

Structure of the Human Ubiquitin Fusion Gene Uba80 (RPS27a) and One of Its Pseudogenes

Lawrence S. Kirschner¹ and Constantine A. Stratakis

Unit on Genetics and Endocrinology, Developmental Endocrinology Branch, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892

Received March 19, 2000

Ubiquitin is a highly conserved 76 amino acid protein that is generated in the cell by proteolysis of larger proteins containing either polyubiquitin chains or ubiquitin fused to carboxyl extension proteins (CEPs). In humans, the two human ubiquitin-CEP genes are Uba80 and Uba52, which code for ubiquitin fused to ribosomal protein S27a and L40, respectively. Working from a recently generated physical map of human chromosome 2p16, we determined the genetic and physical location and the genomic structure of the Uba80 gene in its entirety. A comparison of Uba80 to Uba52 revealed that the two genes share a conserved 5'-end structure, but that the structure of the ubiquitin coding regions was not conserved. Analysis of 400 bp of the promoter of Uba80 revealed strong similarity not only to the Uba52 promoter, but also to the other known human ribosomal gene promoters that have been identified to date. Homology searches also detected the presence of a pseudogene for Uba80, and the structure of this sequence feature is also reported. © 2000 Academic Press

Key Words: genomic structure; genomic mapping; ribosomal proteins; ubiquitin.

Ubiquitin is a 76 amino acid protein that is highly conserved during evolution and plays a key role in the regulated degradation of many cellular proteins by targeting them for degradation in the 26S proteasome (1). Ubiquitin itself is synthesized in the cell as precursor proteins which consist either of polyubiquitin chains that are cleaved into individual ubiquitin moieties (UbB or UbC type) or single ubiquitin moieties fused 5' to unrelated proteins (UbA type). In humans, as in yeast and other mammals studied to date, there are two UbA genes, one coding for ubiquitin fused to a

carboxyl extension protein (CEP) of 76–80 residues, and one coding for ubiquitin fused to a smaller protein of 50–52 residues (1, 2).

Analysis of the CEPs has shown that they contain features typical of nucleic acid binding proteins, including many positively charged residues as well as hydrophobic domains (3), and subsequent work demonstrated that the yeast and human CEPs were identical to components of the ribosome. In the yeast system, the large and small fusion genes are designated UBI3 and UBI2/UBI1, respectively (1). The corresponding genes in humans are Uba80 (HUBCEP80) and Uba52 (HUBCEP52), which code for ubiquitin fused to ribosomal protein S27a (RPS27a), and ribosomal protein L40 (RPL40), respectively (4–6). Because these genes have been studied from both the ubiquitin and ribosomal gene perspectives, there is significant confusion about their nomenclature; both Uba80 and Uba52 have been previously assigned 3 different designations. For example, the gene coding for ubiquitin fused to the 80 amino acid carboxyl extension protein has been designated Uba80, HUBCEP80 and RPS27a by different authors. In fact, the human genome organization (HUGO, URL <http://www.gene.ucl.ac.uk/nomenclature/>) has not resolved this issue either, referring to the ubiquitin-80 residue gene as RPS27a, and the ubiquitin-52 residue gene as Uba52. In this report, we shall refer to these genes by their initial designations as Uba80 and Uba52 (7).

The function of these ubiquitin-ribosomal protein fusion genes has not been fully elucidated, although there is convincing evidence in yeast that the ubiquitin moiety serves to increase the efficiency of the ribosomal protein's transport to the nucleus and incorporation into nascent ribosomes (8). For this reason, the Uba80 and Uba52 genes may be more important as a component of the respective ribosomal genes than as significant sources of ubiquitin within the cell. Although there are approximately 80 ribosomal proteins that have been identified, fewer than one quarter of them have been characterized at the level of gene structure

¹ To whom correspondence should be addressed at Unit on Genetics and Endocrinology, DEB, NICHD, NIH, Building 10, Room 10N262, 10 Center Drive, MSC1862, Bethesda, Maryland 20892-1862. Fax: 301-4020574. E-mail: kirschnl@mail.nih.gov.

(9–20). This report presents the characterization of the Uba80 gene coding for the ubiquitin-RPS27a protein fusion and a pseudogene for this message (ψ -Uba80). Because of its potential role in both protein synthesis by the ribosome and protein degradation by the proteasome, the structure of this gene will be of interest to groups studying both of these processes.

