Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 544–548, January 1988
Genetics

Direct sequencing of enzymatically amplified human genomic DNA

(DNA sequencing/polymerase chain reaction/mutation/oligonucleotide/globin)

DAVID R. ENGELKE*†, PATRICIA A. HOENER*, AND FRANCIS S. COLLINS‡

Departments of *Biological Chemistry, ‡Internal Medicine and Human Genetics, and ‡The Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI 48109

Communicated by James V. Neel, October 5, 1987

ABSTRACT **The polymerase chain reaction is a recently described technique that uses flanking oligonucleotide primers and repeated cycles of enzymatic primer extension to amplify a short segment of DNA by >100,000-fold. By use of sequencing primers located internal to the amplification primers, direct genomic sequence was obtained from enzymatically amplified DNA by using the dideoxynucleotide chain-termination method. The method is relatively simple and offers significant advantages in identifying mutations in genes for which the normal sequence is known. Heterozygous and homozygous mutations in the human β- and γ-globin loci were unambiguously identified in 3 days with <1 μg of genomic DNA.**

The identification of mutations and polymorphisms in human genes by DNA sequencing contributes substantially to our understanding of the molecular nature of disease and has a variety of practical applications in diagnosis. Sequence information is usually obtained by cloning the genes from individuals expected to harbor a mutation and then comparing the sequence of the cloned DNA to that of the normal allele. The cloning process is laborious, however, requiring several weeks of library construction and screening to obtain DNA clones. Moreover, if the individual is not homozygous, several clones may have to be studied to be certain that both alleles are detected.

Several methods of screening for mutations have recently been developed that do not require cloning. Carefully controlled hybridization with mutation-specific oligonucleotide probes (1), for example, can detect the presence of single base changes in short segments of DNA if the exact nature of the mutation is known in advance. The location of mutations in a few genes has also been successfully identified by ribonuclease cleavage of mismatches between a labeled normal antisense RNA and genomic DNA (2) or by cleavage of RNA·RNA duplexes of a labeled antisense RNA and cellular RNA from an affected tissue (3, 4). Mutations in genomic DNA have also been detected by denaturing gradient gel electrophoresis (5). These approaches, however, do not provide the actual sequence at the mutation site, and some mismatches are not detected. Methods have been reported for direct chemical sequencing of mammalian genomic DNA (6), but these have not yet been widely adopted due to the degree of technical difficulty. Direct dideoxynucleotide sequencing (7) of eukaryotic genomic DNA has previously only been achieved in yeast (8).

The development of the polymerase chain reaction (PCR; ref. 9) to selectively amplify a short segment of DNA by >100,000-fold has reduced the problem of sequence complexity and signal strength as impediments to direct analysis of single-copy DNA sequences. The PCR has been used to enrich for segments of the human β-globin (9–11), HLA-DQα (10, 11), and c-*ras* (12) genes and has recently been used to acquire direct sequence from human mitochondrial DNA (12). The misincorporation rate of PCR DNA synthesis is low [1 in 600 base pairs (bp) or less], but in a cloning study only 1–20% of the amplified DNA corresponded to the desired sequence, with the remainder arising from a background of priming from other genomic sequences (10). We now report that this background can be successfully overcome by using a "nested" set of PCR amplification primers (13, 14) and dideoxynucleotide chain-termination sequencing primers that hybridize internally on the amplified DNA.

MATERIALS AND METHODS

Materials. Human DNA was prepared as described (15) from lymphoblastoid cells or peripheral blood leukocytes and digested with restriction endonuclease *Hin*dIII prior to amplification. Oligodeoxyribonucleotides were synthesized on an Applied Biosystems model 380B DNA synthesizer and purified on 15% or 20% polyacrylamide/8 M urea gels (16). Deoxyribonucleotides and dideoxyribonucleotides were obtained from Pharmacia P-L Biochemicals; [γ-^{32}P]ATP was from New England Nuclear; proteinase K was from Beckman Instruments; T4 polynucleotide kinase was from United States Biochemical; and avian myeloblastosis virus (AMV) reverse transcriptase was from Life Sciences (St. Petersburg, FL). The large fragment (Klenow) of *Escherichia coli* DNA polymerase I was prepared from strain CJ155 (17).

