MOLECULAR AND CELLULAR BIOLOGY, Dec. 1985, p. 3337–3344
0270-7306/85/123337-08$02.00/0

Vol. 5, No. 12

Nucleotide Sequence and Expression In Vitro of cDNA Derived from mRNA of *int*-1, a Provirally Activated Mouse Mammary Oncogene

Y.-K. T. FUNG,† G. M. SHACKLEFORD, A. M. C. BROWN, G. S. SANDERS, AND H. E. VARMUS*

Department of Microbiology and Immunology, University of California, San Francisco, California 94143

Received 18 July 1985/Accepted 13 September 1985

The mouse *int*-1 gene is a putative mammary oncogene discovered as a target for transcriptionally activating proviral insertion mutations in mammary carcinomas induced by the mouse mammary tumor virus in C3H mice. We have isolated molecular clones of full- or nearly full-length cDNA transcribed from *int*-1 RNA (2.6 kilobases) in a virus-induced mammary tumor. Comparison of the nucleotide sequence of the cDNA clones with that of the *int*-1 gene (A. van Ooyen and R. Nusse, Cell 39:233–240, 1984) shows the following. (i) The coding region of the *int*-1 gene is composed of four exons. (ii) The splice donor and acceptor sites conform to consensus; however, at least two closely spaced polyadenylation sites are used, and the transcriptional initiation site remains ambiguous. (iii) The major open reading frame is preceded by an open frame 10 codons in length. (iv) The mRNA encodes a 41-kilodalton protein with several striking features—a strongly hydrophobic amino terminus, a cysteine-rich carboxy terminus, and four potential glycosylation sites. (v) There are no differences in nucleotide sequence between the known exons of the normal and a provirally activated allele. The length of the deduced open reading frame was further confirmed by in vitro translation of RNA transcribed from the cDNA clones with SP6 RNA polymerase.

Over the past decade, at least three strategies have been developed for the identification of cellular oncogenes and proto-oncogenes. These include physical detection of cellular progenitors of retroviral oncogenes, functional tests for mutant cellular genes competent to transform cultured cells, and structural analysis of cellular genes repeatedly rearranged in tumor cells as a result of insertion mutations, DNA amplification, or chromosomal translocations (for a review, see references 4 and 33). The mouse *int*-1 gene was among the first to be identified as a putative oncogene solely on the basis of genetic rearrangement and consequent effects upon expression. About three-fourths of mammary carcinomas induced in C3H mice by milk-borne mouse mammary tumor virus (MMTV) harbor proviral insertion mutations within a 30-kilobase (kb) region (21). The proviruses reside on either the 5′ or 3′ side of the *int*-1 transcriptional unit, usually oriented so that transcription of the provirus proceeds in the direction away from *int*-1 (22). These insertions are accompanied by the production of ca. 1 to 10 copies of *int*-1 mRNA per cell. Since the gene is silent in normal mammary tissue, it is likely that the proviral insertions are responsible for activating expression, presumably by an enhancer-like mechanism, and thereby contribute to mammary oncogenesis.

Since *int*-1 has not been previously encountered as the progenitor of a retroviral oncogene and has not yet been shown to have neoplastic effects when introduced into cultured cells, the claim for its status as a cellular oncogene rests principally upon the frequency of transcriptionally activating MMTV insertion mutations of *int*-1 in mouse mammary tumors. To advance our understanding of the *int*-1 gene, we have undertaken the molecular cloning and nucleotide sequencing of cDNA synthesized from *int*-1 mRNA in an MMTV-induced mammary tumor. Comparison of our results with the recently published nucleotide sequence of the normal *int*-1 gene (31) establishes the organization of introns and exons within this locus, confirms the existence of a single open reading frame competent to encode a protein of 370 amino acids, and demonstrates that insertional activation of *int*-1 need not be accompanied by nucleotide substitutions in the coding sequence. We have further validated the deduced open reading frame by synthesis of *int*-1 protein in vitro from SP6 RNA polymerase-generated transcripts of two of our cDNA clones.