MATERIALS AND METHODS

Generation and sequencing of Uba80 PCR fragments. All YACs and BACs used in this study were obtained from commercial sources and manipulated as suggested by the supplier (Research Genetics, Huntsville, AL). CEPH YAC 919_A_6 was selected for analysis based on published EST content data (<http://www-genome.wi.mit.edu/>). The sequences for primers used for PCR amplification were generated using the Whitehead Institute's WWW-accessible program Primer3 (http://www-genome.wi.mit.edu/genome_software/other/primer3.html) and synthesized on an ABI 370 DNA synthesizer (Applied Biosystems, Foster City, CA). Sequences of the primers were: 1L, GAAAACCCCTTACGGGAAGA; 1R, CCACGAAGTCTCAACACAAGA; 1.5L, CAGAGACTGATCTTTGCTGGC; 1.5R, TTTCCAGAATACCTCATTTAAACC; 2L, TCGTGGTGGTGCTAA-GAAAA; 2R, CTTGTCTTCTGGTTTGTGAAA.

PCR conditions used were as described (21). Fragments were cloned using a T-overhang-based cloning system (TA Cloning, Invitrogen, Carlsbad, CA). To eliminate PCR-derived errors, three independent clones were picked for each of the fragments and sequenced in their entirety.

Generation of 5' and 3' genomic structure. To determine additional genomic structure from Uba80, a bacterial artificial chromosome (BAC 388-L-12) was isolated for the RPS27a EST. DNA was isolated as described and used for direct sequencing using primers from the 5' and 3' ends of the full-length cDNA sequence (22).

5' RACE analysis To determine the 5' end of the message, poly-A mRNA from human placenta was reverse transcribed using the Marathon cDNA kit and 5' RACE was performed as directed by the kit's manufacturer (Clontech, Palo Alto, CA). Products were analyzed by gel electrophoresis; distinct bands were cut from the gel, purified, and sequenced directly using fluorescent dideoxy-terminators (Perkin-Elmer, Foster City, CA).

RESULTS AND DISCUSSION

As part of our work on genes located on chromosome 2p16, we have undertaken a cataloging of ESTs localized to the 2p15-2p21 region. The EST RPS27a (23), generated from the Uba80 gene, was reported to be localized in this area, and we have recently confirmed this placement by physical and genetic (radiation hybrid) mapping (21, 24). The EST is located between the polymorphic markers D2S378 and D2S2251, and is located near the recently isolated gene for Doyme's Honeycomb Retinal Dystrophy, EFEMP1 (25). This location, however, rules it out as a candidate for the tumor syndrome Carney Complex, as the genetic locus for this disease lies telomeric to D2S2251 (C.A.S., L.S.K., and S. E. Taymans, in preparation).

To determine the intron-exon structure of the gene, 3 sets of primers (Fig. 1) were designed to generate overlapping cDNA fragments of the Uba80 gene. When

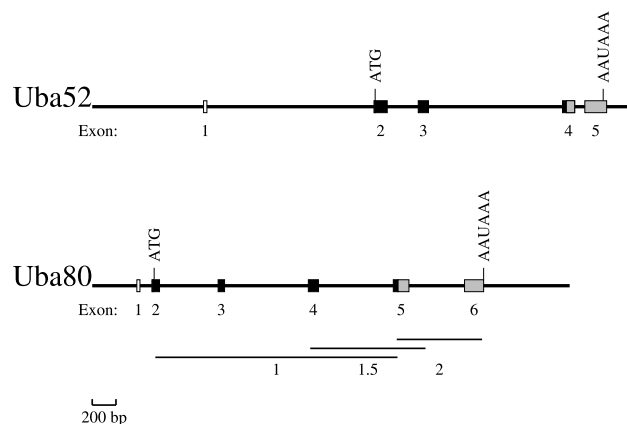


FIG. 1. Comparison of the structure of the human Uba80 and Uba52 genes. The structure of the Uba80 gene was determined by the sequencing of 3 overlapping fragments (1, 1.5, 2) generated by PCR amplification of YAC DNA, as described in the text. Unfilled boxes represent 5'-untranslated exons. Black boxes represent the ubiquitin coding region, and stippled boxes represent the 3' carboxyl terminal extension proteins.

total human genomic DNA from normal individuals was used as a template for PCR amplification, fragments corresponding in size to the cDNA were obtained (data not shown). However, when a 2p16 yeast artificial chromosome was used (CEPH YAC 919_A_6), fragments of a larger size were obtained and subsequently cloned and sequenced (see Materials and Methods). This analysis demonstrated that the human Uba80 gene consists of 6 exons, and that the coding region of the gene spans an entire length of approximately 2.9 kb. A comparison of the structure of the Uba80 gene with the closely related Uba52 gene (26) reveals a number of interesting features (Fig. 1). First, the structure of the 5' end of the gene appears to be well conserved, consisting of a very short, non-coding first exon. The second exon contains the initiator ATG which begins the ubiquitin portion of the protein. As shown in Fig. 1, the size of intron 1 is much greater in the Uba52 (1399 bp) gene than in the Uba80 gene (99 bp).