DNA Amplifications. DNA amplifications were carried out by the method of Saiki *et al* (11) with minor modifications. Amplification reaction mixtures contained (in 100 μl) 2 μg of template DNA, 0.5 μg (each) of purified oligodeoxynucleotide primer, 1.5 mM (each) dATP, dGTP, dCTP, and dTTP, 100 mM Hepes (pH 7.5) buffer, 6 mM $MgCl_2$, and 50 mM NaCl. In our hands the use of Hepes rather than Tris buffer helped maintain the integrity of the DNA. Other variations from the protocol of Saiki are as follows: after the denaturation and centrifugation steps in each round, the sample was incubated only 20 sec in a room temperature water bath before adding 1 μl (20–30 units) of the large fragment of *E. coli* DNA polymerase I. Amplification products were probed in reaction mixtures containing, in 40 μl, 50 mM Hepes (pH 7.9) buffer, 6 mM $MgCl_2$, 40 mM KCl, 0.6 mM (each) dATP, dGTP, dCTP, and dTTP, 1.0 μl of the amplification reaction mixture, and ≈100,000 dpm of the indicated radiolabeled oligonucleotide primer {1000–7000 Ci/mmol (1 Ci = 37 GBq), labeled using T4 polynucleotide kinase and [γ-^{32}P]ATP}. Radiolabeling extension products by incorporation of [α-^{32}P]dNTPs rather than prelabeled primer results in unacceptable background. Reaction mixtures were incubated at 95–100°C for 5 min and then incubated at 50°C for 15 min to hybridize the primer. One microliter (8–12 units) of

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: PCR, polymerase chain reaction; AP, amplification primer; IP, internal primer; HPFH, hereditary persistence of fetal hemoglobin; AMV, avian myeloblastosis virus.

†To whom reprint requests should be addressed.

AMV reverse transcriptase (Life Sciences) was added and the extension reaction was allowed to proceed at 50°C for 20 min before 8 μl of stop mix (1 mg of protease K per ml/2% $NaDodSO_4$/100 mM EDTA, pH 8) was added and the sample was incubated at least 30 min at 50°C. DNA was recovered by precipitation with 110 μl of ethanol, resuspended in loading buffer (98% deionized formamide/10 mM NaOH/1 mM EDTA/0.05% xylene cyanol), and subjected to electrophoresis in polyacrylamide sequencing gels (16) containing 8 M urea. The autoradiographic exposure shown in Fig. 2 was for 2 days on Cronex 4 film (DuPont) with a Lightning Plus intensifying screen.

DNA Sequencing Reactions. Amplified DNA for sequencing reactions was recovered by digestion at 50°C with 0.2 vol of stop mix for at least 2 hr; this was followed by precipitation with 2.5 vol of ethanol and washing the precipitate with 75% ethanol. Dideoxynucleotide chain-termination sequencing reaction mixtures using AMV reverse transcriptase were the same as the radiolabeled probe reaction mixtures described above except that the reaction mixtures contained four times as much template, 0.007 mM dideoxynucleotide triphosphate, 0.022 mM deoxynucleotide triphosphate of the same type, and 0.22 mM of the other three deoxynucleotide triphosphates (8). Autoradiographic exposures varied from 10 hr to 4 days depending on the specific radioactivity of the oligonucleotide primers.