MATERIALS AND METHODS

Tumor-derived nucleic acids. Mammary tumor 103 arose in a multiparous BALB/c mouse that had been foster-nursed by a C3H female in a colony known to transmit MMTV(C3H) in its milk. The affected animal was kindly provided by S. Nandi (Cancer Research Institute, University of California, Berkeley). The tumor was successively transplanted to BALB/c females, and the resulting tumors served as a source of nucleic acid for subsequent studies. A transplantable tumor from a GR mouse, used in initial studies, was obtained from the Mason Research Institute (Worcester, Mass.). Cellular DNA was prepared by previously described methods (8), and poly(A)$^+$ whole cell RNA was isolated as follows. Chilled tumor tissue was minced with a razor blade and homogenized in 5 M guanidinium thiocyanate in the presence of 20 mM Tris hydrochloride (pH 8)–10 mM vanadyl ribonucleoside complex–2% Sarkosyl–5% β-mercaptoethanol. The homogenate was then centrifuged through 3 ml of 5.7 M CsCl in an SW41 rotor at 30,000 rpm for 16 h at 15°C. The RNA pellet was resuspended in 10 mM vanadyl ribonucleoside complex, and poly(A)$^+$ RNA was selected by chromatography on oligo(dT)-cellulose (P-L Biochemicals) (1, 7).

* Corresponding author.
† Present address: Division of Hematology and Oncology, Department of Pediatrics and Microbiology, University of Southern California School of Medicine, Los Angeles, CA 90027.

Synthesis and molecular cloning of cDNA. The following steps were performed according to the detailed, unpublished protocol provided by Thanh Huynh and Tom St. John (Department of Pathology, Stanford University, Stanford, Calif.), with minor modifications and λgt10 (λ*imm*^{434} *b527*) as a cloning vector. DNA complementary to poly(A)$^+$ RNA from tumor 103 was synthesized as follows. A 10-μg quantity of poly(A)$^+$ RNA was treated with 2 mM methyl mecuric hydroxide at room temperature for 5 min and then neutralized with 1 μl of 1.42 M β-mercaptoethanol. To this reaction mixture, all four deoxyribonucleotides and vanadyl ribonucleoside complex were added to final concentrations of 1 mM. A 10-μg quantity of oligo(dT)$_{12-18}$ (P-L Biochemicals) was used as primer, the reaction mixture was adjusted to 50 mM Tris hydrochloride–(pH 8.8)–40 mM KCl–10 mM $MgCl_2$, and the reaction was started by the addition of 100 U of avian myeloblastosis virus reverse transcriptase (Seikagaku, Inc.). A small sample of the reaction was removed and incubated with 10 μCi of [α-^{32}P]dGTP to monitor incorporation. The reaction was allowed to proceed at 46°C for 1 h. After the reaction, 1% of the radioactively labeled cDNA was used to monitor the incorporation by trichloroacetic acid precipitation, and the remainder was subjected to gel electrophoresis to determine the size of the cDNA synthesized. Terminal deoxynucleotidyl transferase (Ratliff Biochemicals) was then used to add approximately 20 residues of dGMP to the 3′ termini of the cDNA-RNA hybrids. RNA was removed by first boiling the cDNA-RNA hybrid for 1 min and then digesting with 5 μg of RNase A. Second strands were synthesized by *Escherichia coli* DNA polymerase I (Boeringer Mannheim), with oligo(dC) (Collaborative Research) as primer. The resulting double-stranded DNA was then methylated with *Eco*RI methylase, *Eco*RI dodecameric linkers were joined to the ends with T4 DNA ligase, and the preparation was digested with *Eco*RI endonuclease (all reagents from New England BioLabs). Molecules approximately 1 kb in length or longer were selected by preparative electrophoresis in a 1.2% agarose gel and ligated to the *Eco*RI cohesive ends of λgt10 DNA. Approximately 2 × 10^7 PFU of phage were obtained per μg of double-stranded DNA after in vitro packaging as described previously (17). The recombinant phage were amplified by the plate lysis procedure, and the amplified library was screened for clones containing *int*-1-specific inserts by using ^{32}P-labeled pMT2.5 (*int*-1 clone C in references 21 and 22) as previously described (2).

Nucleotide sequencing. Fragments of cDNA clones generated by restriction endonuclease digestion or *Bal* 31 deletion were subcloned into M13 vectors for nucleotide sequencing according to standard procedures (17). Sequence analysis was accomplished primarily by the dideoxynucleotide chain termination method (25, 26), with ^{35}S-labeled dATP (Amersham) and buffer gradient gels (3). Some fragments were also sequenced by the chemical cleavage method of Maxam and Gilbert (19). Sequence ambiguities caused by secondary structure in one region (nucleotides 1175 to 1250) were alleviated by the inclusion of 25% formamide in the gel buffer. All *Hae*III and *Hha*I restriction enzyme sites in this region were confirmed by autoradiography of a polyacrylamide gel displaying partial enzymatic digests of an end-labeled fragment encompassing this region (28).

