SHORT COMMUNICATION

cDNA Sequencing and Analysis of POV1 (PB39): A Novel Gene Up-regulated in Prostate Cancer

Kristina A. Cole, Rodrigo F. Chuaqui, Kenneth Katz,* Svetlana Pack, Zhengping Zhuang, Catherine E. Cole, John C. Lyne,† W. Marston Linehan,† Lance A. Liotta, and Michael R. Emmert-Buck[1]

*Laboratory of Pathology, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, 10/2A33, 9000 Rockville Pike, Bethesda, Maryland 20892; *National Center for Biotechnology Information, National Library of Medicine and National Cancer Institute, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894; and †Urologic Oncology Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, 10/2B47, 9000 Rockville Pike, Bethesda, Maryland 20892*

Received February 5, 1998; accepted May 1, 1998

We recently identified a novel gene (PB39) (HGMW-approved symbol POV1) whose expression is up-regulated in human prostate cancer using tissue microdissection-based differential display analysis. In the present study we report the full-length sequencing of PB39 cDNA, genomic localization of the PB39 gene, and genomic sequence of the mouse homologue. The full-length human cDNA is 2317 nucleotides in length and contains an open reading frame of 559 amino acids which does not show homology with any reported human genes. The N-terminus contains charged amino acids and a helical loop pattern suggestive of an srp leader sequence for a secreted protein. Fluorescence *in situ* hybridization using PB39 cDNA as probe mapped the gene to chromosome 11p11.1–p11.2. Comparison of PB39 cDNA sequence with murine sequence available in the public database identified a region of previously sequenced mouse genomic DNA showing 67% amino acid sequence homology with human PB39. Based on alignment and comparison to the human cDNA the mouse genomic sequence suggests there are at least 14 exons in the mouse gene spread over approximately 100 kb of genomic sequence. Further analysis of PB39 expression in human tissues shows the presence of a unique splice variant mRNA that appears to be primarily associated with fetal tissues and tumors. Interestingly, the unique splice variant appears in prostatic intraepithelial neoplasia, a microscopic precursor lesion of prostate cancer. The current data support the hypothesis that PB39 plays a role in the development of human prostate cancer and will be useful in the analysis of the gene product in further human and murine studies. © 1998 Academic Press

Relatively few discovery-oriented studies assessing cancer-associated gene expression as it exists *in vivo* have been done, largely due to the difficulties encountered in working with human tissue samples. However, technological advances in methods for tissue microdissection, differential display, and microarray-based gene expression analysis are now allowing access to the gene expression profiles of microscopic, histopathologically defined cell types. Using a degenerate primer PCR approach to compare pure populations of normal prostate epithelium to invasive tumor microdissected from a patient's prostatectomy specimen, we previously identified an expressed sequence tag (EST), GenBank Accession No. R00504, which was substantially overexpressed in the tumor cells of a 47-year-old gentleman with clinically aggressive prostate cancer (3). Similar analysis of 10 patients with prostate neoplasms showed PB39[2] mRNA to be up-regulated in 50% of the cases, indicating that it may play a role in prostate cancer development and/or progression. In this report the cloning, sequencing, and genomic mapping of the full-length PB39 cDNA are described.

Northern blot analysis using R00504 as a probe showed a transcript of approximately 2.6 kb that was expressed in tissues of the adult colon, small intestine, ovary, prostate, spleen, and pancreas (Fig. 1A); fetal kidney, liver, and lung (Fig. 1B); and adult liver and skeletal muscle (results not shown). The level of expression was highest in adult pancreas tissue (Fig. 1A).

To determine the complete cDNA sequence of PB39, specific 5′ and 3′ PCR products were generated from a pancreas cDNA library using the rapid amplification of cDNA ends (RACE) method (Marathon-Ready cDNA; Clontech). The 3′ and 5′ PB39 RACE primer sequences were chosen from EST clone R00504 (gaccgcatagacttctcaga and gcatgttacaggtagaaaagcc, respectively). A 700-bp 3′ fragment and a 2-kb 5′ fragment were pro-

Sequence data from this article have been deposited with the GenBank Data Library under Accession No. AF045584.