RESULTS

The two segments of genomic DNA chosen for testing are shown schematically in Fig. 1 and were selected because of the presence of known mutations with important *in vivo* consequences that occur in these regions. In the region upstream from the human fetal globin ($^{G}\gamma$ and $^{A}\gamma$) genes a number of single base changes have been identified (18, 19) in association with hereditary persistence of fetal hemoglobin (HPFH). A large number of mutations have been identified within the 5′ portion of the human β-globin gene, which lead to hemoglobinopathies (18, 19). APs were 17–30 nucleotides in length and were chosen to match the normal sequence. PCR amplification was carried out with the Klenow fragment of DNA polymerase I, and the DNA was probed by hybridizing radiolabeled IPs and by extending with AMV reverse transcriptase. As seen in Fig. 2, γ-globin promoter DNA amplified from the γ-AP1 and γ-AP2 primers continued to increase steadily for at least 23 rounds, whereas the β-globin DNA signal was weak and ceased increasing entirely after 15–17 rounds. This difference is unlikely to be due solely to the increased distance between the β primers, since varying the distance between a fixed β-AP1 primer and different β-AP2 primers from 150 to 350 bp resulted in equally poor signal strengths. Similarly, the signal remained strong in promoter amplification when γ-AP1 and γ-AP2 were moved 150 bp apart (data not shown). Cessation of β amplification may be due partially to depletion of nucleotides or primers from the reaction mixture, since dilution of an aliquot into a new reaction mixture after 15 rounds frequently resulted in a moderate increase in signal ("+AP1/AP2"). A larger and more consistent increase was observed, however, when the new reaction mixture contained at least one new amplification primer ("+AP1/AP3") that was nested inside the original set. In this case the signal after 8–10 additional rounds of amplification was nearly equivalent to that of the γ-promoter amplification after 23 rounds (Fig. 2). Although the reasons for the poor two-primer amplification at the β locus remain obscure, it seems possible that the phenomenon is linked to the percentage of the overall events in each round that occur at the desired priming site. As in the sequencing reactions described below, the nested primer in the three-primer amplification would select a subset of the originally amplified DNA, thus greatly increasing the percentage of DNA synthesis comprised by the desired amplification. Several different third primers have all given the same result for the nested β amplification (not shown), indicating that this may prove a generally acceptable method of circumventing poor amplification at any given locus.

An example of the data obtained when the amplified DNA is subjected to dideoxynucleotide chain-termination sequence analysis is shown in Fig. 3. DNA from a homozygote for the sickle mutation (codon 6, GAG to GTG) was amplified by a combination of the β-AP1, β-AP2, and β-AP3 primers. Both strands were subsequently sequenced from the β-IP1 and β-IP2 primers. In repeated experiments the sequence data obtained are of high quality, comparable to that obtained with cloned DNA templates. Occasional am-

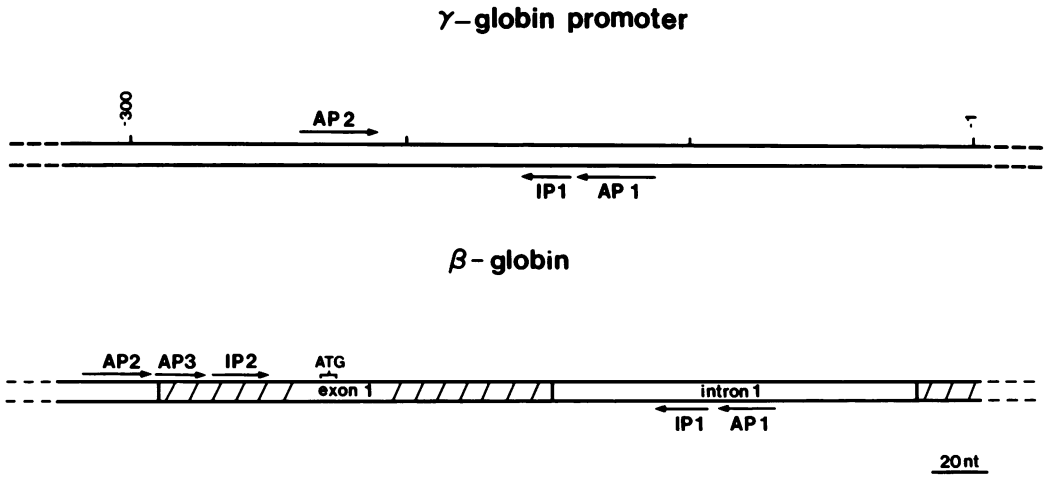

FIG. 1. Amplification and sequencing strategy. The regions of the human β-globin and fetal globin genes chosen for amplification and direct sequencing are shown along with the oligodeoxynucleotide amplification primers (APs) and internal probe primers (IPs) used in the experiments. Arrows indicate the 5′ → 3′ orientation of the primers. The fetal globin primers hybridize to conserved nucleotide sequences (18) in the $^{G}\gamma$ and $^{A}\gamma$ promoter regions at positions −118 to −147 (γ-AP1), −207 to −231 (γ-AP2), and −152 to −168 (γ-IP1) wth respect to the transcription initiation site. The APs for the β-globin genes hybridize at positions 58–78 in the first intron (β-AP1) and positions −55 to −75 (β-AP2) and −38 to −54 (β-AP3) with respect to the translation initiator ATG. β-IP1 and β-IP2 are 17-mers that hybridize immediately adjacent to β-AP1 and β-AP2, respectively. nt, Nucleotides.

