Gene Expression, Vol. 4, pp. 163–175, 1995
Printed in the USA.

1052-2166/95 $3.00 + .00

The Genomic Structure of the Human UBE1L Gene

KLAAS KOK,*[1] ANKE VD BERG,* PATRICK M. J. F. VELDHUIS,* MARION FRANKE,* PETER TERPSTRA,† AND CHARLES H. C. M. BUYS*

**Department of Medical Genetics and †Biomedical Technology Centre, University of Groningen, Groningen, The Netherlands*

The human UBE1L gene, for which the product may well play a role in the ubiquitin system because of its high degree of identity to the ubiquitin activating enzyme, is located at 3p21, a chromosomal region consistently showing loss of heterozygosity in lung cancer. The finding that UBE1L is well expressed in normal lung tissue, but hardly or not in lung cancer-derived cell lines, prompted us to investigate its genomic structure to find an explanation for the lack of expression in lung cancer. The gene has 22 exons distributed over 8.4 kb. Both anchored PCR experiments and mapping of DNase I-hypersensitive sites point to the region immediately upstream of exon 1 as the promoter site. Three moderately to well-informative polymorphisms were found, of which one is easily directly detectable. Cancer-specific mutations were not detected. The lack of expression in lung cancer cell lines correlated with a highly decreased sensitivity towards DNAse I of the promoter region and with an almost complete methylation of the HhaI site in the first exon. 5′-Azacytidine-induced demethylation did not result in a marked increase of the UBE1L mRNA level in the tumor cell lines. This leaves the possibility that mutation or absence of yet unknown transcription factors causes a regulatory block of the UBE1L gene.

Lung cancer Human chromosome 3 Gene structure Gene expression UBE1L

A number of reports have described the frequent loss of heterozygosity in lung cancer at band p21 of the short arm of chromosome 3, suggesting the presence of one or more tumor suppressor genes (Kok et al., 1987; Naylor et al., 1987, Brauch et al., 1987). Recent reports of homozygous deletions delimit the region of interest to that between the loci APEH and D3S1235 (Daly et al., 1993; Kok et al., 1994). The finding that DNA from this very region can suppress tumorigenicity when transfected into tumorigenic cells supports the idea that tumor suppressor genes may be present in the region (Killary et al., 1992). Several genes assigned to the short arm of chromosome 3 have been suggested as candidate genes mainly because of their location [APEH, acyl-peptide hydrolase (Naylor et al., 1989; Erlandson et al., 1990)], or also based on varying expression in tumor cell lines or tumors [APEH (Erlandson et al., 1990); ACY, aminoacylase-1 (Miller et al., 1989; Cook et al., 1993)], or on the nature of the gene product [PTPG, protein-tyrosine phosphatase gamma (LaForgia et al., 1991)].

We have isolated a gene located 150 kb centromeric to D3F15S2 that is among those loci most frequently showing loss of heterozygosity in lung cancer (Carritt et al., 1992). The amino acid sequence as predicted from the sequence of a 3.3 kb cDNA clone had a 47% identity with the human ubiquitin activating enzyme E1 (Kok et al., 1993a), encoded by the UBE1 gene on the X chromosome. Therefore, we named the gene UBE1L (L for like). The function of its protein product, however, is still unknown. The gene is ubiquitously expressed with a messenger RNA of 3.5 kb. It is also well expressed in the lung. In lung cancer-derived cell lines we found that one allele had always been eliminated. Despite the presence of the

Received July 7, 1994; revision accepted September 9, 1994.
[1]Address correspondence to Dr. K. Kok, Department of Medical Genetics, University of Groningen, A. Deusinglaan 4, 9713 AW Groningen, The Netherlands.

remaining allele, gene expression in these cell lines was virtually absent. It could not be detected by Northern blot analysis (Carritt et al., 1992). However, by reverse transcription followed by PCR we could demonstrate that transcripts were present, although at a very low level (Kok et al., 1993b).

To find possible mutations explaining the dramatic reduction of expression in the lung cancer-derived cell lines, we wanted to screen for mutations all exons of the gene as well as its promoter region. Therefore, we needed to know more about the UBE1L gene structure. This article presents the results of our analysis of the genomic structure of UBE1L and its upstream region as well as those of the mutation analysis of our lung cancer-derived cell lines.

MATERIALS AND METHODS

Sequence Analysis

Isolation of cosmid and plasmid DNA, restriction digestion, and cloning of various restriction fragments were carried out according to standard laboratory procedures (Sambrook et al., 1989). The dideoxynucleotide chain-termination method (Sanger et al., 1977) was applied to double-stranded templates and PCR products, using a T7 polymerase-based sequencing kit (Pharmacia). Prior to sequencing, PCR products were purified using the Sephaglass BandPrep Kit (Pharmacia). Primers were homemade by the phosphite triester method using a commercial oligonucleotide synthesizer (Gene Assembler plus, Pharmacia). The complete UBE1L genomic sequence has been submitted to GenBank (accession No. L34170).

Anchored PCR

Total RNA was isolated from an EBV-transformed lymphoblastoid cell line as described (Kok et al., 1993b), and subjected to a 5′-anchored PCR (Frohman et al., 1988), using the 5′RACE system (Gibco BRL), and the procedure recommended by the manufacturers. The primers used were the UBE1L specific primers P1 (5′-AGCTCCTCATCCAGTAGCTTCG-3′) and P2 (5′-GACACAAGGTGCCCACCTTC-3′), whereas first strand cDNA synthesis was started from primer P3 (5′-GTCTCTTGACTTCAGTGATAGCC-3′).

Analysis of DNAse I-Hypersensitive Sites

Cultured cells were harvested and resuspended in cell lysis buffer containing 0.35 M sucrose, 25 mM KCl, 5 mM $MgCl_2$, 1 mM EGTA, 0.15 mM spermidine, 0.1 mM PMSF, 50 mM Tris-HCl, pH 7.5, and 0.2% Triton-X100 (Kok et al., 1985). Nuclei were harvested by centrifugation at 700 × *g* for 10 min, and were washed twice with the cell lysis buffer. The concentration of nuclei was determined by making a 1 : 10 (v/v) dilution of a 20-μl aliquot with a vital dye (Türks solution, Merck) and counting the nuclei using a Büchner hemocytometer. Nuclei were washed once with DNAse I digestion buffer containing 0.35 M sucrose, 25 mM KCl, 5 mM $MgCl_2$, 0.1 mM $CaCl_2$, 10 mM Tris-HCl, pH 7.5, and resuspended in this buffer at a DNA concentration of 0.5 mg/ml. A series of 1-ml aliquots was mixed with increasing amounts of DNAse (Sigma), in a range from 0 to 50 units, and incubated at 37°C for 10 min. The reaction was terminated by the addition of 1 ml buffer containing 1% SDS and 25 mM EDTA. DNA purification, restriction digestion, and subsequent Southern analysis were carried out by standard procedures.