[1] To whom correspondence should be addressed. Telephone: (301) 496-2912. E-mail: mbuck@helix.nih.gov.

[2] The HGMW-approved symbol for the gene described in this paper is POV1.

0888-7543/98 $25.00

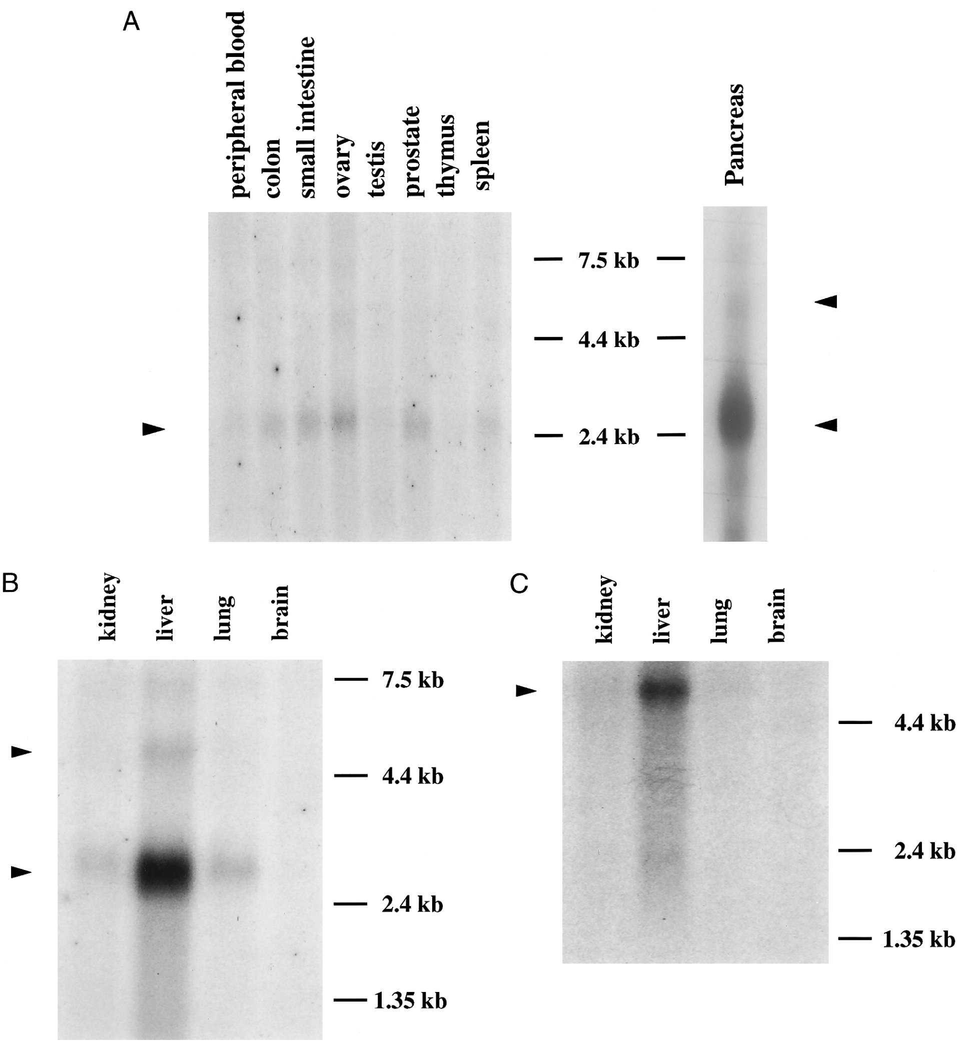

FIG. 1. (**A**) Clontech adult tissue Northerns probed with radiolabeled R00504 insert. The exposure times of the left and right blots were 40 and 6 h, respectively. (**B**) Clontech fetal tissue Northern probed with radiolabeled R00504 insert. The exposure time was 40 h. (**C**) Clontech fetal tissue Northern probed with radiolabeled 5-kb transcript-specific probe. The exposure time was 5 days. The amount of RNA loaded on all of the gels was adjusted to give similar β-actin hybridization signals and represents approximately 2 μg of poly(A)-selected mRNA.