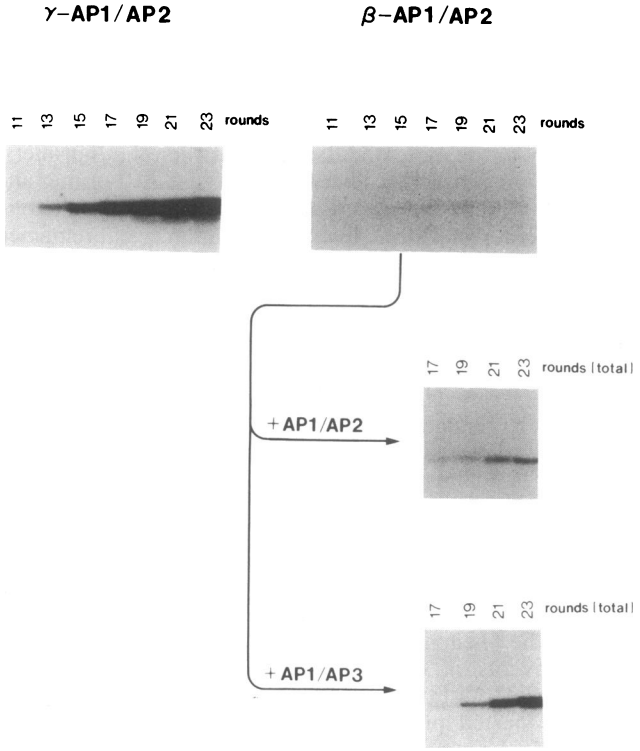

FIG. 2. Two-primer versus three-primer amplification. Two micrograms of leukocyte DNA was amplified in 100-μl reaction mixtures by the PCR method. Initially two reaction mixtures were set up, one containing the γ-AP1 and γ-AP2 primers and one containing the β-AP1 and β-AP2 primers. After 15 rounds of amplification, two 20-μl aliquots of the β-AP1/AP2 reaction mixture were withdrawn and diluted into two new 100-μl reaction mixtures containing fresh buffer, primers, and oligonucleotides. One of the new reaction mixtures contained more added β-AP1 and β-AP2, whereas the other contained β-AP1 and β-AP3. Amplification of the original reactions and the two new reactions was then continued for an additional 8 rounds (rounds 16–23). After the indicated number of rounds, aliquots of 0.2 μl from the original reaction mixtures or 1 μl from the new reaction mixtures were withdrawn. The DNA was probed by hybridizing a radiolabeled internal primer (γ-IP1 or β-IP1), extending with AMV reverse transcriptase, and analyzing the radiolabeled extension products by size on polyacrylamide sequencing gels. The products shown were the only visible bands in the autoradiographs and, as judged by length and subsequent sequence analysis, represent run-off termination at positions at or near the 5′ ends of γ-AP2 (γ-AP1/AP2 amplification), β-AP2 (β-AP1/AP2-only amplifications), or β-AP3 (β-AP1/AP2, then AP1/AP3 amplification). In the bottom panel a minor band slightly longer than the run-off product at the β-AP3 site does not increase in intensity between rounds 17 and 23 and is presumed to arise from run-off at the β-AP2 site due to residual β-AP1/AP2-amplified DNA.