Although the structure of the ubiquitin-coding regions from the two genes is similar in organization, it is interesting to note that the first 34 residues of ubiquitin are contained in a single exon in Uba52, whereas this region is split into two distinct exons in Uba80. The next 28 residues are contained in an exon with a conserved 5' splice acceptor site, but whose 3' splice donor site is shifted 1 bp C-terminal in Uba52. The remainder of the ubiquitin moiety, including the Gly-Gly cleavage site, is contained in the following exon, along with the N-terminal portion of the extension protein. Each gene has a unique final exon coding for the C-terminal tail of the extension protein, and includes a canonical poly(A) cleavage site (AAUAAA). The finding that Uba80 and Uba52 contain introns within the coding region marks another important dif-

A

```

-330          catcttggga caagtagtcc aaggtagaag
-300  aaagcagtgg agaactacag ttcccagaat gcacttccat ttcctacctc
-250  ttcccagaaga ggccccccctc gacctccttt taaaaattct cttagccaag
-200  ttgattgtac gggaaaagcc tttttaaaac atcttttacg ttgcttaaac
-200  ctacagtttc gaaagcatto cgaaggctaa agtgagaaat aagcccaggc
-150  tagggagagg agaaacgaag ttcacgtcct agtctggcac cgggttgat
-100  tgtcgctggg acggcagtca ggcatttggt gtggtgcctc aaggggtggg
-50   tccttcggcg ggagctccgg gaaaccccg ggcctgcgc ggcgttctc
1     CTTTTCGATC CGCCATCTGC Ggtgggtgto tgcacttcgg ctgctctcgg
51   gttagcacc c tatggtgctc tctcttgtga tcctgacct aacctgtctc

```

B

Gene	GB acc.	nt	
Uba80 promoter	1	43	<u>AACTACAGTTCCCAGAATGCA</u>
iNOS promoter	X76303	64	<u>AACTACAGTTCCCAGAATGCA</u>
hPOP1/p14.5 promoter	Y15082	1972	<u>AACTACAGTTCCCAGAATGCC</u>
8p21-p22 genomic seq	AF134726	38825	<u>AACTACAGTTCCCAGAATTC</u>
6p21.3 genomic seq	AB020866	19448	<u>AACTACAGTTCCCAGAAGGCC</u>
Uba52	X56997	2802	<u>ACTCTCAGACTACAACATCCA</u>
RPL7a	X61923	2566	<u>AAATACAGTCTTCAGCTAATG</u>
RPL9	U09954	2584	<u>GCCTTAAGTTCCTTGAAAGGCA</u>
RPL19	X82201	2	<u>CAGTACAGTTGTGAGCCAAAT</u>
RPL38	Z26876	87	<u>CGCGCCAGAGCCCAGCGCGCC</u>
RPS4X	L24369	58	<u>GCGCGCAGTCTCCAGCCCCAA</u>
RPS4Y	L24370	31	<u>AACTAACTTCTCACTCATCT</u>
Consensus			<u>AACTACAGTTCCCAGAAAGC</u>

FIG. 2. (A) Sequence of the human Uba80 promoter. The transcription start site mapped in this report is located at +12 nt, and is indicated with an asterisk. The transcription start site as mapped by Adams *et al.* (29) is used as +1. Exon sequence is shown in bold capitals, with promoter and first intron sequence shown in lower case. The SP-1 element (-46) is fully underlined and the sequence element found in other ribosomal promoters (-34) is underlined with a broken line. The iNOS homology element (-288) is double underlined. (B) A comparison of the iNOS element in the Uba80 gene with other sequences in the GenBank database. The top portion of the figure shows sequences with very high homology to this element, whereas the bottom shows similar sequences from other ribosomal protein genes. The central portion of the element, which appears better conserved in the RP genes, is boxed in the lower half of the figure.

ference between these genes and the other ubiquitin genes (UbB and UbC), which code for polyubiquitin but lack introns in their coding regions (7, 27).