In vitro expression of cDNA clones. The cDNA inserts of λgt10 *int*-1 clones 2 and 26 were subcloned as *Eco*RI fragments into the plasmid pSP65 (Promega Biotech) downstream of the SP6 promoter. Each recombinant plasmid was linearized by digestion with *Mlu*I and transcribed in vitro with SP6 polymerase (Promega) under the conditions described by Krieg and Melton (14). Approximately 500 ng of the resulting RNA was then translated in vitro in 25-μl reactions containing 17.5 μl of rabbit reticulocyte lysate (Promega), 25 μCi of [^{35}S]methionine (Amersham; 1,300 Ci/mmol), and a mixture of unlabeled amino acids (Promega). Two-microliter samples of these translation reactions were then run on 12% polyacrylamide-sodium dodecyl sulfate gels prepared as described previously (14). After treatment with fixative, the gels were soaked in Amplify (Amersham), dried, and exposed to Kodak XAR-5 film for 1 to 12 h.

RESULTS

Selection of a mammary tumor as a source of *int*-1 RNA. Since *int*-1 RNA has thus far only been described in MMTV-induced mammary tumors, it was obligatory to use tumor RNA as a template for synthesis of *int*-1 cDNA. In most mammary tumors with *int*-1 insertion mutations, *int*-1 mRNA is approximately 2.6 kb in length, unlinked to any viral sequences, and presumably generated from normal signals for transcription and RNA processing, in response to an enhancer element in the MMTV provirus (22, 23). However, in some cases, longer transcripts, up to 3.5 kb in length, result from insertions near the 3′ end of the gene that preclude use of the normal *int*-1 polyadenylation site(s) and demand transcription of the long U3 region of the MMTV long terminal repeat (22). In exceptional cases, proviral DNA is positioned close to the 5′ end of the *int*-1 gene, in the same transcriptional orientation, so that it provides a promoter for *int*-1 and donates sequences (R and U5) to the resulting *int*-1 transcripts (31; R. Nusse, personal communication).

To avoid cloning cDNA from such hybrid transcripts, we used previously described strategies (22) to identify tumor 103 from a BALB/cfC3H mouse that contained a single MMTV(C3H) provirus positioned approximately 5 to 7 kb on the 5′ side of the regions of the *int*-1 locus known to be represented in *int*-1 RNA (data not shown). As expected, poly(A)$^+$ RNA from this tumor contained a single, low-abundance species of ca. 2.6 kb that annealed to a labeled subclone from the *int*-1 locus (Fig. 1). RNA prepared in parallel from two other tumors contained either less *int*-1 RNA of 2.6 kb or larger *int*-1 RNA indicative of hybrid transcripts. Tumor 103 offered two additional advantages. It can be readily transplanted into new BALB/c mice without the loss of expression of *int*-1, providing a renewable source of *int*-1 mRNA (unpublished data), and it arose in the same mouse strain (BALB/c) used as a source of normal embryo DNA for cloning and sequencing the *int*-1 gene. Thus, comparisons of the *int*-1 gene in normal and tumor tissue could be made without concern for strain-dependent polymorphisms. Some additional cDNA clones were prepared from a GR mammary tumor that also contained 2.6-kb *int*-1 mRNA.

Isolation and preliminary characterization of *int*-1 cDNA clones. Using the cloning procedure described in Materials and Methods, we obtained several recombinant λgt10 phages bearing inserts with homology to pMT2.5, a subclone of the *int*-1 gene known to hybridize to *int*-1 mRNA in tumor samples (21, 22). Most of these clones contained relatively short inserts (less than 1 kb in length); however, from one library of size-selected cDNAs transcribed from tumor 103 RNA, we isolated 2 clones (from 12 *int*-1-positive clones) with inserts of approximately 2.2 kb in length. The organi-