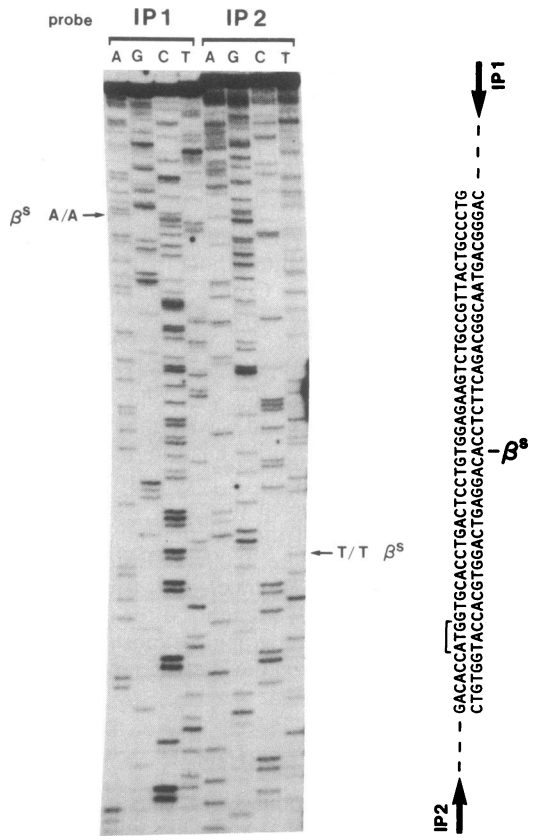

FIG. 3. Sequence analysis of amplified DNA. Two micrograms of DNA prepared from peripheral leukocytes of an individual homozygous for sickle cell anemia was amplified by using the β-AP1/AP2/AP3 three-primer method with the secondary amplification carried out for a total of 10 rounds. Four percent of the recovered DNA was used as a template in each dideoxynucleotide chain-termination sequencing reaction (A, G, C, or T). Sequence was obtained for both strands by using either β-IP1 or β-IP2 as the radiolabeled primer. The position of the homozygous sickle mutation (β^S) in the sequence of each strand is indicated to the sides of the panel. The known nucleotide sequence of both strands (18) in the region of the β^S mutation is given for reference to the right of the panels. Duplicate samples of each reaction mixture were electrophoresed for longer times (not shown) to ensure accurate reading of the β-IP1 reaction at least 25 nucleotides past the β^S position. No ambiguities occurred at the same position in the sequence of both strands.

biguous positions are all resolved by referring to the sequence of the opposite strand. No traces of sequence corresponding to the δ-globin gene, which could be present in the amplified DNA due to significant homology in this region (18), are detected. When DNA from a heterozygote for the sickle mutation was analyzed in comparison with the sickle homozygote (Fig. 4 *Upper*), the the presence of the normal base (adenine) and the sickle mutation base (thymine) was easily identified. Other ambiguities in the sequence do not appear on both strands and are most likely reverse transcriptase artifacts (e.g., the starred position in Fig. 4). We have also used this same set of primers to detect heterozygotes for β-thalassemia caused by either a G → C mutation at position 5 or a T → C mutation at position 6 of intron 1 as well as a heterozygote for a frameshift mutation at codon 8/9 (data not shown). Fourteen of the 34 described β-thalassemia mutations, and more than 50 β-chain hemoglobin variants, would be detected by this set of primers (19).

DNA from the γ-promoter region was sufficiently amplified by the γ-AP1 and γ-AP2 primers that it yielded unambiguous sequence. As shown in Fig. 4 *Lower*, it is possible to identify a heterozygote for $^{G}\gamma$ HPFH, which carries a C → G mutation at position −202 relative to the cap site of the $^{G}\gamma$ gene. Because the $^{G}\gamma$ and $^{A}\gamma$ promoter sequences are normally identical in this region, both promoters are amplified. The mutation is thus present in only 25% of the DNA but is still easily detected.

DISCUSSION

With the widespread success in cloning human genes, a rapid means of detecting disease-causing mutations or polymorphisms in these cloned genes is of increasing importance (for a review, see ref. 20). In addition to providing a means of accurate prenatal and postnatal diagnosis, identification of such sequence changes can offer clues about structure–function relationships and the mechanism of mutation. Molecular cloning of mutant alleles from affected individuals has usually been the method of identification. Orkin and Kazazian

