REFERENCES AND NOTES

1. J. L. King and T. H. Jukes, *Science* **164**, 788 (1969); M. Kimura, *Nature* **267**, 275 (1977).
2. M. Kimura, *Nature* **217**, 624 (1968).
3. T. Miyata, T. Yasunaga, T. Nishida, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7328 (1980); A. C. Wilson, H. Ochman, E. M. Prager, *Trends Genet.* **3**, 241 (1987).
4. W.-H. Li *et al.*, *J. Mol. Evol.* **25**, 330 (1987).
5. P. M. Sharp and W.-H. Li, *ibid.* **28**, 398 (1989).
6. H. Ochman and A. C. Wilson, *ibid.* **26**, 74 (1987).
7. P. M. Sharp and W.-H. Li, *Mol. Biol. Evol.* **4**, 222 (1987).
8. M. Gouy and C. Gautier, *Nucleic Acids Res.* **10**, 7055 (1982); T. Ikemura, *Mol. Biol. Evol.* **2**, 13 (1985).
9. T. Ikemura, in *Population Genetics and Molecular Evolution*, T. Ohta and K. Aoki, Eds. (Japan Scientific Societies Press, Tokyo, 1985), pp. 385–406.
10. The genes are (in increasing order of silent site divergence): *rpsU, hupA, ptsH, metJ, pyrE, rpoB, ompA, ptsI, hupB, crr, ilvM, rpoD, prsA, metB, dnaA, hisG, glnA, ilvY, araA, carA, ompR, envZ, cyaA, hisC, hisA*, orf (5′ to *pyrE*), *crp, cysK, hisP, hisH, dnaB, hisF, araB, araD, ilvA, cheZ, cheR, trpB, metF, hisD, trpE, hisB, trpD, pyrI, hisIE, pyrB, dnaG, pabA, cheB, hisM, cheW, pyrC, mglB, araC, aroA, glgC, cheA*, orf (5′ to *metC*), *tar, trpC, sulA, pabB, cheY, cysB, metC, cysZ, trpA*. Data were taken from the GenBank/EMBL DNA sequence data library wherever possible; all sequence data and sources are available on request.
11. B. J. Bachmann, *Microbiol. Rev.* **47**, 180 (1983).
12. M. Riley and S. Krawiec, in Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 967–981; K. E. Sanderson and J. R. Roth, *Microbiol. Rev.* **52**, 485 (1988).
13. M. B. Schmid and J. R. Roth, *J. Bacteriol.* **169**, 2872 (1987).
14. S. R. Kushner, in Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 1044–1053.
15. W. D. Rupp and P. Howard-Flanders, *J. Mol. Biol.* **31**, 291 (1968); A. K. Ganesan, *ibid.* **87**, 103 (1974); S. C. West, E. Cassuto, P. Howard-Flanders, *Nature* **294**, 659 (1981).
16. R. F. DuBose, D. E. Dykhuizen, D. L. Hartl, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 7036 (1988); R. Milkman and A. Stoltzfus, *Genetics* **120**, 359 (1988).
17. Direct investigation of mutation rates at different map positions in *E. coli* revealed one potential hotspot at 58 to 60 min [J. van Brunt and G. Edlin, *J. Mol. Evol.* **5**, 279 (1975)]. In the related enterobacterium *Serratia marcescens*, there is G+C content heterogeneity among genes, which may be due to mutation pattern variation [M. Nomura, F. Sor, M. Yamagishi, M. Lawson, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 658 (1987)], though it could be explained more simply by the differential action of natural selection on codon choice in the face of a constant mutation pressure to G+C–richness (P. M. Sharp, *Mol Microbiol.*, in press).
18. K. H. Wolfe, P. M. Sharp, W.-H. Li, *Nature* **337**, 283 (1989).
19. J. Filipski, *J. Theor. Biol.* **134**, 159 (1988).
20. C. H. Martin and E. M. Meyerowitz, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8654 (1986).
21. C.-I. Wu and W.-H. Li, *ibid.* **82**, 1741 (1985); R. Kikuno, H. Hayashida, T. Miyata, *Proc. Japan Acad. Ser. B* **61**, 153 (1985); R. J. Britten, *Science* **231**, 1393 (1986).
22. W.-H. Li, C.-I. Wu, C.-C. Luo, *Mol. Biol. Evol.* **2**, 150 (1985).
23. P. M. Sharp and W.-H. Li, *Nucleic Acids Res.* **15**, 1281 (1987).
24. This paper is dedicated to the memory of Alan Robertson, who died on 25 April 1989. Supported in part by grants from the European Community (to P.M.S.) and NIH (to W.-H.L.). We thank D. J. McConnell for his comments on the manuscript.