Single-Strand Conformation Polymorphism (SSCP) Analysis

Reverse transcription was started from 30 μg total RNA in a total volume of 60 μl as previously described (Kok et al., 1993b). One microliter was subsequently amplified for 30 cycles with the appropriate primers. PCR conditions were as described. The PCR products were separated on 1% NA agarose (Pharmacia)/2% NuSieve agarose (Research Organics) gels. The band corresponding to the calculated length of the PCR product was cut out, melted, and mixed with 1 vol of distilled water. One microliter of this mixture (or 100 ng genomic DNA, when DNA instead of RNA was used as the starting material) was subjected to a radioactive PCR, using a buffer in which 2 μM dCTP was replaced by 0.2 μM dCTP and 1 μl [α^{32}P]dCTP. Amplification was for 30 cycles. When the resulting product exceeded 250 bp, 10 units of the appropriate restriction enzyme were added directly to the PCR mixture, followed by an incubation at 37°C for 2 h. The fidelity of the radioactive PCR was checked by electrophoresis of one-fourth of the mixture on a 3% agarose gel. Another 5 μl of the reaction mixture was mixed with 2.5 μl 20 mM EDTA and 2.5 μl, 0.4% SDS, boiled for 5 min, chilled on ice, and applied to either of two nondenaturing acrylamide gels, consisting of 6% acrylamide/bisacrylamide (19 : 1, w/w), 1 × TBE, and either 5% glycerol or 10% glycerol (Orita et al., 1989). During electrophoresis (30 W, 6 h), gels containing 10% or 5% glyc-

erol were kept at a constant temperature of 30°C or 10°C, by the use of a thermostatic plate. For dideoxy fingerprinting (Sarkar et al., 1992), DNA segments were amplified using a primer system in which one of the primers was biotinylated, allowing purification of the PCR product by a biomagnetic separation system (DynaBeads, ITK Diagnostics). The PCR product was subjected to solid-phase DNA sequencing, using a T7 polymerase-based sequence kit (Pharmacia) and dideoxythimidine for chain termination. The resulting products were analyzed as described (Sarkar et al., 1992).

Methylation Analysis

DNA methylation was assessed by digestion of samples of 10 μg DNA with either MspI, HpaII, or HhaI, in combination with the nonmethylation-sensitive enzyme BamHI. Restriction digestion was carried out for 16 h at 37°C using 10 units of enzyme/μg DNA in the buffer conditions recommended by the supplier. Digested samples were electrophoresed on 0.7% agarose gels, and were transferred to nylon membranes as described previously. Southern analysis was carried out as described (Carritt et al., 1992). The small-cell lung cancer (SCLC) cell lines GLC4 and GLC8 were cultured in the presence of 5-azacytidine (5-azaC) to produce partially demethylated DNA (Dobkin et al., 1987). For this, the culture medium was supplemented just before use with 2 μM 5-azaC (Sigma). A 2 mM 5-azaC stock solution was stored in small aliquots at −80°C. Medium was renewed every 3 days. Cells were harvested after 3 weeks, and DNA and RNA were isolated as described. RNA was subjected to a quantitative analysis of the UBE1L mRNA level as previously described (Kok et al., 1993b).

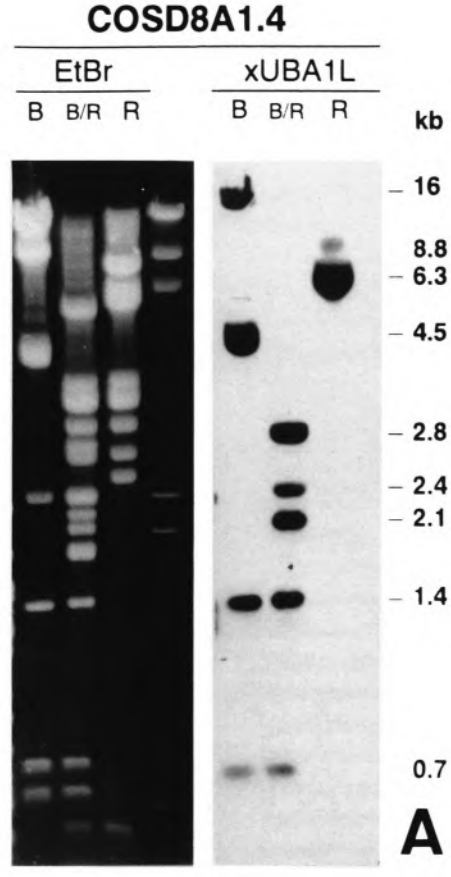
RESULTS

Identification of the Intron–Exon Boundaries of the UBE1L Gene

Initially, a cosmid coded D8A1.4 was isolated using as a probe a 1.6 kb UBE1L cDNA. Upon hybridization of EcoRI digests, BamHI digests, and EcoRI × BamHI double digests of the cosmid DNA with the 3.3 kb UBE1L cDNA, it appeared that the hybridizing fragments covered a total of 9.4 kb (Fig. 1A). The same EcoRI and BamHI fragments were detected upon Southern analysis of several human placenta and lymphocyte DNA samples, indicating that the DNA contained in the cosmid has the germline configuration. A restriction map of the cosmid, giving the restriction sites for EcoRI, BamHI, NruI, and MluI, is shown in Fig. 1B. The fragments that hybridize with the UBE1L cDNA cluster within the central part of the cosmid, indicating that it contains the entire cDNA. All fragments from the EcoRI × BamHI digest containing transcribed sequences were subcloned and sequenced, initially with vector-derived primers, subsequently with exon-derived primers. All introns were sequenced in both directions, with the exception of the largest one, which has only been sequenced in one direction. By comparing the genomic sequences with the cDNA sequence (GenBank accession No. L13852), a total of 21 introns were identified, 11 of them smaller than 100 nucleotides. The intron–exon structure of the gene is schematically depicted in Fig. 1B; the sequences of the intron–exon boundaries are presented in Table 1. All introns start with a GT dinucleotide and end with an AG dinucleotide, in agreement with the consensus sequence (Senepathy et al., 1990). The two genomic PstI fragments that contain the first exon and the polyadenylation signal, respectively, were identified by Southern analysis of cosmid DNA with a 5′ fragment and a 3′ fragment from the UBE1L cDNA, and sequenced almost completely. In the region of nucleotides 7–330, which is about 1 kb upstream from the first exon, an inverted repeat of 56 nucleotides was detected (Fig. 2A). The two elements of the inverted repeat, separated by 210 nucleotides, differed at only four positions.