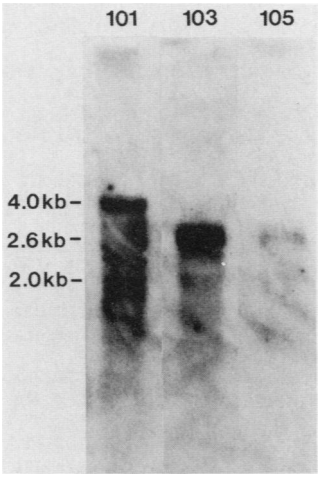


FIG. 1. *int*-1 RNA in mammary tumor 103. Poly(A)$^+$ RNA was prepared from three mammary tumors (101, 103, and 105) arising in BALB/cfC3H mice, and 5 μg was analyzed by agarose gel electrophoresis in 2.2 M formaldehyde, transfer to nitrocellulose, and hybridization with ^{32}P-labeled *int*-1 DNA (pMT2.5) (21, 22). Sizes were determined by reference to rRNA detected by staining the gel with ethidium bromide. RNA from tumor 103 was used as template for synthesis of cDNA for cloning.

zation of these two long clones, numbered 2 and 26, and of two shorter clones from a GR tumor is shown in relation to the deduced structure of the *int*-1 gene and its mRNA (Fig. 2).

Nucleotide sequence of *int*-1 cDNAs. A combination of dideoxynucleotide and Maxam-Gilbert sequencing techniques was used to derive a nucleotide sequence for the *int*-1 domains represented in the several cDNA clones (Fig. 3B). An annotated version of the amalgamated sequence is presented in Fig. 3A, and important features with respect to coding potential are summarized in Fig. 4.

(i) Potential reading frames. The longest open reading frame is present in frame 1, as drawn in Fig. 4A, and it contains a methionine codon near its 5′ end. Since a stop codon is present in the same frame 117 nucleotides further upstream, it is likely that the ATG at position 185 is the initiation codon for the synthesis of *int*-1 protein. This conjecture is further supported by the sequence immediately surrounding the ATG. The proposed start site, AGGC CATGG, conforms closely to the start site consensus deduced by Kozak (11, 13), i.e., ccA/GccATGG (with capitalized letters strongly favored).

The initiation codon at position 185 is not, however, the closest to the 5′ end of *int*-1 mRNA; clone 2 contains an ATG in the same reading frame at position 38, 30 nucleotides upstream of the termination codon that precedes the long open frame. Thus, a decapeptide could also be synthesized from *int*-1 mRNA in the bicistronic manner proposed for certain other eucaryotic mRNAs (5, 9, 10, 12, 16, 18); however, the context of the ATG at position 38 is much less favorable for translational initiation than the context of the ATG at position 185, and we have no direct evidence for use of the short 5′ open reading frame. The only other lengthy region that lacks termination codons (in frame 2) also lacks any initiation codons.

(ii) Properties of the predicted *int*-1 protein. The deduced amino acid sequence of the *int*-1 protein (Fig. 3A) is identical to that deduced from the sequence of the *int*-1 gene, in conjunction with S1 mapping of exons, by van Ooyen and Nusse (31). The molecular weight of *int*-1 protein is calculated to be 41,185.5. Although the overall amino acid composition is not particularly biased, the sequence is notably rich in cysteines in the carboxy-terminal domain (11 of the