21 June 1989; accepted 22 September 1989

Genomic Sequencing and Methylation Analysis by Ligation Mediated PCR

GERD P. PFEIFER, SABINE D. STEIGERWALD, PAUL R. MUELLER, BARBARA WOLD, ARTHUR D. RIGGS*

Genomic sequencing permits studies of in vivo DNA methylation and protein-DNA interactions, but its use has been limited because of the complexity of the mammalian genome. A newly developed genomic sequencing procedure in which a ligation mediated polymerase chain reaction (PCR) is used generates high quality, reproducible sequence ladders starting with only 1 microgram of uncloned mammalian DNA per reaction. Different sequence ladders can be created simultaneously by inclusion of multiple primers and visualized separately by rehybridization. Relatively little radioactivity is needed for hybridization and exposure times are short. Methylation patterns in genomic DNA are readily detectable; for example, 17 CpG dinucleotides in the 5′ region of human X-linked PGK-1 (phosphoglycerate kinase 1) were found to be methylated on an inactive human X chromosome, but unmethylated on an active X chromosome.

METHYLATION OF CpG DINUCLEOTIDES in critical regions of many vertebrate genes may be part of a gene silencing mechanism involved in cell differentiation, X chromosome inactivation, and genomic imprinting (*1, 2*). Methylation-sensitive restriction endonucleases are commonly used to determine in vivo methylation patterns, but this limits the analysis to a small subset of all CpG dinucleotides. Another method for methylation analysis is genomic sequencing (*3*), a method that retains information normally lost during cloning, such as the location of 5-methylcytosines (*3*) and DNA-protein interactions (*4*). Genomic sequencing has, however, been difficult, requiring large amounts of radioactivity and long autoradiographic exposures (*5*). Primer extension has been used to simplify genomic sequencing, but these procedures still require the special preparation of primers labeled to extremely high specific activity and up to 50 μg of DNA per sequencing lane (*6*).

We now describe a genomic sequencing method in which we use a ligation mediated polymerase chain reaction (PCR) procedure [see figure 1 in (*7*)]. Briefly, step 1 of our genomic sequencing procedure is base-specific chemical cleavage of DNA samples (*8*) at either G, G+A, T+C, or C (*9*), generating 5′ phosphorylated molecules. Step 2 is gene-specific primer extension of an oligonucleotide (primer 1) by a DNA polymerase to give molecules that have a blunt end on the side opposite the primer (*10*). Step 3 is the ligation of an unphosphorylated linker to the blunt ends (*11*). Step 4 is the exponential amplification of the linker-ligated fragments with the use of the longer oligonucleotide of the linker (as a linker-primer) and a second gene-specific primer (primer 2) in a PCR reaction (*12*). After undergoing 15 to 18 amplification cycles, the DNA fragments are separated on a sequencing gel, transferred by electroblotting to nylon membranes (*13*), and hybridized with a single-stranded gene-specific probe (*14*). This procedure works well for all bases, sensitivity is improved, and the background is minimized by the transfer and hybridization steps. Moreover, several different sequences can be analyzed in a single experiment by rehybridization of the membrane.

The human X-linked phosphoglycerate kinase (PGK-1) gene is a housekeeping gene that is subject to X inactivation. The 5′ region is a CpG-rich island (*15*), but, unlike most autosomal CpG islands that are characteristically unmethylated, the Hpa II sites in the region shown (Fig. 1) are methylated on the inactive X chromosome (*16, 17*).

In an experiment with HeLa cell DNA, two different primer sets (Fig. 1, D and E) were included simultaneously in the primer extension and amplification reactions. The sequence defined by primer set D was visualized first (Fig. 2A) by hybridization with an Eco RI–Dde I hybridization probe. After stripping of the first probe from the membrane and rehybridization with an Xma III–

G. P. Pfeifer, S. D. Steigerwald, A. D. Riggs, Molecular Biology Section, Beckman Research Institute of the City of Hope, Duarte, CA 91010.
P. R. Mueller and B. Wold, California Institute of Technology, Division of Biology, Pasadena, CA 91125.