Localization of the Transcription Initiation Site

The complete sequence of the first exon, as well as some 1150 nucleotides directly upstream, is shown in Fig. 2A. The underlined CCA triplet marks the beginning of the 3.3 kb UBE1L cDNA clone. The start site of transcription was determined by an anchored PCR (Frohman et al., 1988) on the total RNA isolated from an EBV-transformed lymphoblastoid cell line with a high UBE1L expression. First strand cDNA synthesis was started with primer P3 in exon 7, at 915 nucleotides downstream from the CCA triplet (see Materials and Methods). Following tailing, the cDNA strands were amplified with the anchored primer (P*an*) in combination with one of two different primers (P1 and P2, respectively) in separate experiments. P1 lies in the first exon; its position is indicated in Fig. 2A. P2 lies in the fourth exon 377 nucleotides downstream from P1 according to the cDNA sequence. The resulting PCR products were



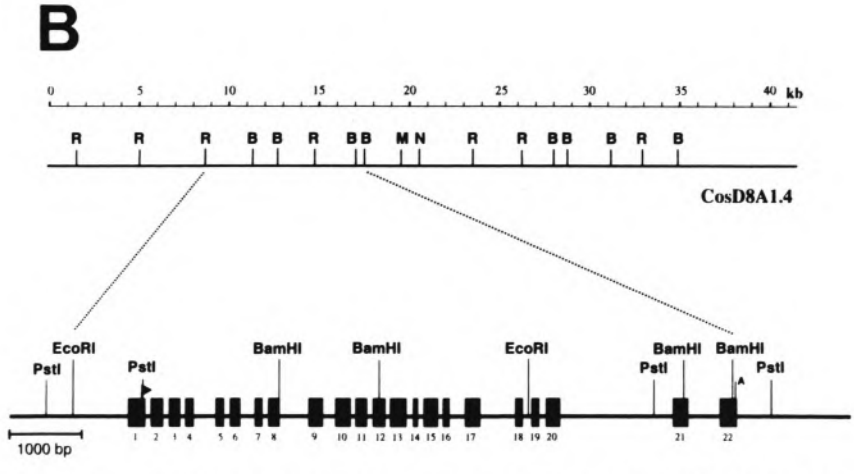


FIG. 1. Structure of the UBE1L gene. (A) Restriction digestion of cosmid COSD8A1.4 with BamHI (B), with EcoRI (R), or with both enzymes (B/R). The left part shows a photograph of the ethidium bromide-stained agarose gel, the right part shows the autoradiograph after hybridization with the 3.3 kb UBE1L cDNA. The length of the hybridizing fragments (in kilobases) is indicated. HindIII digested lambda DNA is used as a size marker (λ). (B) Restriction map of cosmid COSD8A1.4, showing the EcoRI (R), BamHI (B), MluI (M), and NruI (N) sites. Below the map, the coding region of the UBE1L gene is enlarged. Exons, numbered 1–22, are indicated by solid blocks. The arrowhead above exon 1 indicates the translation start site. The A above the last exon indicates the polyadenylation signal. The PstI sites defining the fragments containing the first and the last exon, respectively, are also indicated.

TABLE 1
INTRON-EXON BOUNDARIES OF THE UBE1L GENE

```
                                                    ttcacttttc ttttctccag AGGAGCCAGG        (1;  220)

CA AGA CAG CT gtgaggcccg aggtgggggg....(1;  84)....cttctgcttc ccatccacag G TAT GTG CTG    (2;  169)
S   R   Q   L                                                            Y   V   L

G GCT GCC CAG gtaagtgtcc tggggctatg....(2;  91)....tgcactggct ctctccctag TTT CTC CTC T    (3;  135)
  A   A   Q                                                              F   L   L

G GAC TTC CAG gtcagctcag gcctgcagcc....(3;  80)....tgatcactgt gtccacccag GTG GTG GTG C    (4;  107)
  D   F   Q                                                              V   V   V

CC CTC GTG GG gtgagtaaga ctgcctgccc....(4; 323)...cccctaccac cccatctcag G CAG TTG TTC    (5;   91)
A   L   V   G                                                            Q   L   F

C ATC TCC CAG gtgggtgctg ctgagctgta....(5; 104)...ccaagcctcc tcctacccag GGC TCC CCT G    (6;  136)
  I   S   Q                                                              G   S   P

CAC GTG CGG G gtaagccaat cccattccaa....(6; 201)...ctagactgga actccctcag AG GAT GGG TC    (7;   98)
 H   V   R                                                              E   D   G   S

T GTG AGA CAT gtgagtgcaa gtccatctga....(7;  92)....tgagccaggt gcccttgcag AAG TCC CTG G    (8;  144)
  V   R   H                                                              K   S   L

C TGG GAT CCT gtgagtagtc ctgttgctcc....(8; 414)...ccacagagtc ctaccaacag GTT GAT GCA G    (9;  172)
  W   D   P                                                              V   D   A

TGC CCA GGA A gtgctgaagg tgggcagagg....(9; 191)...gcctgaagtg ctgcccctag GC AAT CTC CA    (10;  188)
 C   P   G                                                              S   A   E   G

C TAC CTC CTG gtgagctgtg gggtgagact...(10;  79)....aaggctgttc ctgccaacag GTG GGC GCT G    (11;  156)
  Y   L   L                                                              V   G   A

G GAC GTT GGT gtgagtgctg acccctctcc...(11;  88)....tcttcctgcc ttcttcccag AGA CCC AAG G    (12;  165)
  D   V   G                                                              R   P   K

TTC CAG GCC C gtgagtgctt gacttcggag...(12;  79)....acccactgct cccytgccag GGC GCT ATG T    (13;  206)
 F   Q   A                                                              R   R   Y

C ACC CTG CAG gtaggaagca ccctggagac...(13;  98)....tcttcctccc tcctccacag TGG GCC CGG C    (14;   65)
  T   L   Q                                                              W   A   R

AC CAC CAA CA gtaaggccac caacagaggc...(14;  89)....acttcctcct ctctctgcag G GCA CAC ACT    (15;  184)
H   H   Q   Q                                                              A   H   T

A CCT AAT AAA gtgtgtggct aggggttggg...(15;  77)....gcttctactt acctacctag GTG CTT GAG G    (16;   75)
  P   N   K                                                              V   L   E

C ACC AAC CAA gtgagtggga ttctgtaggg...(16; 235)...tgctcttggt ctggctgcag GAC ACA CAC C    (17;  184)
  T   N   Q                                                              D   T   H

GCT GAG TTT G gtgaggctcc tggccctggc...(17; 507)...aacatcccct tccctgacag GC CCT GAG CA    (18;   83)
 A   E   F                                                              G   P   E   Q

G TTT GAG AAG gtgggtgccc aagtggcagt...(18; 139)...ctttgacttg ggccttacag GAT GAT GAC A    (19;   93)
  F   E   K                                                              D   D   D

C CGT GCC CAG gtaaccccac cccttgaggc...(19;  92)....aaatctcttg tccttggcag AGC AAG CGA A    (20;  192)
  R   A   Q                                                              S   K   R

C ATC CAG ACG gttgagccca tgatacccca...(20; 1570)..ctcctctacc tcatcccaag TTC CAT CAC C    (21;  194)
  I   Q   T                                                              F   H   H

TG CCC CTC AG gtgagcccac ttgggcttta...(21; 451)...cctaaaccca cccctaccag G GTG ACA GAA    (22;  231)
L   P   L   R                                                              V   T   E

TAAAGGAAGG cattgcagaga ggacggacg
```