Using these techniques, we have mapped the start site to a C residue located within a polypyrimidine tract, the same site which had been mapped previously (28). However, another report placed the 5' end of the message 12 nucleotides further upstream within the same polypyrimidine tract (29). This minor discrepancy likely reflects multiple start sites for the Uba80 gene, and may reflect a bias of the 5'-RACE technique that was used in this report. Because the different start sites are not predicted to affect the coding region of the message, this longer 5' end will be designated as +1 in discussion of the Uba80 promoter (Fig. 2).

Analysis of the promoter of the Uba80 gene reveals a GC-rich region that lacks canonical TATA or CAAT

boxes upstream from transcription initiation (Fig. 2A). The region contains an SP1 box at bp -46, which is followed directly at bp -34 by a match for an element (tCCGGGAAA vs a/gCCGGAAa/g) that has been found in the promoter for many of the clones ribosomal genes including the human RPS4X, RPS4Y, and snRNP E, and mouse RPL32, RPL7, and RPL30 promoters (20). This combination of findings suggests that the Uba80 promoter functions in a manner analogous to the other ribosomal protein genes. Incidentally detected was another 21 bp element (AACTACAGTTCCCAGAATGCA) that is an exact match for a sequence element present in the 5' promoter of the human endothelial nitric oxide synthase gene (NOS3) (30, 31), and in the bi-directional promoter found in the p14.5/hPOP1 gene cluster (32). This element was also noted in two other genomic regions that have been sequenced but not