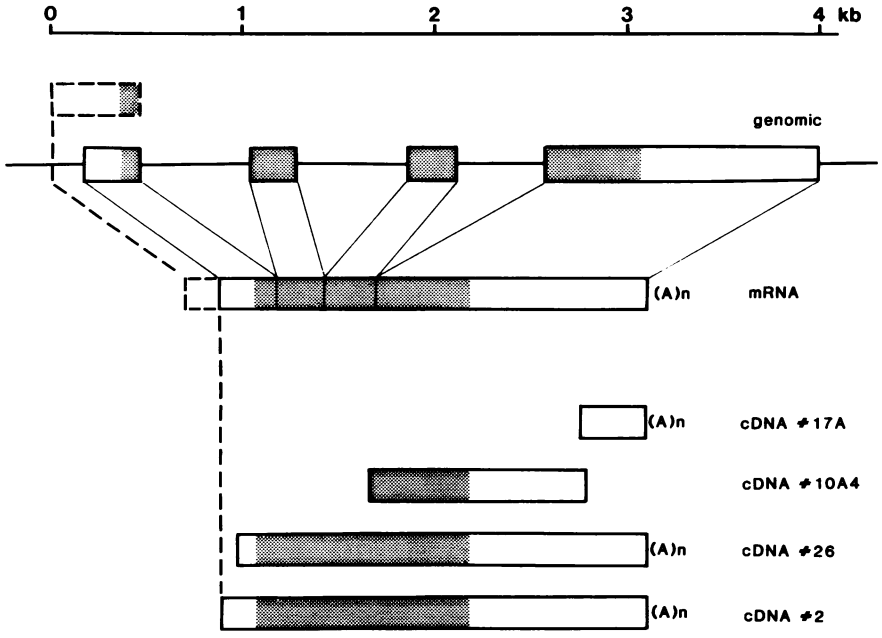
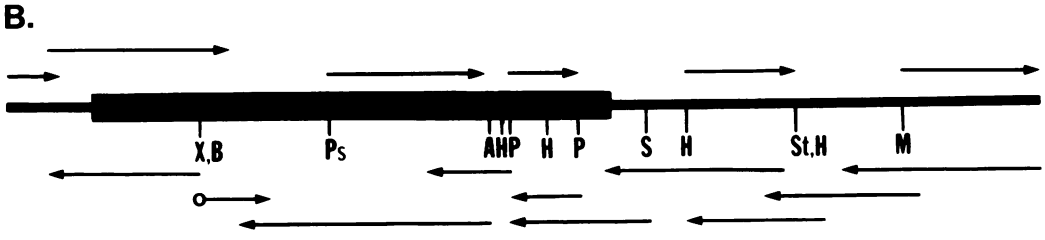


FIG. 2. The structure of *int*-1 cDNA clones in relation to the proposed organization of the mouse *int*-1 gene and its mRNA. Boxed areas represent exon regions; their coding regions are shaded. A possible alternate first exon (31) is outlined in dashes. The four cDNA clones used for sequence determination are shown; clones 2 and 26 are derived from the BALB/cf/C3H tumor 103, and clones 10A4 and 17A are derived from a GR tumor.