Intron sequences are indicated by lower case letters and exon sequences by capitals. Numbers of the introns (or exons) and their total length (including the sequences shown) are indicated in parentheses. Translated sequences are presented as triplets with the respective amino acids indicated by their one-letter symbol.

analyzed on an agarose gel (Fig. 2B). In both experiments, several bands were produced, as can be seen from lanes 1 and 2 in Fig. 2B. The longest fragments in each lane indeed contained the 5′ part of the UBE1L cDNA, because they hybridized with a 900 bp EcoRI × PstI fragment containing the first 140 bp of the UBE1L cDNA. The PCR fragment resulting from the amplification with P1 is 260 bp. Subtraction of the length of the -artificial- anchored primer (48 bp) results in a cDNA fragment of 212 bp (see Fig. 2C). The downstream end point of this fragment is defined by the P1 primer (i.e., nucleotide 1351). Assuming absence of an additional intron, the upstream end point will thus map at nucleotide 1139 (double underlined triplet in Fig. 2A). The same calculation can be made for the 630 bp PCR product resulting from the amplification with P2. Now the upstream end point of the cDNA part maps at nucleotide 1146. The difference in length between the longest PCR products in each of the lanes is thus almost fully accounted for by the different position of the primer used for the amplification. We determined the nucleotide sequence of the longest PCR product from lane 1 using the P1 primer to initiate the reaction. The sequence

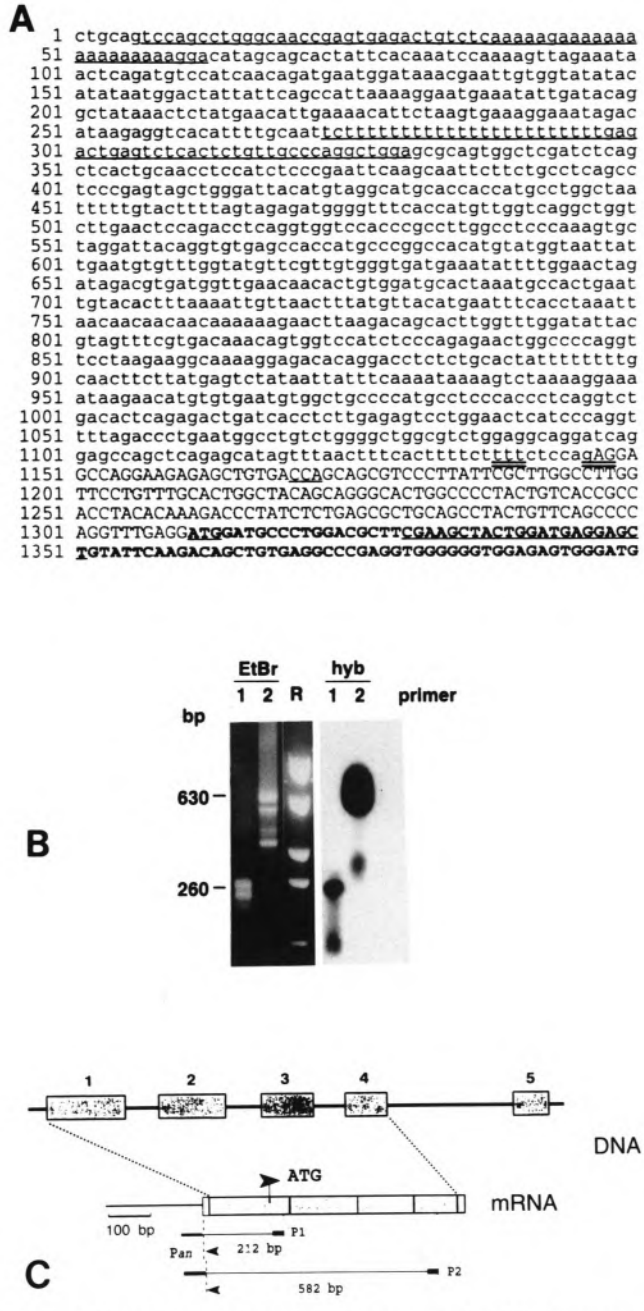


FIG. 2. Identification of the transcription start site. (A) Sequence of the PstI fragments that contain the first exon, and 1146 bp directly upstream from it. Underlined is the inverted repeat (positions 7–62 and 273–330); also underlined are the first triplet of the 3.3 kb UBE1L cDNA clone (CCA 1170–1172), the translation start site (ATG 1311–1313), and the primer P1. Double underlined are the end points of the anchored PCR products. Transcribed sequences, as revealed by the anchored PCR, are indicated by capitals. (B) Agarose gel electrophoresis of anchored PCR products after amplification with primers P1 (lane 1) and P2 (lane 2). The right part shows the autoradiograph after hybridization with a 900 bp genomic EcoRI × PstI fragment containing 140 bp of the first exon. Plasmid pUC18 digested with MboI is used as a size marker (R). (C) Schematic presentation of the anchored PCR analysis. Exon sequences are indicated by boxes. The position of the primers P1 and P2, as well as the length and the position of the PCR products, is indicated. The universal anchored primer is indicated by a solid bar and the letters Pan. The grey boxes indicate mRNA sequences present in UBE1L cDNA; the white box indicates transcribed sequences not present in UBE1L cDNA.

TABLE 2
SEQUENCE VARIANTS IN THE UBE1L GENE

```
deletion/insertion polymorphism at the 5'-border of exon 1:
      allele 1: 117 bp: tttcttttctccagAGGAGCCAGGAAGAGAGCT (53/92)
      allele 2: 108 bp: tttcttttct.........CCAGGAAGAGAGCT (39/92)

base substitution polymorphism in intron 6:
      allele 1:  aaggtgtggga  (23/38)
      allele 2:  -----t-----  (15/38)

base substitution polymorphism in intron 12:
      allele 1:     gtcccacttctgacccactgctcccctgccagGGCGCTAT  (31/46)
      allele 2:     ------c-----------------t--------------   (7/46)
      allele 3:     ------t-----------------t--------------   (8/46)
```

Dots and dashes indicate missing and identical nucleotides, respectively. Capitals indicate exon sequences. Allele frequencies are given in parentheses.

turned out to be identical to the genomic sequence, starting from position 1260 up to position 1147, after which point the sequence became unspecific. This strongly indicates that there exists no additional intron. Within the limits of this technique, the UBE1L transcription initiation site will therefore reside between nucleotides 1138 and 1147.