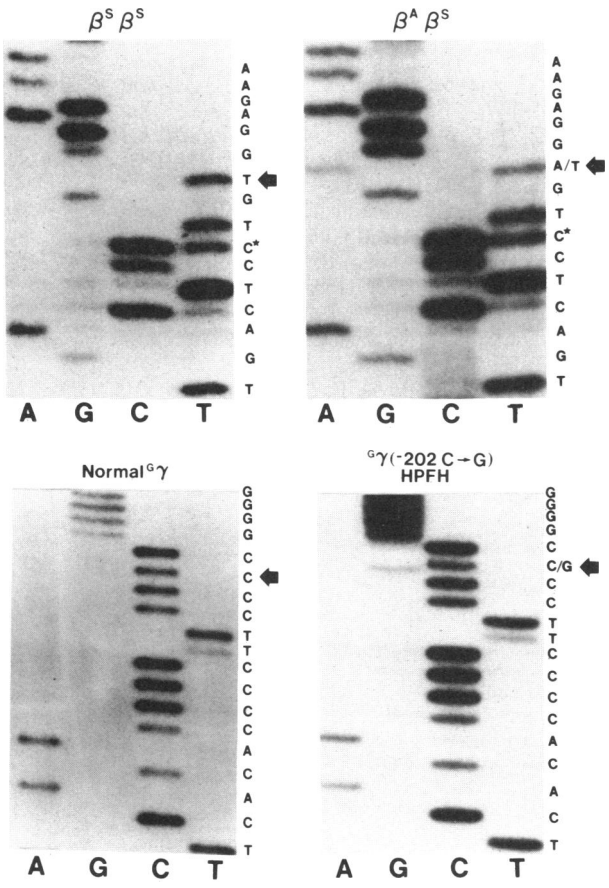

FIG. 4. Detection of heterozygous mutations. Leukocyte DNA from a homozygote ($\beta^S\beta^S$) or a heterozygote ($\beta^A\beta^S$) for sickle cell anemia was amplified by using β-AP1/AP2/AP3. Normal DNA or DNA from an individual previously shown to be heterozygous for $^G\gamma$ HPFH, carrying the C → G mutation at −202 relative to the cap site, was amplified by using the γ-AP1 and γ-AP2 primers. The sequence of the sense strand is displayed on the right of each panel with the positions of the β^S and $^G\gamma$ HPFH −202 mutations denoted by heavy arrows. For the amplified β-globin DNA, sequence from the β-IP2 primer is shown for the region of the β^S mutation. The A and T bands at the β^S position in the heterozygote DNA are consistently of roughly equal intensity. The heterogeneity at the starred C position in the β sequences is an artifactual reverse transcriptase pause and does not appear in the sequence of the complementary strand (see Fig. 3). Portions of the sequence from the γ-IP1 primer are shown in the cases of the amplified γ-globin DNA in the −200 region. In independent experiments, the wild-type C band at position −202 in the $^G\gamma$ HPFH DNA was more intense than the G band at the same position, as expected since the $^G\gamma$ and $^A\gamma$ promoters are both amplified in this experiment and therefore only one sequence out of four carries the C → G mutation in this HPFH heterozygote.

(21), for example, have used this approach to identify a large number of point mutations and frameshifts that lead to β-thalassemia. An advantage of the cloning approach is that one can then place the mutant allele into an expression vector to study its mechanism. In many situations, however, the mechanism is obvious or the behavior of the altered protein can be directly analyzed. In such settings, a method of detecting mutations that does not require the time-consuming generation and screening of a genomic library would be advantageous. Recently described methods of point mutation detection that depend on denaturing gradient gels (5) or ribonuclease cleavage of mismatched RNA·DNA or RNA·RNA duplexes (2–4) avoid the need for cloning but do not detect all possible single-base mutations. These methods also indicate only the approximate position of a mutation and not its actual sequence.

Direct genomic sequencing avoids many of these difficulties but has in the past been technically challenging because of the complexity of the human genome and the consequent signal-to-noise problem (6). We show here that use of the PCR to selectively amplify the region of interest greatly reduces this problem and allows direct characterization of heterozygous and homozygous mutations in human genomic DNA from <1 μg of starting material and in a time period of about 3 days. Homozygous mutations and mutations in males with X-linked conditions will be particularly easy to detect as will frameshift mutations since these result in a "double image" sequence above the site of the frameshift. Direct genomic sequencing may also provide a convenient approach to the analysis of evolutionary differences between closely related species, to the frequency of germ-line and somatic mutations at the DNA level, and to the detection of sequence polymorphisms for linkage analysis. As with conventional sequencing, both strands of amplified DNA should be sequenced and a normal genomic DNA should be run in parallel in order to avoid being misled by sequencing artifacts (an example is shown in Fig. 4 *Upper*). The PCR using the Klenow fragment is known to have an error rate of ≈1/600 bp, as determined by cloning of amplified DNA (10). This should not present a problem for direct sequencing, however, since any given nucleotide misincorporation will not comprise a significant proportion of the amplified DNA.