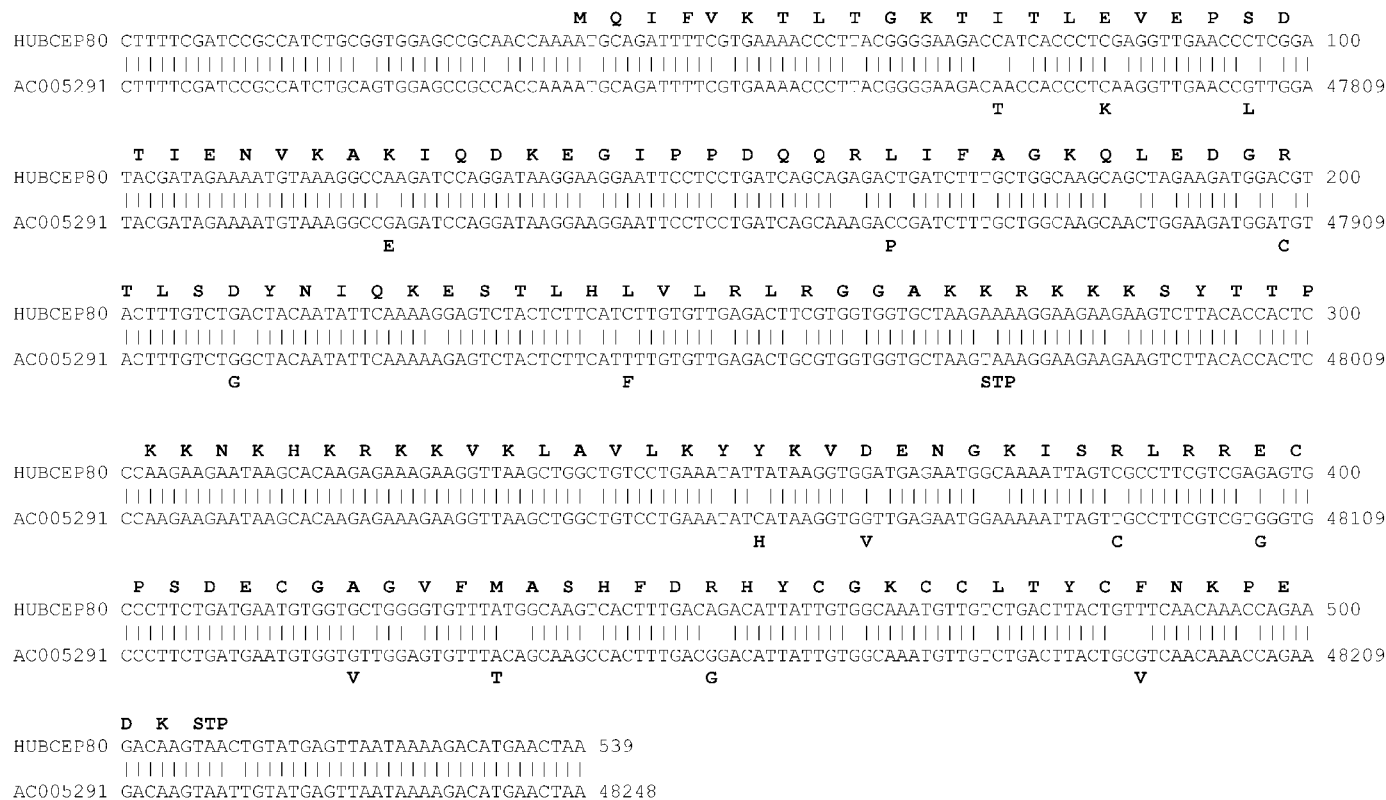


FIG. 3. Comparison of the structure of the cDNA sequence encoded by the bona fide Uba80 gene (HUBCEP80) with that of a pseudogene from chromosome 17 (AC005291). Sequence changes are identified, and changes in predicted amino acid sequences are also shown. Note that the pseudogene encodes a stop codon after amino acid 43.

analyzed, one on chromosome 6p21.3 and one on chromosome 8p21.3-p22 (Fig. 2B). Interestingly, the central portion of this element appears at least partially conserved in a number of other ribosomal protein genes, as indicated in the lower portion of Fig. 2B. The functional significance of this element is unclear, although the fact that it is found in a number of promoters and ribosomal protein genes suggests that it may function as a regulatory element in the control of constitutively active genes.

Lastly, use of the genomic structure of Uba80 for a BLAST search of the non-redundant sequence databases detected the presence of a pseudogene of Uba80 on a BAC from chromosome 17 (Fig. 3). The pseudogene starts exactly at the transcription start site and is homologous to the actual gene coding sequences up to the signal for cleavage and polyadenylation, which lies 12 bp downstream from the AAUAAA signal. Analysis of the sequence of the pseudogene reveals a small number of sequence changes; however, these are enough to disrupt the coding sequence by the introduction of a number of extra stop codons. The presence of a pseudogene of Uba80 was expected from the initial PCR studies described above. This finding was not surprising in light of the fact that the human genome contains many pseudogenes not only for most ribosomal protein

genes (6), but also for the ubiquitin genes (27, 33). Previous analysis of the Uba52 gene had already demonstrated that this gene also is also represented by multiple pseudogenes (34).

In summary, we report here the structure of the gene encoding human Uba80, also known as HUBCEP80 and RPS27a. The entire structure of this gene has been determined, as well as sequences both 5' and 3' of the gene. We show that the gene structure is quite similar to the other human ubiquitin fusion protein, Uba52, although the ubiquitin-coding portion of the gene has been split into an additional exon in this gene. Although 2 ribosomal genes contain TATA and/or CAAT elements in their promoters have been described (10, 11), the large majority of ribosomal gene promoters are similar to that reported here. Specifically, the regulatory elements of these genes consist of a TATA-less promoter within a CpG-rich island, with the transcriptional start site located within a polypyrimidine tract. We also report the structure of one of the Uba80 pseudogenes which was detected by comparison of our sequence to available online databases. Because of Uba80's unusual position as a ubiquitin-ribosomal protein fusion, the data presented here should help future researchers studying the regulation and mechanism of

ribosome synthesis by allowing genomic studies of the gene.