duced, subcloned into a plasmid PCR vector, and cycle sequenced. To validate the sequence, gene-specific PCR products amplified from the pancreas library were directly sequenced and were verified by 10 independent sequencing reactions. Assembly of the entire set of sequences produced a 2317-nucleotide cDNA with 76 nucleotides of 5′ untranslated sequence, a 1677-nucleotide open reading frame (559 amino acids), and 564 nucleotides of 3′ untranslated sequence (Fig. 2A). The consensus Kozac sequence GCCGCCATGG placed the translation initiation methionine at nucleotide position 77 (9).

Basic Local Alignment Search Tool analysis (1) of PB39 sequence against the public human EST database showed multiple PB39 EST clones from diverse tissue types. Interestingly, several PB39 homologous clones (Fig. 2B) showed an identical divergence within the coding sequence (at nucleotide 1610) with introduction of additional nucleotide sequence. To analyze this longer, alternative form of PB39 further, a PCR primer specific for the inserted sequence (tctgcaaagtggctgagatgag) was designed and used to amplify cDNA from the pancreas library with a PB39-specific 5′ primer (cctgccttatctttctgaactgcacc). The amplified product was isolated and directly sequenced. Subsequent analysis of the open reading frame showed the addition of 48 new amino acids at nucleotide position 1613 followed by a stop codon. Thus, the larger transcript of PB39 encodes a 560-amino-acid protein that has replaced the 47 C-terminal amino acids found in the 2.3-kb PB39 with 48 new amino acids (Figs. 2A and 2B).

Northern blot analysis using a probe specific for the inserted sequence showed a 5-kb PB39 transcript expressed in adult pancreas and fetal liver tissue. As expected, this probe did not hybridize to the 2.3-kb PB39 (Fig. 1C). A longer exposure of the R00504-probed Northerns did reveal a less intense transcript at 5 kb, which would be expected since this sequence is common to both transcripts (Figs. 1A and 1B).