*To whom correspondence should be addressed.

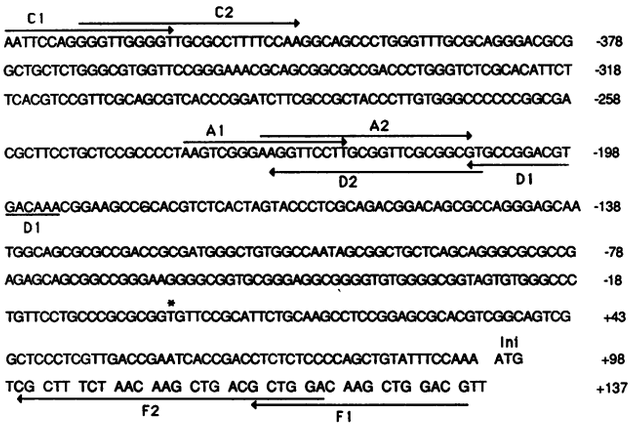

Fig. 1. Sequence of the CpG island containing the promoter and first exon of the human PGK-1 gene. Arrows indicate the oligonucleotide primer positions and the direction of reading. The asterisk indicates the major transcription start site. "Ini" marks the translation start point.

Pvu II probe, the sequence defined by primer set F became visible (Fig. 2B). As shown, both sequences could be read unambiguously; the correct calls would be made for a total of 336 nucleotides, even if the sequence were not known.

5-Methylcytosine appears as a gap in the cytosine ladder of a Maxam-Gilbert sequencing gel. To test the procedure as an assay for methylated cytosines, we studied human-hamster hybrid cell lines that contained either an inactive or an active human X chromosome (*17*). Primer set A (Fig. 1) allowed reading of the sequence toward the transcription start site, and the data indicate that the human inactive X chromosome is methylated at all 17 CpG dinucleotides in the investigated region (Fig. 3, lane X_i), whereas the active X_i chromosome is unmethylated at the same 17 sites (Fig. 3, lane X_a). By comparison with cloned DNA, we found that HeLa DNA has no detectable methylation at 52 CpG dinucleotides analyzed in the PGK-1–associated CpG-rich island, a result consistent with this cell line not having an inactive X chromosome.

Although the band patterns shown in Figs. 2 and 3 are reproducible, band intensities vary from fragment to fragment within a single lane; a few bands are weak or even missing on short exposures. For example, three bands in a region of about 100 nucleotides are not seen in the sequence shown in Fig. 3. Band intensity may vary because (i) the chemical reactivity of individual bases may differ; (ii) small effects of sequence on amplification efficiency may play some role, although PCR can be very quantitative and reproducible [±4% SD through 25 cycles (*18*)]; (iii) certain sequences may not be extended to a blunt end as efficiently in primer extension prior to ligation; (iv) the ligation step may introduce some bias. Overexposed autoradiograms can often be used to obtain sequence information even from weak bands. This is the case for most of the sites that we have studied, such as the CpG site at nucleotide position −119. The intensities of only the weak bands were increased by increasing the time for ligation or the temperature of the first primer extension reaction (*19*). Thus, the steps before the PCR may be very efficient for strong bands.

Exponential amplification works so well that band detection probably no longer limits its sensitivity. Comparing band intensities with the standard genomic sequencing method applied to the same PGK-1 sequence studied with primer set D, we estimate an amplification of 3000- to 6000-fold after 15 PCR cycles (10 to 20% overall efficiency). One microgram of HeLa cell DNA per reaction gave a strong signal with short exposures (2 hours, Fig. 2B). However, a statistical limit should be considered. The lowest limit of the DNA needed for sequence analysis is a single molecule for each band in the sequence ladder. For optimum efficiency of molecule usage, there should be one chemical cleavage per length of sequence to be analyzed, for example one cut per 200 nucleotides (nt). To avoid band intensity variation of more than 10% standard deviation that is due only to statistical sampling fluctuation, about 100 founder molecules are needed for each band. If ligation and first primer extension for an unfavorable fragment is only about 10% efficient, then 2×10^5 molecules ($200 \times 100 \times 10$) should give a band of invariant intensity. This then would be a statistical limit for experiments where visualization of this band is important. Efforts at further increasing the sensitivity of the method should focus on weak bands and on the steps that precede PCR, such as ligation and first primer extension. Increasing ligation time was helpful; however, attempts to increase

