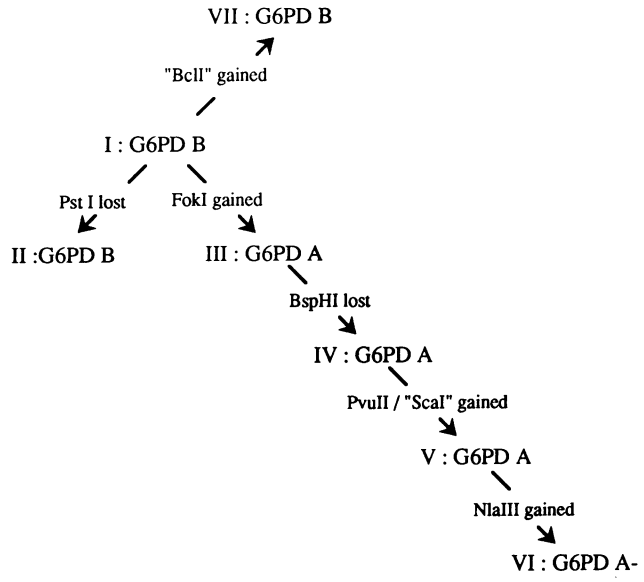

FIG. 4. Postulated evolutionary sequence of the different mutations in the *G6PD* gene based on the most simple progression from haplotype to haplotype (indicated by the Roman numerals), starting from the most common G6PD B haplotype.

B-like (17). The sequential order of the mutations is unambiguous, except that we do not know whether "*Sca* I" has appeared before or after *Pvu* II; this could be established by testing more samples.

Since the same haplotype is found in a G6PD A− gene from Southern Italy [the G6PD "Matera" sequenced by Chen *et al.* (14)], in a sample of G6PD Betica (not shown), which is the same as G6PD A− and is polymorphic in Spain (18), and in samples of G6PD A− from white men from Corsica and North Carolina (T.J.V. and L.L., unpublished), it seems reasonable to assume that it has spread to these places from Africa. The same is also true of the G6PD A− that has been observed in white people in Mexico (17).

A classical problem in population genetics is whether the spread of a gene is due to selection or to neutral mechanisms such as drift and founder effects. Sites that may themselves be neutral, but are closely linked to an allele that is selected for, may be enriched through this association. Alternatively, old and neutral mutations may remain in their own right in an expanded and mixed population because they had become fixed in the small population groups in which they arose. We hope that the haplotypes described here, which involve the G6PD A− allele that is thought to be selected for due to the relative protection it provides against malaria mortality in heterozygous females (19) as well as other alleles that are more ancient and presumed to be neutral, will help in the future investigation of these possibilities.

We thank Jean-Luis Vives-Corrons for supplying a DNA sample from a subject with G6PD Betica and Viola Calabro for reviewing the manuscript. This work was supported by a Medical Research Council (U.K.) Programme Grant and by a European Economic Community Contract (SC1 CT90 0573). A.O. was supported by a medical fellowship from the Commonwealth Scholarship Commission.

1. Beutler, E. (1990) *Sem. Haematol.* **27,** 137–164.
2. Luzzatto, L. & Mehta, A. (1989) in *The Metabolic Basis of Inherited Disease*, eds. Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D. (McGraw-Hill, New York), pp. 2237–2266.
3. D'Urso, M., Luzzatto, L., Peroni, L., Ciccodicola, A., Gentile, G., Peluso, I., Persico, M. G., Pizzella, T., Toniolo, D. & Vulliamy, T. J. (1988) *Am. J. Hum. Genet.* **42,** 735–741.
4. Yoshida, A., Takizawa, T. & Prchal, J. T. (1988) *Am. J. Hum. Genet.* **42,** 872–876.
5. Fey, M. F., Wainscoat, J. S., Mukwala, E. C., Falusi, A. G., Vulliamy, T. J. & Luzzatto, L. (1990) *Hum. Genet.* **84,** 471–472.
6. Beutler, E. & Kuhl, W. (1990) *Hum. Genet.* **85,** 9–11.
7. Vulliamy, T. J., Othman, A., Town, M., Nathwani, A., Falusi, Y. & Luzzatto, L. (1991) *Gene Geogr.*, in press.
8. Sykes, B. C. (1983) *Lancet* **ii,** 787–788.
9. Saiki, R., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239,** 487–491.
10. Kurdi-Haidar, B., Mason, P. J., Berrebi, A., Ankra-Badu, G., Al-Ali, A., Oppenheim, A. & Luzzatto, L. (1990) *Am. J. Hum. Genet.* **47,** 1013–1019.
11. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143,** 161–178.
12. Vulliamy, T. J., D'Urso, M., Battistuzzi, G., Estrada, M., Foulkes, N. S., Martini, G., Calabro, V., Poggi, V., Giordano, R., Town, M., Luzzatto, L. & Persico, M. G. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 5171–5175.
13. Vulliamy, T. J., Wanachiwanawin, W., Mason, P. J. & Luzzatto, L. (1989) *Nucleic Acids Res.* **17,** 5868.
14. Chen, E. Y., Cheng, A. L., Kuang, W.-J., Weigelt, L., Green, P., Schlessinger, D., Ciccodicola, A. & D'Urso, M. (1991) *Genomics* **10,** 792–800.
15. Beutler, E. & Kuhl, W. (1990) *Am. J. Hum. Genet.* **47,** 1008–1012.
16. Orkin, S. H., Kazazian, H. H., Antonarakis, S. E., Goff, S. C., Boehm, C. D., Sexton, J. P., Waber, P. G. & Giardina, P. J. V. (1982) *Nature (London)* **296,** 627–631.
17. Beutler, E., Kuhl, W. & Vives-Corrons, J.-L. (1989) *Blood* **74,** 2550–2555.
18. Vives-Corrons, J.-L. & Pujades, A. (1982) *Hum. Genet.* **60,** 216–222.
19. Luzzatto, L. (1979) *Blood* **54,** 961–976.