RT-PCR analysis was performed to study the expres-

```
A     1 ccggggctggagggggcaagcgggttccgaggtgcaaagcctgg
        tgccccgagccctgcggagctcggggccagc
     77 atggcccccacgctgcaacaggcgtaccggaggcgctggtggatg
         M  A  P  T  L  Q  Q  A  Y  R  R  R  W  W  M    15
    122 gcctgcacggctgtgctggagaacctcttcttctctgctgtactc
         A  C  T  A  V  L  E  N  L  F  F  S  A  V  L    30
    167 ctgggctggggctccctgttgatcattctgaagaacgagggcttc
         L  G  W  G  S  L  L  I  I  L  K  N  E  G  F    45
    212 tattccagcacgtgcccagctgagagcagcaccaacaccacccag
         Y  S  S  T  C  P  A  E  S  S  T  N  T  T  Q    60
    257 gatgagcagcgcaggtggccaggctgtgaccagcaggacgagatg
         D  E  Q  R  R  W  P  G  C  D  Q  Q  D  E  M    75
    302 ctcaacctgggcttcaccattggttccttcgtgctcagcgccacc
         L  N  L  G  F  T  I  G  S  F  V  L  S  A  T    90
    347 accctgccactggggatcctcatggaccgctttggcccccgaccc
         T  L  P  L  G  I  L  M  D  R  F  G  P  R  P   105
    392 gtgcggctggttggcagtgcctgcttcactgcgtcctgcaccctc
         V  R  L  V  G  S  A  C  F  T  A  S  C  T  L   120
    437 atggccctggcctcccgggacgtggaagctctgtctccgttgata
         M  A  L  A  S  R  D  V  E  A  L  S  P  L  I   135
    482 ttcctggcgctgtccctgaatggctttggtggcatctgcctaacg
         F  L  A  L  S  L  N  G  F  G  G  I  C  L  T   150
    527 ttcacttcactcacgctgcccaacatgtttgggaacctgcgctcc
         F  T  S  L  T  L  P  N  M  F  G  N  L  R  S   165
    572 acgttaatggccctcatgattggctcttacgcctcttctgccatt
         T  L  M  A  L  M  I  G  S  Y  A  S  S  A  I   180
    617 acgttcccaggaatcaagctgatctacgatgccggtgtggccttc
         T  F  P  G  I  K  L  I  Y  D  A  G  V  A  F   195
    662 gtggtcatcatgttcacctggtctggcctggcctgccttatcttt
         V  V  I  M  F  T  W  S  G  L  A  C  L  I  F   210
    707 ctgaactgcaccctcaactggcccatcgaagcctttcctgcccct
         L  N  C  T  L  N  W  P  I  E  A  F  P  A  P   225
    752 gaggaagtcaattacacgaagaagatcaagctgagtgggctggcc
         E  E  V  N  Y  T  K  K  I  K  L  S  G  L  A   240
    797 ctggaccacaaggtgacaggtgacctcttctacacccatgtgacc
         L  D  H  K  V  T  G  D  L  F  Y  T  H  V  T   255
    842 accatgggccagaggctcagccagaaggcccccagcctggaggac
         T  M  G  Q  R  L  S  Q  K  A  P  S  L  E  D   270
    887 ggttcggatgccttcatgtcaccccaggatgttcggggcacctca
         G  S  D  A  F  M  S  P  Q  D  V  R  G  T  S   285
    932 gaaaaccttcctgagaggtctgtccccttacgcaagagcctctgc
         E  N  L  P  E  R  S  V  P  L  R  K  S  L  C   300
    977 tcccccactttcctgtggagcctcctcaccatgggcatgacccag
         S  P  T  F  L  W  S  L  L  T  M  G  M  T  Q   315
   1022 ctgcggatcatcttctacatggctgctgtgaacaagatgctggag
         L  R  I  I  F  Y  M  A  A  V  N  K  M  L  E   330
   1067 taccttgtgactggtggccaggagcatgagacaaatgaacagcaa
         Y  L  V  T  G  G  Q  E  H  E  T  N  E  Q  Q   345
   1112 caaaaggtggcagagacagttgggttctactcctccgtcttcggg
         Q  K  V  A  E  T  V  G  F  Y  S  S  V  F  G   360
   1157 gccatgcagctgttgtgccttctcacctgcccccctcattggctac
         A  M  Q  L  L  C  L  L  T  C  P  L  I  G  Y   375
```

FIG. 2. (**A**) Nucleotide and amino acid sequence of PB39. The nucleotide sequence is numbered on the left and the amino acid sequence numbered on the right. The underlined ATG start is at nucleotide position 77. When two sequences are present at nucleotide position 1613, the upper nucleotide and amino acid sequences refer the 2.3-kb transcript. The lower nucleotide and amino acids sequences refer to the 5-kb transcript. (**B**) Sequence overlap and divergence between 2.3- and 5-kb transcripts (top and bottom, respectively). Open reading frame (between arrowheads), 5′UTR (vertical line pattern), 3′UTR (area downstream of arrowhead), inserted sequence of the 5-kb transcript (black), and position of divergence (arrow). The white area corresponds to the same sequence in both transcripts. Representative EST clones are identified below each transcript.

sion of the 5-kb PB39 transcript in human prostate tissue as previously described (3, 6). RT-PCR using primers directed against the inserted sequence in microdissected normal and invasive prostate epithelium showed a product in 4/4 tumor samples, but only 1/4 corresponding normal samples (results not shown). One of the cases overexpressing the 5-kb transcript did not show overexpression of the 2.3-kb form of PB39. Our previous report showed overexpression of PB39 in 5 of 10 tumor samples; thus it appears that both the 2.3- and the 5-kb forms of PB39 are up-regulated in unique subsets of tumors. The physiological significance of this finding is not yet clear.