SSCP Analysis of the UBE1L Gene in SCLC-Derived Cell Lines

Southern analysis of EcoRI- or BamHI-digested DNA from 15 SCLC-derived cell lines and 40 lung tumor samples with the 3.3 kb UBE1L cDNA did not reveal hybridizing fragments of aberrant size in any of them (data not shown). Next, 15 SCLC-derived cell lines were screened by SSCP (Orita et al., 1989) for small mutations in the UBE1L coding sequence. For exons 15 to 22, SSCP analysis was carried out on total cellular RNA following reverse transcription and PCR (RT-PCR) with exon-specific primers. All products were analyzed under at least two different conditions. No aberrant bands were detected in any of the analyzed samples. Exons 1 to 14 were analyzed starting from genomic DNA instead of RNA, because RT-PCR gave relatively poor yields for these exons. In this region of the UBE1L gene, three variants were detected. A subsequent analysis of a substantial number of DNA samples from unrelated normal controls revealed that all three were neutral sequence polymorphisms. In each case, the nucleotide sequence of the alleles was determined by a direct analysis of the PCR product. One of the polymorphisms was a deletion/insertion of nine nucleotides in the region where we think the transcription starts. The deletion removes five nucleotides that are spacing a 4 bp direct repeat as well as one of the repeat units itself. Allele sequences and allele frequencies are given in Table 2. Although originally detected by SSCP analysis, this variation could readily be analyzed directly on standard agarose gels (Fig. 3). Another

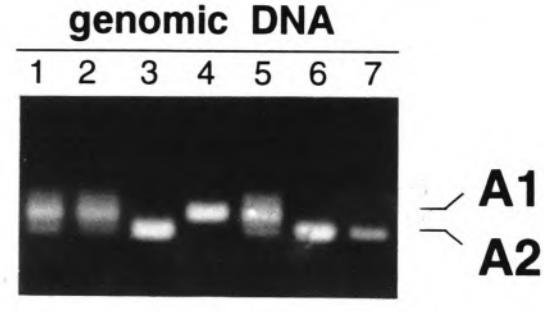


FIG. 3. Deletion/insertion polymorphism. DNA of seven unrelated individuals was amplified with primers PL (5′-GGAGCCAGCTCAGAGCATAG-3′) and PR (5′-CCA-GTGCAAACAGGAACCAAG-3′), and analyzed on a 1% NA agarose (Pharmacia)/2% NuSieve agarose (Research Organics) gel. Allele A1 and A2 are 117 bp and 108 bp, respectively. The blurred bands just above A1, present only in the heterozygous samples, are probably caused by heteroduplex structures. Allele frequencies are given in Table 2.

variant was a G ←→ T transition in the middle of intron 6 (Table 2). The third polymorphism is actually a double one consisting of two C ←→ T transitions, spaced by 18 unaffected nucleotides, in intron 12. The downstream transition mapped in the pyrimidine stretch of the splice acceptor site. Because the transition causes no interruption of this pyrimidine stretch, it is unlikely that splicing is affected. Although two transitions may cause four haplotypes, which we here denote as alleles, we detected only three of them (Table 2). All lung cancer cell lines analyzed were apparently homozygous for all three polymorphisms, in agreement with heterozygous loss of the gene. A DNA fragment consisting of the first exon and about 500 nucleotides directly upstream to it was analyzed not only by SSCP, but also by dideoxy fingerprinting (Sarkar et al., 1992). Apart from the above-mentioned polymorphism in the region of the transcription initiation site, no additional mutations were detected.

Identification of Sites of Transcriptional Regulation

No clearly recognizable promoter elements were found in the region immediately upstream of the initiation start site (Locker and Buzard, 1990). The identification of DNAse I hypersensitive sites is an alternative approach to localize a promoter, because these sites often coincide with regulatory sequences of genes (Gross and Garrard, 1988). The 5′ end of the UBE1L gene lies in a 15 kb BamHI fragment, within 2 kb from one of the BamHI sites. DNAse I-hypersensitive sites in the UBE1L upstream region were mapped relative to this BamHI site using as a probe a 1.15 kb SmaI-BamHI fragment (Fig. 4B). DNAse I treatment of nuclei isolated from the lymphoblastoid cell line exhibiting a high UBE1L expression resulted in two fragments in addition to the 15 kb BamHI fragment. They are represented by the prominent bands at 3.3 kb and 2.2 kb in Fig. 4A and mark two hypersensitive sites. The one corresponding to the 3.3 kb band maps 1.3 kb upstream from the putative initiation start site. The other hypersensitive site, corresponding to the band of 2.2 kb, maps between nucleotides 1075 and 1175 in Fig. 2, and may coincide with the putative promoter region. This band was flanked by two much weaker bands of about 2.0 and 2.4 kb, respectively. In the SCLC-derived cell line GLC4, the region directly upstream to the first exon turned out to be only weakly sensitive towards DNAse I.

Methylation Pattern of the UBE1L Upstream Region

We analyzed the methylation status of the UBE1L upstream region by hybridizing the 5′ EcoRI-BamHI fragment of 2.8 kb (Fig. 1B, also indicated in Fig. 5C) to genomic DNA, double-digested with either HhaI, HpaII, or MspI, each in combination with BamHI. HhaI cuts DNA at its target sequence GCGC only if the internal cytosine is unmethylated. MspI and HpaII both cut the target sequence CCGG. HpaII, but not MspI, digestion is inhibited by methylation of the internal cytosine. The methylation pattern of two SCLC-derived cell lines was compared with the methylation pattern of the lymphoblastoid cell line LB (Fig. 5A). Appearance of a fragment indicates that the site involved is to a large extent unmethylated. In the lane containing HhaI × BamHI digested DNA from the lymphoblastoid cell line (LB), only one hybridizing fragment, of 2.0 kb, can be seen. This indicates that the genomic DNA is completely unmethylated at the HhaI site, which, according to the sequence interpretation, lies in the first exon. This site is indicated in Fig. 5C, which also shows the position of the further HhaI and MspI/HpaII sites up to 4.5 kb upstream of the UBE1L coding region. The 2.0 kb fragment is absent in GLC8, and only faintly visible in GLC4 (Fig. 5A), indicating that the HhaI site in exon 1 is almost completely methylated in the two SCLC-derived cell lines. In both the LB and GLC lanes containing HpaII × BamHI digested DNA, two or more bands can be seen, in each case the shortest band being 2.7 kb. The presence of longer bands in the same lanes indicates that the HpaII site at 0.6 kb upstream from the first exon is only partly demethylated. The absence from the HpaII × BamHI digests of the 1.5 kb band, as present in the BamII × MspI digests, and the absence of shorter bands, indicates complete methylation in all cell lines of the MspI/HpaII sites in the UBE1L coding region itself (Fig. 5C). The same HpaII methylation pattern was seen in several SCLC-derived cell lines, and UBE1L-expressing tissues (data not shown).