Several features of the procedure require further investigation. It is unclear why the β-globin amplifications were only successful when a third nested amplification primer was used, though the γ-globin promoter amplifications were quite successful with two primers (Fig. 2). Sequence preference of the Klenow enzyme, secondary structure of the renatured genomic DNA, and mishybridizing of the primers to other related genomic sequences are possible causes. It is not yet clear how far apart the amplification primers can be placed before efficiency drops to unacceptable levels; we noted no loss of efficiency when the primers were placed up to 350 bp apart, but even larger segments of amplification would allow more convenient analysis of larger genes. However, even with present protocols, the synthesis of amplification and sequencing primers for the promoter, exons, and exon–intron boundaries of small- to moderate-sized genes is feasible. The use of alternative polymerase enzymes to perform the amplification and/or sequencing should also be investigated. By performing the first primer extension reaction with reverse transcriptase, for example, it should also be possible to carry out this procedure with RNA.

Note Added in Proof. McMahon *et al.* (22) have recently reported on direct genomic sequencing of sequences in the c-Ki-*ras* gene. In addition, PCR using the heat-stable *Thermus aquaticus* polymerase followed by direct genomic sequencing has recently been used to characterize several β-thalassemia mutations (23).

We thank D. Ginsburg and D. Gumucio for many helpful discussions, A. Krikos for critical reading of the manuscript, C. Bruzdzinski, S. L. Thein, and G. Atweh for genomic DNA, N. Grindley for supplying the CJ155 strain, and D. Rennert for typing the manuscript. This research was supported by National Science Foundation Grant DMB-8603115 and National Institutes of Health Grant RO1 GM34869 to D.R.E. and by Grants from the Cooley's Anemia Foundation and the March of Dimes (5-499) to F.S.C., who also gratefully acknowledges support from the Howard Hughes Medical Institute.

1. Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L. & Wallace, R. B. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 278–282.
2. Myers, R. M., Larin, Z. & Maniatis, T. (1985) *Science* **230,** 1242–1246.
3. Gibbs, R. A. & Caskey, C. T. (1987) *Science* **236,** 303–305.

4. Forrester, K., Almoguera, C., Han, K., Grizzle, W. E. & Perucho, M. (1987) *Nature (London)* **327,** 298–303.
5. Myers, R. M., Lumelsky, N. & Maniatis, T. (1985) *Nature (London)* **313,** 495–498.
6. Church, G. M. & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1991–1995.
7. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
8. Huibregtse, J. M. & Engelke, D. R. (1986) *Gene* **44,** 151–158.
9. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230,** 1350–1354.
10. Scharf, S. J., Horn, G. T. & Erlich, H. A. (1986) *Science* **233,** 1076–1078.
11. Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1986) *Nature (London)* **324,** 163–166.
12. Bos, J. L., Fearon, E. R., Hamilton, S. R., Verlaan-deVries, M., Van Boom, J. H., Van der Eb, A. J. & Vogelstein, B. (1987) *Nature (London)* **327,** 293–297.
13. Wrischnik, L. A., Higuchi, R. G., Stoneking, M., Erlich, H. A., Arnheim, N. & Wilson, A. C. (1987) *Nucleic Acids Res.* **15,** 529–542.
14. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. & Erlich, H. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51,** 263–273.
15. Goossens, M. & Kan, Y. W. (1981) *Methods Enzymol.* **76,** 805–817.
16. Sanger, F. & Coulson, A. R. (1978) *FEBS Lett.* **87,** 107–110.
17. Joyce, C. M. & Grindley, N. D. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1830–1834.
18. Collins, F. S. & Weissman, S. M. (1984) *Prog. Nucleic Acids Res. Mol. Biol.* **31,** 315–462.
19. Bunn, H. F. & Forget, B. G. (1986) *Hemoglobin: Molecular, Genetic and Clinical Aspects* (Saunders, Philadelphia).
20. Caskey, C. T. (1987) *Science* **236,** 1223–1229.
21. Orkin, S. H. & Kazazian, H. H. (1984) *Annu. Rev. Genet.* **18,** 131–171.
22. McMahon, G., Davis, E. & Wogan, G. N. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4974–4978.
23. Wong, C., Dowling, C. E., Saiki, R. K., Higuchi, R. G., Erlich, H. A. & Kazazian, H. H. Jr. (1987) *Nature (London)* **330,** 384–386.