REFERENCES

- Jentsch, S., Seufert, W., and Hauser, H. P. (1991) *Biochim. Biophys. Acta* **1089**(2), 127–139.
- Ozkaynak, E., Finley, D., Solomon, M. J., and Varshavsky, A. (1987) *EMBO J.* **6**(5), 1429–1439.
- Redman, K. L., and Rechsteiner, M. (1988) *J. Biol. Chem.* **263**(10), 4926–4931.
- Redman, K. L., and Rechsteiner, M. (1989) *Nature* **338**(6214), 438–440.
- Webb, G. C., Baker, R. T., Coggan, M., and Board, P. G. (1994) *Genomics* **19**(3), 567–569.
- Kenmochi, N., Kawaguchi, T., Rozen, S., Davis, E., Goodman, N., Hudson, T. J., Tanaka, T., and Page, D. C. (1998) *Genome Res.* **8**(5), 509–523.
- Wiborg, O., Pedersen, M. S., Wind, A., Berglund, L. E., Marcker, K. A., and Vuust, J. (1985) *EMBO J.* **4**(3), 755–759.
- Finley, D., Bartel, B., and Varshavsky, A. (1989) *Nature* **338**(6214), 394–401.
- Seshadri, T., Uzman, J. A., Oshima, J., and Campisi, J. (1993) *J. Biol. Chem.* **268**(25), 18474–18480.
- Rhoads, D. D., Dixit, A., and Roufa, D. J. (1986) *Mol. Cell Biol.* **6**(8), 2774–2783.
- Fan, W., Christensen, M., Eichler, E., Zhang, X., and Lennon, G. (1997) *Genomics* **46**(2), 234–239.
- Annino, T., Laan, M., Stahl, J., and Metspalu, A. (1995) *Gene* **165**(2), 297–302.
- Chen, I. T., and Roufa, D. J. (1988) *Gene* **70**(1), 107–116.
- Colombo, P., Yon, J., and Fried, M. (1991) *Biochim. Biophys. Acta* **1129**(1), 93–95.
- Davies, B., and Fried, M. (1993) *Genomics* **15**(1), 68–75.
- Davies, B., and Fried, M. (1995) *Genomics* **25**(2), 372–380.
- De Falco, S., Russo, G., Angiolillo, A., and Pietropaolo, C. (1993) *Gene* **126**(2), 227–235.
- Espinosa, L., Martin, M., Nicolas, A., Fabre, M., and Navarro, E. (1997) *Biochim. Biophys. Acta* **1354**(1), 58–64.
- Fisher, E. M., Beer-Romero, P., Brown, L. G., Ridley, A., McNeil, J. A., Lawrence, J. B., Willard, H. F., Bieber, F. R., and Page, D. C. (1990) *Cell* **63**(6), 1205–1218.
- Zinn, A. R., Alagappan, R. K., Brown, L. G., Wool, I., and Page, D. C. (1994) *Mol. Cell Biol.* **14**(4), 2485–2492.
- Kirschner, L. S., Taymans, S. E., Pack, S., Pak, E., Pike, B. L., Chandrasekharappa, S. C., Zhuang, Z., and Stratakis, C. A. (1999) *Genomics* **62**(1), 21–33.
- Kirschner, L. S., and Stratakis, C. A. (1999) *BioTechniques* **27**(1), 72–74.
- Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Hudson, T. J., et al. (1996) *Science* **274**(5287), 540–546.
- Taymans, S. E., Kirschner, L. S., Giatzakis, C., and Stratakis, C. A. (1999) *Genomics* **56**, 344–349.
- Stone, E. M., Lotery, A. J., Munier, F. L., Heon, E., Piguet, B., Guymer, R. H., Vandeburgh, K., Cousin, P., Nishimura, D., Swiderski, R. E., Silvestri, G., Mackey, D. A., Hageman, G. S., Bird, A. C., Sheffield, V. C., and Schorderet, D. F. (1999) *Nat. Genet.* **22**(2), 199–202.
- Baker, R. T., and Board, P. G. (1991) *Nucleic Acids Res.* **19**(5), 1035–1040.
- Baker, R. T., and Board, P. G. (1987) *Nucleic Acids Res.* **15**(2), 443–463.
- Lund, P. K., Moats-Staats, B. M., Simmons, J. G., Hoyt, E., D'Ercole, A. J., Martin, F., and Van Wyk, J. J. (1985) *J. Biol. Chem.* **260**(12), 7609–7613.
- Adams, S. M., Sharp, M. G., Walker, R. A., Brammar, W. J., and Varley, J. M. (1992) *Br. J. Cancer* **65**(1), 65–71.
- Nadaud, S., Bonnardeaux, A., Lathrop, M., and Soubrier, F. (1994) *Biochem. Biophys. Res. Commun.* **198**(3), 1027–1033.
- Marsden, P. A., Heng, H. H., Scherer, S. W., Stewart, R. J., Hall, A. V., Shi, X. M., Tsui, L. C., and Schappert, K. T. (1993) *J. Biol. Chem.* **268**(23), 17478–17488.
- Schmiedeknecht, G., Buchler, C., and Schmitz, G. (1997) *Biochem. Biophys. Res. Commun.* **241**(1), 59–67.
- Baker, R. T., and Board, P. G. (1987) *Nucleic Acids Res.* **15**(10), 4352.
- Baker, R. T., and Board, P. G. (1992) *Genomics* **14**(2), 520–522.