To obtain an overall reduction of CpG methylation, GLC8 and GLC4 were cultured in the presence of 5′-azacytidine for 3 weeks, equal to at least two cell divisions. As can be seen in Fig. 5B, the methylation status of these cells had indeed changed. Relatively shorter bands appear upon hybridization of BamHI × HpaII digested genomic DNA with the 2.8 kb EcoRI-BamHI fragment. RNA was isolated from these cells, and sub-

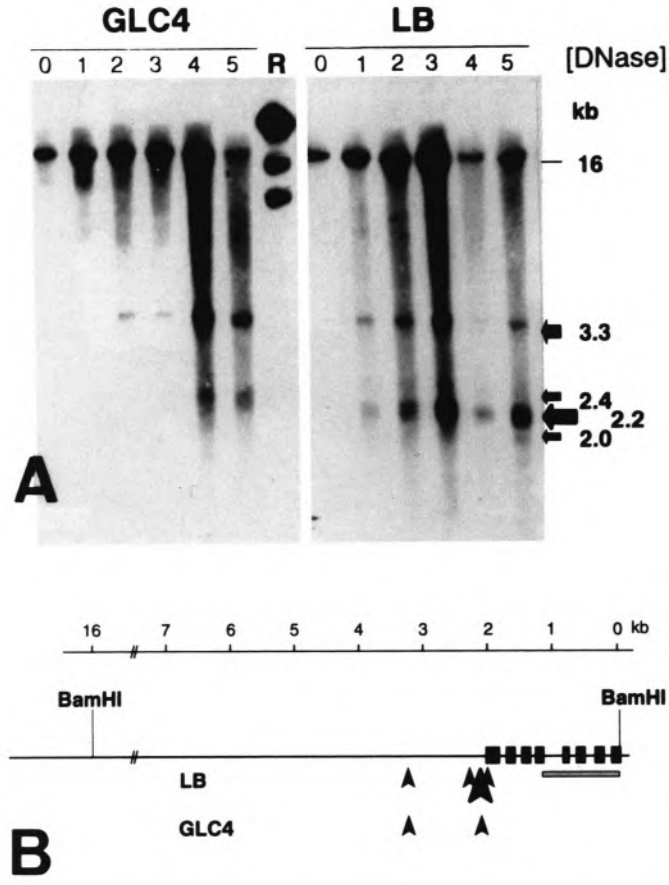


FIG. 4. Localization of DNAse I-hypersensitive sites. (A) Nuclei, isolated from the SCLC-derived cell line GLC4 and from a lymphoblastoid cell line (LB), were digested with DNAse I: lane 0–5; 0, 1, 2.5, 5, 10, and 25 unit/ml, respectively. Southern blots of BamHI-digested DNA were hybridized with a 1.15 kb SmaI-BamHI fragment. The length of the hybridizing fragments is indicated in kilobases. HindIII-digested lambda DNA is used as a reference (R). (B) Map of the hypersensitive sites (arrowheads) on the 16 kb BamHI fragment.

jected to a quantitative analysis of UBE1L gene expression (Kok et al., 1993b). The overall reduction of CpG methylation had no effect on the UBE1L mRNA level in GLC4, and increased the UBE1L mRNA level in GLC8 with no more than a factor 2 (from 10 to about 20 copies per cell).

DISCUSSION

The longest UBE1L cDNA clone we could isolate had a nucleotide sequence of 3.3 kb. It recognizes a mRNA of 3.5 kb (Carritt et al., 1992). However, the difference in length between cDNA and transcript may, to a large extent, be explained by the presence of a poly(A) tail in the latter. In total, a DNA segment of approx. 10 kb was sequenced, extending from 1.2 kb upstream of the first exon identified to 300 bp downstream of the polyadenylation site. The coding sequences lie on a genomic fragment of approximately 8380 nucleotides. The gene consists of 22 exons and 21 introns, 11 of which are less than 100 bp (i.e., relatively short when taking into account the distribution of intron lengths in vertebrates) (Hawkins, 1988). The extended splice donor and acceptor sites of the UBE1L gene shown in Table 1 are largely in agreement with the consensus sequences (Senapathy et al., 1990). We confirmed by anchored PCR that the 5′ end of the 3.3 kb UBE1L cDNA is indeed part of the UBE1L mRNA and extended the first exon with 23 nucleotides. It contains a TGA stopcodon (nucleotides 1167–1169) in frame with the open reading frame of the 3.3 kb UBE1L cDNA. Thus, translation has to start downstream of this triplet, probably at

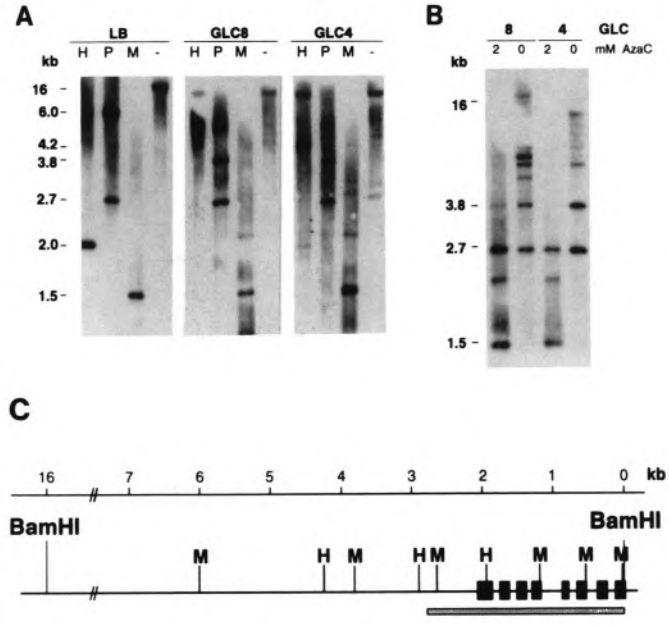


FIG. 5. Methylation of HhaI and HpaII sites. (A) DNA from a lymphoblastoid cell line (LB) and from two SCLC-derived cell lines (GLC4 and GLC8) was digested with BamHI, and in addition with HhaI (H), HpaII (P), or MspI (M). Southern blots were hybridized with the 2.8 kb EcoRI-BamII fragment. (B) DNA from GLC8 and GLC4 before (0) and after (2) culturing the cells for 14 days in the presence of 2 mM azacytidine was digested with BamHI and HpaII. The Southern blot was hybridized with the same probe as sub (A). (C) Restriction map of the 16 kb BamHI fragment. Exon sequences are indicated by grey boxes. The 2.8 kb EcoRI-BamHI fragment is indicated as a hatched bar. The position of the HhaI sites (H) and MspI/HpaII sites (M) is indicated. HhaI sites andMspI/HpaII sites upstream of the MspI site at 6.5 kb have not been mapped.

ATG 1311–1313 (numbers according to Fig. 2A). Translation initiation at this point would result in a protein with a predicted molecular weight of 112 kDa. In vivo transcription/translation of the 3.3 kb UBE1L cDNA in COS7 cells indeed resulted in a protein of this molecular weight (W. Helfrich, K. Kok, C. H. C. M. Buys, L. de Leij, manuscript in preparation).