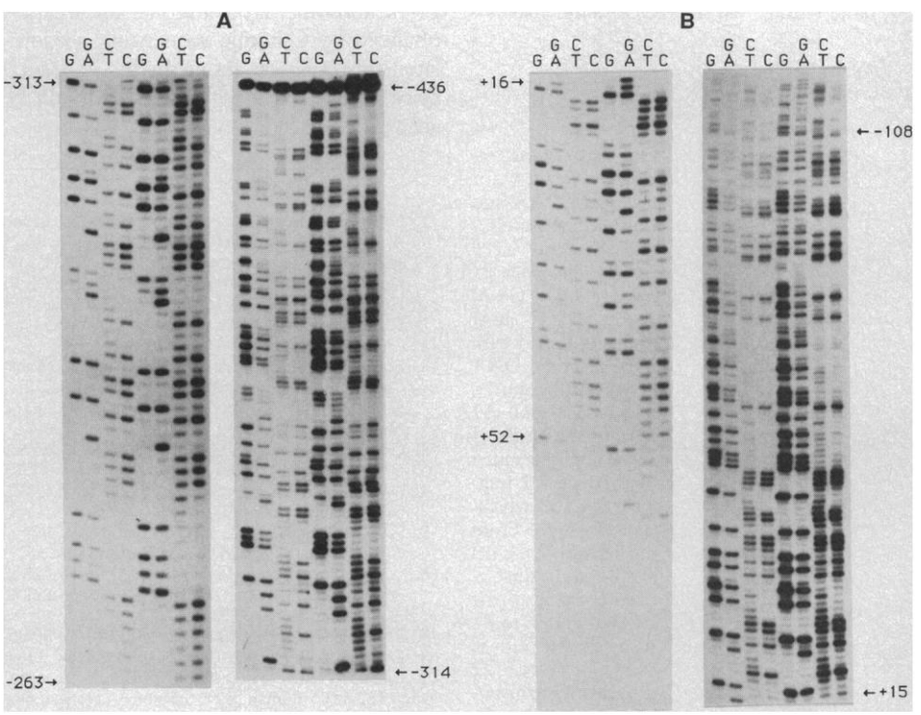

Fig. 2. Genomic sequence data for the 5′ end of PGK-1 in HeLa DNA. Primer sets D and F were used and two sequences were simultaneously amplified from 1 μg of DNA per base-specific reaction. The amplified DNA was then split for gel analysis with the four rightmost lanes of each panel receiving 2.5 times more DNA than the first four left lanes in order to visualize the weakest bands. (**A**) A single-stranded Eco RI–Dde I fragment was used as a hybridization probe to visualize the sequence defined by primer set D. (**B**) Rehybridization of the nylon membrane with a single-stranded Xma III–Pvu II fragment allowed visualization of the sequence defined by primer set F. The strong signal in all lanes at position −436 of A corresponds to the Eco RI site at which the DNA had been cut. The sequence corresponds to the sequence in (*15*), except at position −430, where there is an extra A.

ligation efficiency by adding polyethylene glycol (*20*) or hexamine cobalt (*21*) to the ligation mixtures were unsuccessful. The optimum temperature and Mg^{2+} concentrations for PCR depend on the primer and the gene region to be studied. Optimization of conditions was more critical for certain primers than for others. We obtained good sequence ladders for each of the four primer sets in Fig. 1, even though the region is more than 70% G+C.

Instead of directly labeling the fragments of the sequence ladder (*7*), we transferred to nylon membranes and, to visualize the sequence, hybridized with a single-stranded probe located 3′ to the gene-specific primers. Transfer to membranes makes it possible to use relatively low levels of radioactivity, results in minimal radiation exposure to laboratory workers, and permits analysis of multiple sequences by rehybridization of the membrane (Fig. 2, A and B). We do not yet know what the limit is for simultaneous ladder formation and sequential visualization, but a membrane containing a sequence ladder can be rehybridized up to 50 times (*22*).