A.

```
                (G)AGTGGCTGCTTCAGCCCAGCAGCCAGGACAGCGAACCATGCTGCCTGCGGCCCGCCTCCAGACTTATTAGAGCCAGCCTGGG   82

AACTCGCATCACTGCCCTCACCGCTGTGTCCAGTCCCACCGTCGCGGACAGCAACCACAGTCGTCAGAACCGCAGCACAGAACCAGCAAGGCCAGGCAG   181

GCC ATG GGG CTC TGG GCG CTG CTG CCC AGC TGG GTT TCT ACT ACG TTG CTA CTG GCA CTG ACC GCT CTG CCC GCA   256
    Met Gly Leu Trp Ala Leu Leu Pro Ser Trp Val Ser Thr Thr Leu Leu Leu Ala Leu Thr Ala Leu Pro Ala
                                   ▼ splice
GCC CTG GCT GCC AAC AGT AGT GGC CGA TGG TGG GGC ATC GTG AAC ATA GCC TCC TCC ACG AAC CTG TTG ACG GAT   331
25  Ala Leu Ala Ala Asn Ser Ser Gly Arg Trp Trp Gly Ile Val Asn Ile Ala Ser Ser Thr Asn Leu Leu Thr Asp

TCC AAG AGT CTG CAG CTG GTG CTC GAG CCC AGT CTG CAG CTG CTG AGC CGC AAG CAG CGG CGA CTG ATC CGA CAG   406
50  Ser Lys Ser Leu Gln Leu Val Leu Glu Pro Ser Leu Gln Leu Leu Ser Arg Lys Gln Arg Arg Leu Ile Arg Gln

AAC CCG GGG ATC CTG CAC AGC GTG AGT GGA GGG CTC CAG AGC GCT GTG CGA GAG TGC AAA TGG CAA TTC CGA AAC   481
75  Asn Pro Gly Ile Leu His Ser Val Ser Gly Gly Leu Gln Ser Ala Val Arg Glu Cys Lys Trp Gln Phe Arg Asn
                                                                  ▼ splice
CGC CGC TGG AAC TGC CCC ACT GCT CCG GGG CCC CAC CTC TTC GGC AAG ATC GTC AAC CGA GGC TGC CGA GAA ACA   556
100 Arg Arg Trp Asn Cys Pro Thr Ala Pro Gly Pro His Leu Phe Gly Lys Ile Val Asn Arg Gly Cys Arg Glu Thr

GCG TTC ATC TTC GCA ATC ACC TCC GCC GGG GTC ACA CAT TCC GTG GCG CGC TCC TGC TCC GAA GGC TCC ATC GAG   631
125 Ala Phe Ile Phe Ala Ile Thr Ser Ala Gly Val Thr His Ser Val Ala Arg Ser Cys Ser Glu Gly Ser Ile Glu

TCC TGC ACC TGC GAC TAC CGG CGG CGC GGC CCT GGG GGC CCC GAC TGG CAC TGG GGG GGC TGC AGT GAC AAC ATC   706
150 Ser Cys Thr Cys Asp Tyr Arg Arg Arg Gly Pro Gly Gly Pro Asp Trp His Trp Gly Gly Cys Ser Asp Asn Ile

GAT TTT GGT CGC CTC TTT GGC CGA GAG TTC GTG GAC TCC GGG GAG AAG GGG CGG GAC CTA CGC TTC CTC ATG AAC   781
175 Asp Phe Gly Arg Leu Phe Gly Arg Glu Phe Val Asp Ser Gly Glu Lys Gly Arg Asp Leu Arg Phe Leu Met Asn
                                  ▼ splice
CTT CAC AAC AAC GAG GCA GGG CGA ACG ACC GTG TTC TCT GAG ATG CGC CAA GAG TGC AAA TGC CAC GGG ATG TCC   856
200 Leu His Asn Asn Glu Ala Gly Arg Thr Thr Val Phe Ser Glu Met Arg Gln Glu Cys Lys Cys His Gly Met Ser

GGC TCC TGC ACG GTG CGC ACG TGT TGG ATG CGG CTG CCC ACG CTG CGC GCT GTG GGC GAC GTG CTG CGC GAC CGC   931
225 Gly Ser Cys Thr Val Arg Thr Cys Trp Met Arg Leu Pro Thr Leu Arg Ala Gly Val Asp Val Leu Arg Asp Arg

TTC GAC GGC GCC TCC CGC GTC CTT TAC GGC AAC CGA GGC AGC AAC CGC GCC TCG CGG GCG GAG CTG CTG CGC CTG   1,006
250 Phe Asp Gly Ala Ser Arg Val Leu Tyr Gly Asn Arg Gly Ser Asn Arg Ala Ser Arg Ala Glu Leu Leu Arg Leu

GAG CCC GAA GAC CCC GCG CAC AAG CCT CCC TCC CCT CAC GAC CTC GTC TAC TTC GAG AAA TCG CCC AAC TTC TGC   1,081
275 Glu Pro Glu Asp Pro Ala His Lys Pro Pro Ser Pro His Asp Leu Val Tyr Phe Glu Lys Ser Pro Asn Phe Cys

ACG TAC AGT GGC CGC CTG GGC ACA GCT GGC ACA GCT GGA CGA GCT TGC AAC AGC TCG TCT CCC GCG CTG GAC GGC   1,156
300 Thr Tyr Ser Gly Arg Leu Gly Thr Ala Gly Thr Ala Gly Arg Ala Cys Asn Ser Ser Ser Pro Ala Leu Asp Gly

TGT GAG CTG CTG TGC TGT GGC CGA GGC CAC CGC ACG CGC ACG CAG CGC GTC ACG GAG CGC TGC AAC TGC ACC TTC   1,231
325 Cys Glu Leu Leu Cys Cys Gly Arg Gly His Arg Thr Arg Thr Gln Arg Val Thr Glu Arg Cys Asn Cys Thr Phe

CAC TGG TGC TGC CAC GTC AGC TGC CGC AAC TGC ACG CAC ACG CGC GTT CTG CAC GAG TGT CTA TGA GGTGCCGCGCC   1,308
350 His Trp Cys Cys His Val Ser Cys Arg Asn Cys Thr His Thr Arg Val Leu His Glu Cys Leu ***

TCCGGGAACGGGAACGCTCTCTTCCAGTTCTCAGACACACTCGCTGGTCCTGATGTTTGCCCACCCTACCGCGTCCAGCCACAGTCCCAGGGTTCATAG   1,407

CGATCCATCTCTCCCACCTCCTACCTGGGGACTCCTGAAACCACTTGCCTGAGTCGGCTCGAACCCTTTTGCCATCCTGAGGGCCCTGACCCAGCCTAC   1,506

CTCCCTCCCTCTTTGAGGGAGACTCCTTTTGCACTGCCCCCCAATTTGGCCAGAGGGTGAGAGAAAGATTCTTCTTCTGGGGTGGGGGTGGGGAGGTCA   1,605

ACTCTTGAAGGTGTTGCGGTTCCTGATGTATTTTGCGCTGTGACCTCTTTGGGTATTATCACCTTTCCTTGTCTCTCGGGTCCCTATAGGTCCCTTGAG   1,704

TTCTCTAACCAGCACCTCTGGGCTTCAAGGCCTTTCCCCTCCCACCTGTAGCTGAAGAGTTTCCGAGTTGAAAGGGCACGGAAAGCTAAGTGGGAAAGG   1,803

AGGTTGCTGGACCCAGCAGCAAAACCCTACATTCTCCTTGTCTCTGCCTCGGAGCCATTGAACAGCTGTGAACCATGCCTCCCTCAGCCTCCTCCCACC   1,902

CCTTCCTGTCCTGCCTCCTCATCACTGTGTAAATAATTTGCACCGAAATGTGGCCGCAGAGCCACGCGTTCGGTTATGTAAATAAAACTATTTATTGTG   2,001

CTGGGTTCCAGCCTGGGTTGCAGAGACCACCCTCACCCCACCTCACTGCTCCTCTGTTCTGCTCGCCAGTCCTTTTGTTATCCGACCTTTTTTCTCTTT   2,100

TACCCAGCTTCTCATAGGCGCCCTTGCCCACCGGATCAGTATTTCCTTCCACTGTAGCTATTAGTGGCTCCTCGCCCCCACCAATGTAGTATCTTCCTC   2,199

TGAGGAATAA                                                                                              2,230
```