Nucleotides 1131–1148 directly upstream of the start of the UBE1L cDNA clone revealed a striking homology with the splice acceptor consensus in consisting of an AG dinucleotide preceded by a 14 nucleotide perfect pyrimidine stretch. Moreover, the "A" nucleotide at position 1124 could well serve as a branch point for the splicing machinery (Senapathy et al., 1990). This, in combination with the absence of a clearly identifiable promoter sequence, initially suggested to us the presence of a possible upstream exon that was missing from the 3.3 kb cDNA clone. The anchored PCR analysis, however, demonstrated that this is not the case, and that transcription starts between nucleotides 1138 and 1147 (Fig. 2A). No exact position for the transcription initiation site could be identified. Many genes are known, however, for which transcription can start at several sites, either highly clustered or spread over a somewhat larger region (Sehgal et al., 1988; Means and Farnham, 1990).

Promoter regions of RNA-polymerase II-transcribed genes can be subdivided into three groups (Smale and Baltimore, 1989). The TATA box-containing promoters constitute one group. Genes with such a promoter have a well-defined single transcription start site, 27–31 bp downstream of the first "T" of the TATA box. In the vast majority of the genes, transcription starts at a CA dinucleotide (Corden et al., 1980). The second group are the so-called GC-rich promoters (reviewed by Sehgal et al., 1988). Genes with this type of promoter are characterized by the occurrence of one or multiple binding sites for transcription factors like SP1, AP1, AP2, and by the presence of several transcription initiation sites distributed over a 15–20 bp region (Sehgal et al., 1988). These genes mainly express themselves at

low levels and the proteins coded for are found in all cells. Some of these genes are characterized by the HIP1 binding site consensus sequence [ATTTCN(1-30)GCCA] at their site of transcription initiation (Means and Farnham, 1990). Transcription can start at multiple sites, but is dependent on the presence of binding sites for the above-mentioned transcription factors. A third group of promoters is associated with genes that have neither a TATA box element, nor a GC-rich region. Most of these genes are not constitutively active but rather are regulated during differentiation or development (reviewed by Smale and Baltimore, 1989). Computer analysis of the UBE1L upstream sequence revealed a weak homology with a TATA box 430 bp upstream of the first exon, suggesting transcription to start at 30 nucleotides downstream from that position. However, this would imply presence of a noncoding exon upstream of exon 1. When we tried to amplify the mRNA using a primer within the putative first exon and primer P1, after first strand cDNA synthesis starting from primer P2, we failed to obtain a product, ruling out the presence of a functional TATA box at the region where the homology occurred. On the whole, the region upstream of the first exon is not very GC rich, the GC content of the 1.3 kb PstI fragment (Fig. 1B) being 44%, although a few GC-rich (70%) stretches occur that are less than 50 bp long. The sequence at position 1136 to 1153 (Fig. 2) has some homology with the HIP1 (housekeeping initiation protein 1) binding site consensus, and indeed closely resembles the HPRT transcription initiation box (Means and Farnham, 1990). This type of promoter element requires the presence of SP1-like factors to regulate the efficiency of transcription. However, sequences homologous to the consensus binding sites of Sp1, Ap1, and Ap2 (Locker and Buzard, 1990) could not be detected. Still, this does not exclude the presence of a promoter, as some transcription factors have been described that bind to GC-rich regions lacking a SP1 consensus sequence (Kageyama et al., 1989) and as there is the third group of promoters occurring with genes regulated during differentiation or development.

DNAse I-hypersensitive sites often coincide with regulatory sequences of genes. Therefore, the mapping of such sites can also be of help to identify regions involved in the regulation of transcription (Kok et al., 1985; reviewed by Gross and Garrard, 1988). The results obtained for the UBE1L gene point to the region directly upstream of the first exon as a site for transcription regulation, thus supporting the results of the anchored PCR experiments and leading to the conclusion that this region may indeed contain the UBE1L promoter. Hypersensitive sites representing elements involved in transcriptional regulation may also occur in the first intron of some genes (Chrysogelos, 1993). Therefore, verification of the presence of UBE1L upstream sequences that control transcription and their identification awaits transient expression assays with constructs containing various parts of the region upstream to a suitable reporter gene.

The similarity of the amino acid sequence of the UBE1L product to that of the ubiquitin-activating enzyme E1 might suggest similar roles for E1 and the UBE1L gene product (Kok et al., 1993a). By activating and transporting a ubiquitin molecule to one of a group of ubiquitin-conjugating enzymes, E1 catalyses the first step of the ubiquitin pathway that ligates ubiquitin to intracellular target proteins. In human only one ubiquitin-activating enzyme has been identified. The gene encoding it has been cloned (Handley et al., 1991) and mapped to Xp11.2-p11.4 (Takahashi et al., 1992). There exist at least two isoforms of E1 in mammals (Cook and Chock, 1992). Moreover, using immunofluorescence techniques, E1 has been detected in various cellular compartments (Trausch et al., 1993). Neither study can exclude that proteins are involved encoded by different genes. Possibly, more than one gene exists that codes for E1-like products that each have a preferred set of target-conjugating enzymes. Occurrence of more than one activating enzyme would allow for regulation already at the ubiquitin activation step. The ubiquitin pathway has been implicated in a number of fundamental processes, including selective degradation of abnormal and short-living proteins, cell cycle progression, and DNA repair. A strongly decreased level of a ubiquitin-activating enzyme could conceivably contribute to the development of cancer, as has been suggested with respect to a role of the ubiquitin system in degrading nuclear oncoproteins like N-myc, c-myc, and c-fos (Ciechanover et al., 1991).

Most likely, the low level of UBE1L mRNA in lung cancer-derived cell lines (Kok et al., 1993b) cannot be attributed to mutant instable mRNA molecules, because in 15 SCLC cell lines analyzed no mutations of the UBE1L gene could be detected. This leaves the possibility that it is due to a regulatory block of expression of the UBE1L gene.

The only mutations we found were three sequence polymorphisms with heterozygosities between 40% and 50%. These polymorphisms, espe-

cially the deletion/insertion polymorphism that is easy to detect directly, may be a welcome addition to the already known polymorphisms (Carritt et al., 1986; Kok et al., 1991) of the wider UBE1L region, which is known to be involved in loss of heterozygosity of many different types of tumors.

5-Methylcytosine is the only modified base occurring in vertebrate DNA. It is mainly present in the dinucleotide 5′-CpG-3′ (Bird, 1986). Hypomethylation of CpG sequences, especially those associated with hypersensitive regions, correlates positively with gene expression (Gross and Garrard, 1988). For UBE1L, this was only true for the HhaI site in the first exon, which was unmethylated in a lymphoblastoid cell line expression UBE1L at a high level, but almost fully methylated in the SCLC cell lines analyzed that virtually lack expression of UBE1L (Kok et al., 1993b). Maybe not coincidentally, it is this site that is included in the transcription-dependent DNAse I-hypersensitive region. However, an overall demethylation, as induced by culturing the cells in the presence of 5′-azacytidine (Dobkin et al., 1987), did not result in a marked increase of UBE1L mRNA levels in the SCLC cell lines. Thus, although the methylation status of this HhaI site may mark the level of expression, it does not directly affect it.