By improving the quality of the data and by decreasing the amount of DNA needed, genomic sequencing by ligation mediated PCR should aid in vivo footprinting studies (*7*) and studies on DNA cytosine methylation. The procedure also appears adequate for sequencing unknown DNA regions adjacent to any known region by primer extension into the unknown region. Step by step genomic sequencing, in which the primers are selected from the previous sequence determination, would avoid any cloning procedure.

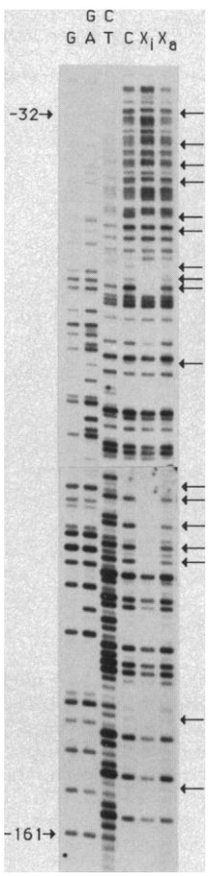

Fig. 3. Genomic sequence data for HeLa cell DNA and Chinese hamster–human hybrid cell DNA. The lanes labeled G, G+A, T+C, and C show results obtained with 2 μg of HeLa DNA per lane. Oligonucleotides A1 and A2 were used as primers and a single-stranded Mae III–Apa I fragment was the hybridization probe. From the bottom of the gel to the top, the sequence reads the lower strand 3′ to 5′. Also shown are C lanes for DNA (2 μg) from hamster-human hybrid cells containing either an active (X_a) or an inactive (X_i) human X chromosome. Methylated cytosines, which appear as a gap in the sequence ladder, are indicated by arrows.

Most methylation studies to date have used DNA 5-methylcytosine–sensitive restriction endonucleases such as Hpa II. Methylation analysis with PCR after Hpa II cleavage is more sensitive than genomic sequencing, needing only a few molecules (*18*), but the Hpa II–PCR procedure assays only about 6% of all methylated sites. Genomic sequencing allows analysis of all sites, and we found that the inactive human X chromosome in Chinese hamster–human hybrid cells faithfully retains a high level methylation at 17 CpG sites in the PGK-1 upstream region, whereas an active X chromosome is unmethylated at these sites. It has been suggested that more than 50% of methylated cytosines are not at CpG sites (*23*), and in vitro DNA methyltransferase will occasionally methylate cytosine at other than CpG sites (*24*). It is therefore of interest that we have not seen methylated cytosines in dinucleotide sequences other than CpG, even in the heavily methylated inactive X chromosome. As additional information obtained by genomic sequencing accumulates, this question about the in vivo specificity of DNA methyltransferase should be answered.