B.



final 56 amino acids are cysteines) (Fig. 4C). A plot of hydropathicity values determined by the method of Kyte and Doolittle (15) reveals that the first 46 amino acids form a strongly hydrophobic region; in general, the remainder of the protein is rather hydrophilic (Fig. 4B). Since the amino terminus seems likely to affiliate with membranes and could constitute a signal sequence characteristic of a secreted protein, we also examined the amino acid sequence for potential sites for N-linked glycosylations. Four were found, three within the cysteine-rich region and one near the amino terminus (Fig. 3A and 4C).

(iii) Signals for RNA processing. Comparison of the genomic and cDNA sequences of *int*-1 allows the unambiguous definition of three introns and hence three sets of splice donor and acceptor sites. The sequences at these sites conform to published consensus sequences (20), and the sites are identical to those predicted by van Ooyen and Nusse on the basis of S1 nuclease mapping of mRNA (31). At least two of our cDNA clones bear tracts of oligo(dT) at different positions in the *int*-1 sequence, i.e., at the positions of CA and GA found 16 and 20 nucleotides downstream from the consensus polyadenylation signals, AATAAA (23). This implies at least two nearby sites are polyadenylated during the processing of *int*-1 mRNA.

(iv) 5′ and 3′ noncoding regions. The final exon of *int*-1 is the longest (1419 to 1422 base pairs), and all but 486 nucleotides follow the translation termination site. Thus, almost half of the mRNA is composed of a 3′ noncoding sequence of uncertain function. At least 184 nucleotides precede the start of the designated *int*-1 coding region. However, neither our study nor the report of van Ooyen and Nusse (31) clarifies the 5′ boundary of the first coding exon. S1 mapping of *int*-1 mRNA suggests the existence of two 5′ ends to this exon (Fig. 2), but the genomic sequences in the vicinity of the ends are compatible with splice acceptor sites or with transcriptional start sites (31). One of our cDNA clones, clone 2, ends very near one of the two sites judged to represent a 5′ boundary for the exon, suggesting that the cDNA may be a complete copy of a transcript initiated from this site, which is 33 base pairs downstream from the sequence TATAAGA in the genomic clone. An alternative interpretation is implied by the first nucleotide in the cDNA clone following the homopolymeric deoxycytidylic acid tract. The G residue (shown in parentheses in Fig. 3A) is not predicted from the genomic sequence in this region, and it could be derived from an unidentified upstream exon that ends with GAG. The upstream exon could be joined to the sequence that follows GAG in the cDNA clone by splicing, mediated by the possible splice acceptor site that is apparent in the corresponding genomic sequence upstream of position 3 in the cDNA sequence. Perhaps more likely, the unexplained G residue may represent an erroneous addition near the 5′ end of *int*-1 mRNA (e.g., at the position of a cap nucleotide).