The absence of yet unknown transcription factors might be responsible for the low UBE1L expression in lung cancer-derived cell lines. In this study, some possible target regions for these proteins have been identified. These are now under further analysis.

ACKNOWLEDGEMENT

This work was supported by grant GUKC90-15 from the Dutch Cancer Society.

REFERENCES

A. P. Bird (1986), Nature 321, 209–213.

H. Brauch, B. Johnson, J. Hovis, T. Yano, A. Gazdar, O. S. Pettingill, S. Graziano, G. D. Sorenson, B. J. Poiesz, J. Minna, M. Linehan, and B. Zbar, N Engl J Med 317, 1109–1113.

B. Carritt, K. Kok, A. van den Berg, J. Osinga, A. Pilz, R. Hofstra, M. B. Davis, A. Y. van der Veen, P. H. Rabbitts, K. Gulati, and C. H. C. M. Buys (1992), Cancer Res 52, 1536–1541.

B. Carritt, H. M. Welch, and N. J. Parry-Jones (1986), Am J Hum Genet 38, 428–436.

S. A. Chrysogelos (1993), Nucleic Acids Res 21, 5736–5741.

A. Ciechanover, J. A. DiGuiseppe, B. Bercovich, A. Orian, J. D. Richter, A. L. Schwartz, and G. M. Brodeur (1991), Proc Natl Acad Sci USA 88, 139–143.

J. C. Cook and P. B. Chock (1992), J Biol Chem 267, 24315–24321.

R. M. Cook, B. J. Burke, D. L. Buchhagen, J. D. Minna, and Y. E. Miller (1993), J Biol Chem 268, 17010–17017.

J. Corden, B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedinger, and P. Chambron (1980), Science 209, 1406–1414.

M. C. Daly, R.-H. Xiang, D. Buchhagen, C. H. Hensel, D. K. Garcia, A. M. Killary, J. D. Minna, and S. L. Naylor (1993), Oncogene 8, 1825–1832.

C. Dobkin, C. Ferrando, and W. T. Brown (1987), Nucleic Acids Res 15, 3183.

R. Erlandsson, U. S. R. Bergerheim, F. Boldog, Z. Marcsek, K. Kunimi, B. Y.-T. Lin, S. Ingvarsson, J. S. Castresana, W.-H. Lee, G. Klein, and J. Sümegi (1990), Oncogene 5, 1207–1211.

M. A. Frohman, M. K. Dush, and G. R. Martin (1988), Proc Natl Acad Sci USA 85, 8998–9002.

D. S. Gross and W. T. Garrard (1988), Annu Rev Biochem 57, 159–197.

P. M. Handley, M. Mueckler, N. R. Siegel, A. Ciechanover, and A. L. Schwartz (1991), Proc Natl Acad Sci USA 88, 258–262, and erratum 88, 7456.

J. D. Hawkins (1988), Nucleic Acids Res 16, 9893–9908.

R. Kageyama, G. T. Merlino, and I. Pastan (1989), J Biol Chem 264, 15508–15514.

A. M. Killary, M. E. Wolf, T. A. Giambernardi, and S. L. Naylor (1992), Proc Natl Acad Sci USA 89, 10877–10881.

K. Kok, L. Snippe, G. AB, and M. Gruber (1985), Nucleic Acids Res 13, 5185–5202.

K. Kok, J. Osinga, B. Carritt, M. B. Davis, A. H. van der Hout, A. Y. van der Veen, R. M. Landsvater, L. F. M. H. de Leij, H. H. Berendsen, P. E. Postmus, S. Poppema, and C. H. C. M. Buys (1987), Nature 330, 578–581.

K. Kok, M.-Z. Fan, A. Jonas, A. van den Berg, B. Carritt, and C. H. C. M. Buys (1991), Nucleic Acids Res 19, 4797.

K. Kok, R. Hofstra, A. Pilz, A. van den Berg, P. Terpstra, C. H. C. M. Buys, and B. Carritt (1993a), Proc Natl Acad Sci USA 90, 6071–6075.

K. Kok, A. van den Berg, D. L. Buchhagen, B. Carritt, and C. H. C. M. Buys (1993b), Eur J Hum Genet 1, 156–163.

K. Kok, A. van den Berg, P. M. J. F. Veldhuis, A. Y. van der Veen, M. Franke, L. de Leij, E. P. M. Schoenmakers, M. M. F. Hulsbeek, W. van de Ven, and C. H. C. M. Buys (1994), Cancer Res 54, 4183–4187.

S. LaForgia, B. Morse, J. Levy, G. Barnea, L. A. Cannizzaro, F. Li, P. C. Nowell, L. Boghosian-Sell, J. Glick, A. Weston, C. C. Harris, H. Drabkin, D. Patterson, C. M. Croce, J. Schlessinger, and K. Heubner (1991), Proc Natl Acad Sci USA 88, 5036–5040.

L. Locker and G. Buzard (1990), J DNA Sequencing Mapping 1, 3–11.

A. L. Means and P. J. Farnham (1990), Mol Cell Biol 10, 653–651.

Y. E. Miller, J. D. Minna, and A. Gazdar (1989), J Clin Invest 83, 2120–2124.

S. L. Naylor, B. E. Johnson, J. D. Minna, and A. Y. Sakaguchi (1987), Nature 329, 451–454.

S. L. Naylor, A. Marshall, C. Hensel, P. F. Martinez, B. Holley, and A. Y. Sakaguchi (1989), Genomics 4, 355–361.

M. Orita, Y. Suzuki, T. Sekiya, and K. Hayashi (1989), Genomics 5, 874–879.

J. Sambrook, E. F. Fritsch, and T. Maniatis (1989), Molecular Cloning: A Laboratory Manual. Cold Spring Harbor University Press, Cold Spring Harbor, NY.

F. Sanger, S. Nicklen, and A. R. Coulson (1977), Proc Natl Acad Sci USA 74, 5463–5467.

G. Sarkar, H.-S. Yoon, and S. S. Sommer (1992), Genomics 13, 441–443.

A. Sehgal, N. Patil, and M. Chao (1988), Mol Cell Biol 8, 3160–3167.

P. Senapathy, M. B. Shapiro, and N. L. Harris (1990), Methods Enzymol 183, 252–278.

S. T. Smale and D. Baltimore (1989), Cell 57, 103–113.

E. Takahashi, D. Ayusawa, S. Kaneda, Y. Itoh, and T. Hori (1992), Cytogenet Cell Genet 59, 268–269.

J. S. Trausch, S. J. Grenfell, P. M. Handley-Gearhart, A. Ciechanover, and A. L. Schwartz (1993), Am J Physiol 264, 93–102.