REFERENCES AND NOTES

1. For reviews on DNA methylation, see A. Razin and A. D. Riggs, *Science* **210**, 604 (1980); A. D. Riggs and P. A. Jones, *Adv. Cancer Res.* **40**, 1 (1983); W. Doerfler, *Annu. Rev. Biochem.* **52**, 93 (1983); R. Holliday, *Science* **238**, 163 (1987); H. Cedar, *Cell* **53**, 3 (1988); S. Grünwald and G. P. Pfeifer, *Prog. Clin. Biochem. Med.* **9**, 61 (1989).
2. For reviews on genomic imprinting, see M. Monk, *Genes Dev.* **2**, 921 (1988); D. Solter, *Annu. Rev. Genet.* **22**, 127 (1988); W. Reik and M. A. Surani, *Nature* **338**, 112 (1989).
3. G. M. Church and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1991 (1984).
4. A. Ephrussi, G. M. Church, S. Tonegawa, W. Gilbert, *Science* **227**, 134 (1985).
5. H. P. Saluz and J. P. Jost, *A Laboratory Guide to Genomic Sequencing* (Birkhauser, Boston, 1987).
6. ———, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2602 (1989); M. M. Becker, Z. Wang, G. Grossmann, K. A. Becherer, *ibid.*, p. 5315.
7. P. R. Mueller and B. Wold, *Science*, **780** (1989).
8. DNA from human-hamster hybrid cell lines that contained either an inactive (X8-6T2) or an active (Y162-11C) human X chromosome (*17*) was provided by R. S. Hansen and S. M. Gartler. Nuclei were isolated from HeLa (S315) cells by published procedures [J. Wijnholds, J. N. J. Philipsen, G. Ab, *EMBO J.* 7, 2757 (1988)], and DNA was purified from the isolated nuclei as described (*5*). DNA's were usually cleaved with Eco RI to reduce viscosity.
9. Base-specific cleavage: DNA, usually a 10-μg sample, was subjected to the Maxam-Gilbert reactions [A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980)] as modified by Saluz and Jost (*5*). After piperidine cleavage, DNA was precipitated with 2.5 volumes of ethanol, washed twice with 80% ethanol, and dissolved in 200 μl of water. Traces of piperidine were removed by vacuum drying the sample overnight. The DNA pellet was dissolved in water (1 μg/μl).
10. First primer extension: 1 to 2 μg of chemically cleaved DNA was mixed (final volume, 15 μl) with 0.6 pmol of a gene-specific primer (A1, D1, or F1, in Fig. 1) in 40 m*M* tris-HCl (*p*H 7.7) and 50 m*M* NaCl, heated at 95°C for 3 min, and then at 45°C for 30 min. The solution was adjusted to 25 m*M* tris-HCl (*p*H 7.7), 30 m*M* NaCl, 6 m*M* $MgCl_2$, 6 m*M* dithiothreitol, 80 μ*M* dNTP's, and a final volume of 24 μl. Sequenase™ (4.5 units, USB) was added, and the reaction mixture was incubated at 45°C for 15 min. The reaction was stopped by adding 6 μl of ice-cold 310 m*M* tris-HCl (*p*H 7.7), and the enzyme was inactivated at 67°C for 15 min. The ligation step (*11*) followed without further treatment.
11. Ligation: The structure of the linker and the annealing and ligation conditions were as described in (*7*).
12. PCR amplification: Primer extended, ligated molecules were amplified with Taq polymerase (Amplitaq™, Perkin-Elmer Cetus) in the presence of the longer oligonucleotide of the linker and a gene-specific primer (A2, D2, or F2; Fig. 1). The fragments were amplified in 100 μl of 10 m*M* tris-HCl (*p*H 8.9), 40 m*M* NaCl, 2 m*M* $MgCl_2$, gelatin (0.01%), 0.25 m*M* dNTP's, 10 pmol of each primer, with 3 units Taq polymerase, according to the Perkin-Elmer Cetus protocol. Conditions for the 15 to 18 thermal cycles were 1 min at 95°C, 2 min at 65°C, and 3 min at 75°C. Fresh enzyme (1 unit) was added and incubation was continued for 10 min at 74°C; EDTA and sodium acetate (*p*H 5.2) were added to 20 m*M* and 300 m*M*, respectively, and the reaction mixture was extracted with phenol and $CHCl_3$, and subsequently precipitated with ethanol and centrifuged. The DNA was then dissolved in electrophoresis sample dye (*5*), and, just before being placed on the gel, it was heated to 95°C for 2 min.
13. Gel electrophoresis and electroblotting: DNA fragments were separated on a sequencing gel (95 by 0.08 or 0.04 cm) consisting of 8% polyacrylamide (acrylamide-bisacrylamide, 29:1) and 7*M* urea; and then transferred to GeneScreen nylon membrane by electroblotting (*5*). A simple transfer apparatus was constructed with porous steel plates of a Bio-Rad gel dryer as electrodes (Bio-Rad 200/2.0 power supply). The details for this and other procedures are available on request. After the transfer, the membranes were air-dried, baked at 80°C for 20 min in a vacuum oven, then subjected to ultraviolet crosslinking [P. B. Becker, S. Ruppert, G. Schütz, *Cell* **51**, 435 (1987)].
14. Hybridization: To make probes the cDNA procedure [F. Weih, A. F. Stewart, G. Schütz, *Nucleic Acids Res.* **16**, 1628 (1988)] was used. Small (100 to 200 bp) restriction fragments from the 5′ region of PGK-1 (see below) were subcloned into Bluescript SK+ (Stratagene), and then RNA was produced from the inserts on a preparative scale; either T3 or T7 RNA polymerase was used. Labeled cDNA was made from the RNA (0.5 μg) by reverse transcriptase with the use of T7 or T3 primers and 100 μCi of [^{32}P]dCTP (6000 Ci/mmol). After ethanol precipitation, the single-stranded cDNA probe (1×10^9 to 5×10^9 dpm/μg) was used directly for hybridization. Hybridization probes were made from three subclones that contained fragments Eco RI–Dde I (197 bp; positions −436 to −240); Mae III–Apa I (181 bp; −199 to −19); or Xma III–Pvu II (150 bp; −69 to +81). Hybridization and washing buffers were as described (*5*). Preliminary hybridization (25 ml of hybridization buffer, 10 min) and hybridization (50 μCi per gel in 5 ml of hybridization buffer, overnight) were performed in rolling cylinders at 68°C in a hybridization oven. Nylon filters were exposed for 2 to 7 hours to Kodak XAR-5 film with intensifying screens at −70°C. For rehybridization experiments, probes were stripped from the membrane by incubation in 0.2*M* NaOH at 45°C for 30 min.
15. J. Singer-Sam *et al.*, *Gene* **32**, 409 (1984).
16. D. H. Keith, J. Singer-Sam, A. D. Riggs, *Mol. Cell. Biol.* **6**, 4122 (1986).