Cloning and sequence analysis of full-length human PB39 cDNA has shown it to be a unique gene with little homology to previously reported human genes. Both

```
1202 atcatggactggcggatcaaggactgcgtggacgccccaactcag
      I  M  D  W  R  I  K  D  C  V  D  A  P  T  Q  390
1247 ggcactgtcctcggagatgccagggacggggttgctaccaaatcc
      G  T  V  L  G  D  A  R  D  G  V  A  T  K  S  405
1292 atcagaccacgctactgcaagatccaaaagctcaccaatgccatc
      I  R  P  R  Y  C  K  I  Q  K  L  T  N  A  I  420
1337 agtgccttcaccctgaccaacctgctgcttgtgggttttggcatc
      S  A  F  T  L  T  N  L  L  L  V  G  F  G  I  435
1382 acctgtctcatcaacaacttacacctccagtttgtgacctttgtc
      T  C  L  I  N  N  L  H  L  Q  F  V  T  F  V  450
1427 ctgcacaccattgttcgaggtttcttccactcagcctgtgggagt
      L  H  T  I  V  R  G  F  F  H  S  A  C  G  S  465
1472 ctctatgctgcagtgttcccatccaaccactttgggacgctgaca
      L  Y  A  A  V  F  P  S  N  H  F  G  T  L  T  480
1517 ggcctgcagtccctcatcagtgctgtgttcgccttgcttcagcag
      G  L  Q  S  L  I  S  A  V  F  A  L  L  Q  Q  495
1562 ccacttttcatggcgatggtgggaccccctgaaaggagagcccttc
      P  L  F  M  A  M  V  G  P  L  K  G  E  P  F  510
1607 tgggtgaatctgggcctcctgctattctcactcctgggattcctg
           agagcgagggttggtgtgggggagcaggagccactctc
      W  V  N  L  G  L  L  L  F  S  L  L  G  F  L  525
            R  A  R  V  G  V  G  G  A  G  A  T  L  525
1652 ttgccttcctacctcttctattaccgtgcccggctccagcaggag
     ctgggggcaggggtagggccttgtatgtggtgccatccctcactc
      L  P  S  Y  L  F  Y  Y  R  A  R  L  Q  Q  E  540
      L  G  A  G  V  G  P  C  M  W  C  H  P  S  L  540
1697 tacgccgccaatgggatgggcccactgaaggtgcttagcggctct
     atctcagccagaggcacctcagaggtctctaatctgcaggtttcc
      Y  A  A  N  G  M  G  P  L  K  V  L  S  G  S  555
      I  S  A  R  G  T  S  E  V  S  N  L  Q  V  S  555
1742 gaggtgaccgcatag     1756
     aagttgtctgccttttag  1759
      E  V  T  A  * 559
      K  L  S  A  F  *#560

     acttctcagaccaagggacctggatgacaggcaatcaaggcctga
     gcaaccaaaaggagtgccccatatggcttttctacctgtaacatg
     cacatagagccatggccgtagatttataaataccaagagaagttc
     tatttttgtaaagactgcaaaaaggaggaaaaaaaaccttcaaaa
     acgcccccctaagtcaacgctccattgactgaagacagtccctatc
     ctagaggggttgagctttcttcctccttgggttggaggagaccag
     ggtgcctcttatctccttctagcggtctgcctcctggtacctctt
     ggggggatcggcaaacaggctacccctgaggtcccatgtgccatg
     agtgtgcacaacatgcaatgtgtctgtgtatgtgtgccatgaatg
     tgagaaaaacacagccctcctttcagaaggaaaggggcctgaggg
     ctgtgtcctgggttaggggttgggggtcggccccttccagggcca
     ggaaggcaggttccctctctggtgctgctgcttgcaagtcttaga
     ggaaataaaaagggaagtgag aaaaaaaaa
```

#there are approximately an additional 2300 nucleotides at this site contributing to the 5 kb transcript's 3′UTR.

FIG. 2—*Continued*

the 2.3- and the 5-kb PB39 cDNA sequences were compared against available genomic data in GenBank, and two mouse chromosome 11 matches were found (GenBank Accession Nos. AC000400 and AC002121). No matches with human genomic sequence were identified. The mouse PB39 genomic sequence contains several regions of strong homology with the human PB39 coding sequence and appears to contain at least 14 exons that code for 455 amino acids. Since no murine cDNA sequence is available, precise assignment of intron–exon boundaries in the genomic mouse DNA sequence is not possible. The available mouse genomic DNA sequence terminates and does not contain sequence homologous to the C-terminus of human PB39, including the 48 C-terminal amino acids that differ between the 2.3- and the 5-kb splice variants of human PB39. Comparison of the available 455-amino-acid mouse sequence with the corresponding human sequence shows an overall identity of 67% and similarity of 83%.

PHDhtm analysis (12, 14) of PB39 protein highlighted the N-terminal region between residues 22 and 39 as likely to encode a helical transmembrane domain. This region is highly conserved between mouse and human: 41 of the first 47 amino acids encoded are identical between these species, and 3 of the 6 nonidentical residues represent conservative substitutions. Because the residues surrounding this putative trans-

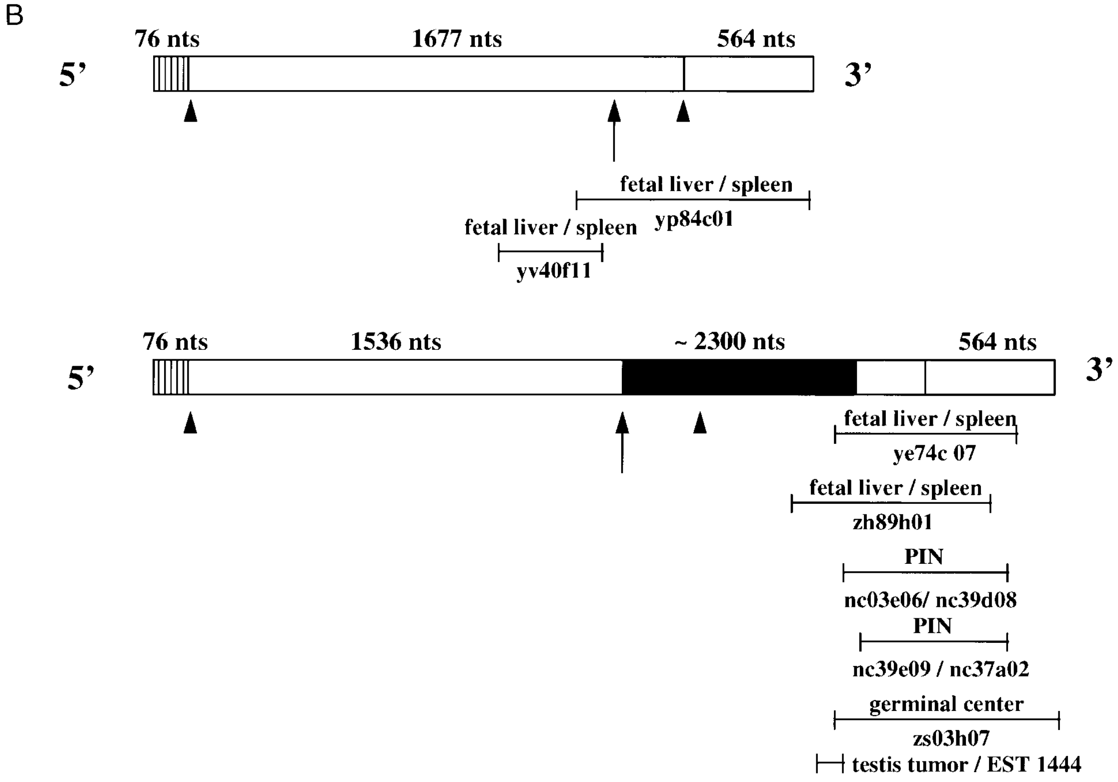

FIG. 2—*Continued*

membrane domain are predominantly charged, it is possible that the amino-terminal portion of PB39 encodes a signal sequence. We further analyzed this N-terminal sequence using SignalP (10), a computational tool designed to identify probable secretory signal peptides. The results suggested the existence of a secretory signal peptide within the first 50 amino acid residues (cleavage between residues 44 and 45, both mouse and human). We caution that many different types of N-terminal signals found in proteins share some or all of these general features (amino-terminal helical domain bordered by charged residues), e.g., nuclear targeting signals and signal peptides of secretory proteins as well as those targeting proteins to mitochondria. In addition to the secretory signal, a leucine zipper motif was observed from peptide 149 to 170.