17. R. S. Hansen, N. A. Ellis, S. M. Gartler, *ibid.* **8**, 4692 (1988).
18. J. Singer-Sam *et al.*, in *Nucleic Acid Methylation*, G. Clawson, D. Willis, A. Weissbach, P. A. Jones, Eds. (Liss, New York, in press); this is volume 128 of *UCLA Symposia on Molecular and Cellular Biology, New Series*.
19. G. P. Pfeifer, S. D. Steigerwald, P. R. Mueller, B. Wold, A. D. Riggs, unpublished results.
20. B. H. Pheiffer and S. B. Zimmerman, *Nucleic Acids Res.* **11**, 7853 (1983).
21. J. R. Rusche and P. Howard-Flanders, *ibid.* **13**, 1997 (1985).
22. G. M. Church and S. Kieffer-Higgens, *Science* **240**, 185 (1988).
23. D. M. Woodcock, P. J. Crowther, W. P. Diver, *Biochem. Biophys. Res. Commun.* **145**, 888 (1987).
24. D. Simon *et al.*, *Nature* **304**, 275 (1983); G. P. Pfeifer, E. Spiess, S. Grünwald, T. L. J. Boehm, D. Drahovsky, *EMBO J.* **4**, 2879 (1985).
25. Supported by National Institute of Aging grant AG08196 (A.D.R.), NIH grants RR07003 and GM355262BW (B.W.), and a fellowship from the Deutsche Forschungsgemeinschaft (Pf212/1-1) (G.P.P.).

11 July 1989; accepted 26 September 1989

Isolation of Human Transcribed Sequences from Human-Rodent Somatic Cell Hybrids

PU LIU, RANDY LEGERSKI, MICHAEL J. SICILIANO*

A method was developed for selectively isolating genes from localized regions of the human genome that are contained in interspecific hybrid cells. Complementary human DNA was prepared from a human-rodent somatic cell hybrid that contained less than 1% human DNA, by using consensus 5′ intron splice sequences as primers. These primers would select immature, unspliced messenger RNA (still retaining species-specific repeat sequences) as templates. Screening a derived complementary DNA library for human repeat sequences resulted in the isolation of human clones at the anticipated frequency with characteristics expected of exons of transcribed human genes—single copy sequences that hybridized to discrete bands on Northern (RNA) blots.

HUMAN-RODENT SOMATIC CELL hybrids are often highly reduced for human genomic content, with the human fraction representing <1% of the total DNA. Regions of the human genome may be isolated by screening DNA libraries made from such hybrids for human-specific repetitive sequences. Whereas this has been an important step in reverse genetic approaches to the cloning of biomedically important genes, subsequent identification of the genes themselves has proven problematic. Preparation of cDNA libraries from mature mRNA produced by such cells, though direct, is not helpful because the repeat sequences that are needed to detect the low percentage of clones containing human sequences are primarily located in introns (which are removed during RNA message maturation) (*1*).

We have developed a strategy using immature transcripts (hnRNA) as templates for cDNA synthesis. Many such cDNAs would contain exon coding sequences as well as introns with human repeat sequences. To enrich for cDNA transcripts of hnRNA in a total RNA population, we began cDNA synthesis with consensus 5′ splice sequences as primer. Such a primer will initiate cDNA synthesis at the 5′ end of an intron (present in hnRNA but not mature mRNA). Complementary DNA synthesis will proceed copying the adjacent 5′ exon and then into the next 5′ intron, which may contain a repeat sequence. Cloning of this cDNA population results in a library in which human repetitive sequences are present in clones representing structural gene transcripts.