A Whitehead Institute STS marker WI-17004 (GenBank Accession No. G22380) maps PB39 to 291.1 cR from the telomere of the short arm of chromosome 11 (13). To confirm this result, *in situ* hybridization was performed following protocols previously described (7, 11). The chromosomal localization of the gene was determined by hybridization of the 2.3-kb 5′ RACE PCR product to metaphase chromosomes and converting DAPI banding to "G-banding" using IP Lab Spectrum software for chromosomal identification. The result indicated that the PB39 gene maps to human chromosome 11p11.1–p11.2. A total of 50 cells were examined to determine the precise chromosomal location of the probe. In all metaphases scored, clear signals were seen on the short arm of chromosome 11 (Figs. 3A and 3B). DAPI banding unambiguously showed the position of the signal in the region 11p.11.1–p11.2 (Figs. 3A–3D). Both the STS primers and the FISH probe were directed against sequence common to both the 2.3- and the 5-kb transcripts. In both cases hybridization to only one chromosomal location was identified. Interestingly, the human chromosome 11p11–p12 region has been postulated to harbor one or more metastasis-suppressor genes, including KAI1, and has also been shown to be deleted in 70% of advanced prostate cancers (5, 8).

The 5-kb PB39 transcript was found to match EST clones primarily from fetal and tumor tissue libraries (Fig. 2B). This transcript was also highly expressed in a cDNA library from a microdissected prostatic intraepithelial neoplasia focus that was sequenced as part of the Cancer Genome Anatomy Project (http://www.ncbi.nlm.nih.gov/ncicgap/). Thus, PB39 represents one of the first identified genes that have been shown to be increased early in prostate cancer development. The cellular regulation of PB39 mRNA splice variants, their precise expression levels during prostate tumorigenesis, and the functional significance of altering the C-terminal 47 amino acids remain to be determined.

The predicted N-terminal amino acid sequence of PB39 suggests the presence of an srp sequence for a secreted protein. Certainly, identification of a potential serum protein that is increased early in the progression of prostate cancer, in prostatic intraepithelial neoplasia (PIN) for example, would be a use-

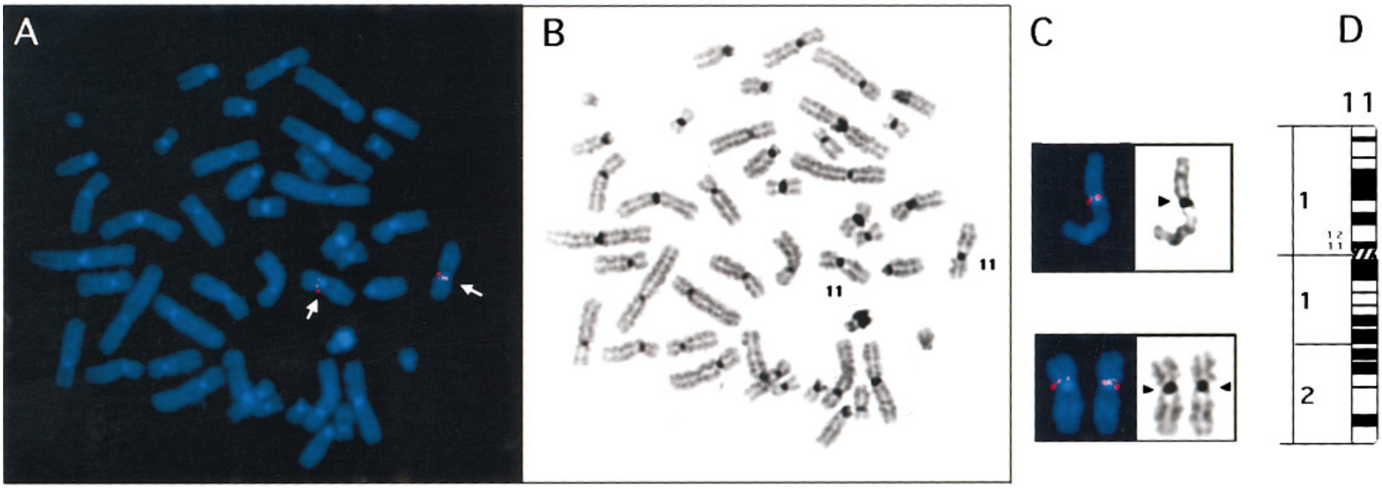

FIG. 3. Regional localization of PB39. (**A**) Localization to chromosome 11p. DAPI-counterstained metaphase showing the location of the PB39 gene (red) on the short arm of chromosome 11. (**B**) G-banded chromosomal analysis. An example of chromosome 11 DAPI image (A) that was converted to a black and white G-band image to show the position of PB39 on the short arm of chromosome 11 close to the centromere. The position of PB39 is identified at 11p11.1–p11.2. (**C**) DAPI and simulated G-banded image of chromosome 11 after FISH; arrows show location of PB39. (**D**) Idiogram of chromosome 11.

ful tool for the early detection of prostate cancer. Epidemiologic studies have shown that PIN precedes the development of prostate cancer by several decades in most men, thus a marker of early malignancy could identify those men who develop PIN lesions early in life and are at greatest risk for developing clinically significant disease (2).

REFERENCES

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.
2. Bostwick, D. G., Pacelli, A., and Lopez-Beltran, A. (1996). Molecular biology of prostatic intraepithelial neoplasia. *Prostate* **29:** 117–134.
3. Chuaqui, R., Englert, C., Strup, S., Vocke, C., Zhuang, Z., Duray, P., Bostwick, D., Linehan, M., Liotta, L., and Emmert-Buck, M. R. (1997). Identification of a novel transcript upregulated in a clinically aggressive prostate carcinoma. *Urology* **50:** 302–307.
4. Dalbey, R. E., Lively, M. O., Bron, S., and van Dijl, J. M. (1997). The chemistry and enzymology of the type I signal peptidases. *Protein Sci.* **6:** 1129–1138.
5. Dong, J. T., Lamb, P. W., Rinker-Schaeffer, C. W., Vukanovic, J., Ichikawa, T., Isaacs, J. T., and Barrett, J. C. (1995). KAI1, a metastasis suppressor gene for prostate cancer on human chromosome 11p11.2. *Science* **268:** 884–886.
6. Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zuang, Z., Goldstein, S. R., Weiss, R. A., and Liotta, L. A. (1996). Laser capture microdissection. *Science* **274:** 998–1001.
7. Hirai, M., Suto, Y., and Kanoh, M. (1994). A method for simultaneous detection of fluorescent G-bands and in situ hybridization signals. *Cytogenet. Cell Genet.* **66:** 149–151.
8. Kawana, Y., Komiya, A., Ueda, T., Nihei, N., Kuramochi, H., Suzuki, H., Yatani, R., Imai, T., Dong, J. T., Imai, T., Yoshie, O., Barrett, J. C., Isaacs, J. T., Shimazaki, J., Ito, H., and Ichikawa, T. (1997). Location of KAI1 on the short arm of human chromosome 11 and frequency of allelic loss in advanced prostate cancer. *Prostate* **32:** 205–213.
9. Kozac, M. (1991). Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* **266:** 19867–19870.
10. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6. [www server: http://www.cbs.dtu.dk/services/SignalP/]
11. Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J., and Gray, J. W. (1988). Fluorescence in situ hybridization with human chromosome-specific libraries: Detection of the trisomy 21 and translocations of chromosome 4. *Proc. Natl. Acad. Sci. USA* **85:** 9138–9142.
12. Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1997). Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4:** 521–533.
13. Schuler, G. D. (1997). Sequence mapping by electronic PCR. *Genome Res.* **7:** 541–550.
14. Walker, D. R., and Koonin, E. V. (1997). SEALS: A system for easy analysis of lots of sequences. *ISMB* **5:** 333–339.