We used a human–Chinese hamster ovary (CHO) cell hybrid (20XP3542-1-4) containing the region of human chromosome 19q that carries markers closely linked to the myotonic dystrophy gene (*DM*) (*2*). The human DNA content in the hybrid was determined to be approximately 25 Mb (<1% of total genomic) by slot blot analysis (*3*). Chromosomal mapping of human genomic DNA clones obtained from cosmid libraries made with the hybrid DNA indicates approximately one-half of the human DNA to be from the *DM* region of human chromosome 19q, with much of the remainder DNA from a portion of chromosome 17 (*4*). A cDNA library was constructed from this line by established procedures (*5*) modified according to the strategy outlined above. To selectively use hnRNA as template, as opposed to mature mRNA, and to enrich for coding sequences, we used hexamers complementary to the 5′ splice site as primers. Because of sequence divergence of the 5′ splice site, four hexamers were used that represent the majority of the 5′ splice

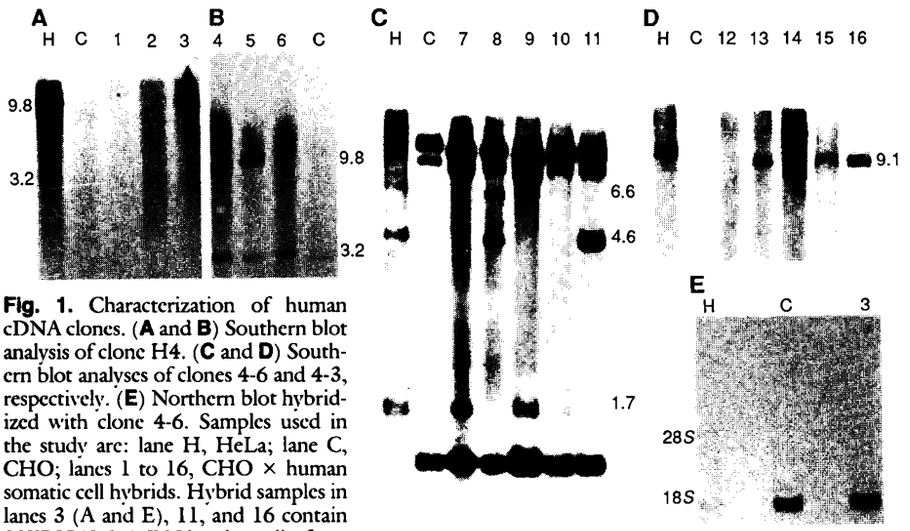

Fig. 1. Characterization of human cDNA clones. (**A** and **B**) Southern blot analysis of clone H4. (**C** and **D**) Southern blot analyses of clones 4-6 and 4-3, respectively. (**E**) Northern blot hybridized with clone 4-6. Samples used in the study are: lane H, HeLa; lane C, CHO; lanes 1 to 16, CHO × human somatic cell hybrids. Hybrid samples in lanes 3 (A and E), 11, and 16 contain 20XP3542-1-4 DNA—the cells from which the cDNA library was made. Putative human cDNA-containing plasmid clones were identified by screening the library with total human DNA. The plasmid DNA was purified, digested with Pst I, and separated by agarose gel electrophoresis. Insert bands were excised and DNA was isolated from agarose by the squeeze-freeze method (*12*). Insert DNA was labeled with a random prime labeling kit (Amersham). For both Southern and Northern blots, probe DNAs were hybridized with a molar excess of sheared human placental DNA to block hybridization by the repetitive region of the probe (*13*). The Southern blot hybridization was performed according to Amasino (*14*). For the blot shown in (B), the last two washes were done at 60°C, instead of the usual 65°C, to enhance the cross hybridization to CHO DNA. Gel separation of poly(A)$^+$ RNA, Northern blot transfer, and filter hybridization were done by standard procedures (*15*).

Department of Molecular Genetics, The University of Texas, M. D. Anderson Cancer Center, Houston, TX 77030.

*To whom correspondence should be addressed.