(v) Comparison of normal and provirally activated alleles. We found no differences other than the ambiguous nucleotide mentioned above between the sequences of our cDNA clones and the exonic regions of the BALB/c *int*-1 gene sequenced by van Ooyen and Nusse (31). This shows that proviral activation of *int*-1 is not necessarily accompanied by mutations in the *int*-1 exons.

Synthesis of *int*-1 protein in vitro. To confirm the presence of the deduced open reading frame for *int*-1 protein in cDNA clones 2 and 26, the entire insert was subcloned in the plasmid vector pSP65, downstream from a recognition site for the RNA polymerase of bacteriophage SP6, in an orientation intended to generate a transcript with the same chemical polarity as *int*-1 mRNA. The RNA synthesized in vitro by the SP6 polymerase was then translated in a rabbit reticulocyte lysate in the presence of [^{35}S]methionine, producing a single major species of labeled polypeptide that migrated in sodium dodecyl sulfate-polyacrylamide gels with an apparent molecular mass of approximately 37 kilodaltons, close to the theoretical molecular mass (41 kilodaltons) predicted from the nucleotide sequence (Fig. 5). The protein was not synthesized from RNA transcribed from a clone with *int*-1 cDNA in the opposite orientation (Fig. 5, lane C). The identity of the labeled protein has been further substantiated by immunoprecipitation with sera from rabbits immunized with peptides whose synthesis was based upon the sequence of *int*-1 clones (unpublished data).

DISCUSSION

The *int*-1 gene. The results reported here, in conjunction with the recent studies by van Ooyen and Nusse (31), show that the *int*-1 gene contains four coding exons. The final exon is much larger than the others, and it contributes ca. 935 nucleotides of 3′ noncoding sequence to *int*-1 mRNA. The end of this exon is variable, since we have found cDNA clones copied from mRNAs that appear to be poly(A)$^+$ at two nearby but different sites. The 5′ end of the *int*-1 gene remains less well defined. There may be variable 5′ boundaries to the first coding exon, the transcriptional initiation site(s) has not been identified, and the possibility of an additional 5′ (noncoding) exon has not been excluded. Furthermore, it is important to recall that all attempts to describe transcription of the *int*-1 gene have been confined to RNA from tumors in which the gene is transcriptionally activated by nearby insertion mutations. It is possible that transcription is initiated at different sites in those still unidentified cells in which the gene is normally expressed.

Translational initiation also appears to be unusual. The only long open reading frame with a favorable initiation site

FIG. 3. The sequence of *int*-1 cDNA. (A) The nucleotide sequence of *int*-1 cDNA with deduced amino acid sequence of *int*-1 protein. The combined nucleotide sequences of the cDNA clones shown in Fig. 2 are presented as the sequence of the longest cDNA, clone 2. The numbering of nucleotides is indicated at the right and begins with the first base in an extended sequence identical to the published sequence of the *int*-1 gene (31); a single G residue between the synthetic poly(C) tract and the *int*-1 sequence is placed in parentheses and discussed in the text. Splice junctions are indicated by large vertical arrowheads, polyadenylation sites are indicated by small vertical arrowheads, and a putative polyadenylation signal is indicated by dashed underlining. The deduced amino acid sequence is provided from the first ATG in the long open reading frame (nucleotides 185 to 187) and numbered on the left-hand margin; a preceding stop codon in the same frame is overlined (positions 68 to 70). Four potential glycosylation sites are underlined. (B) The sequencing strategy. Arrows indicate direction of sequencing from various restriction endonuclease sites. Abbreviations: X, *Xho* I; B, *Bam*HI; Ps, *Pst*I; A, *Acc*I; H, *Hin*dIII; P, *Pvu*II; S, *Sac*I; St, *Stu*I; M, *Mlu*I. All determinations were done by the dideoxy method, save one with an open circle at the start of the arrow. The latter was performed by chemical methods and 5′-end labeling (see Materials and Methods). At least one strand of clone 2 or 26 was sequenced throughout, and the data were supplemented with results from clones 17A and 10A4 in the 3′ half of the sequence.

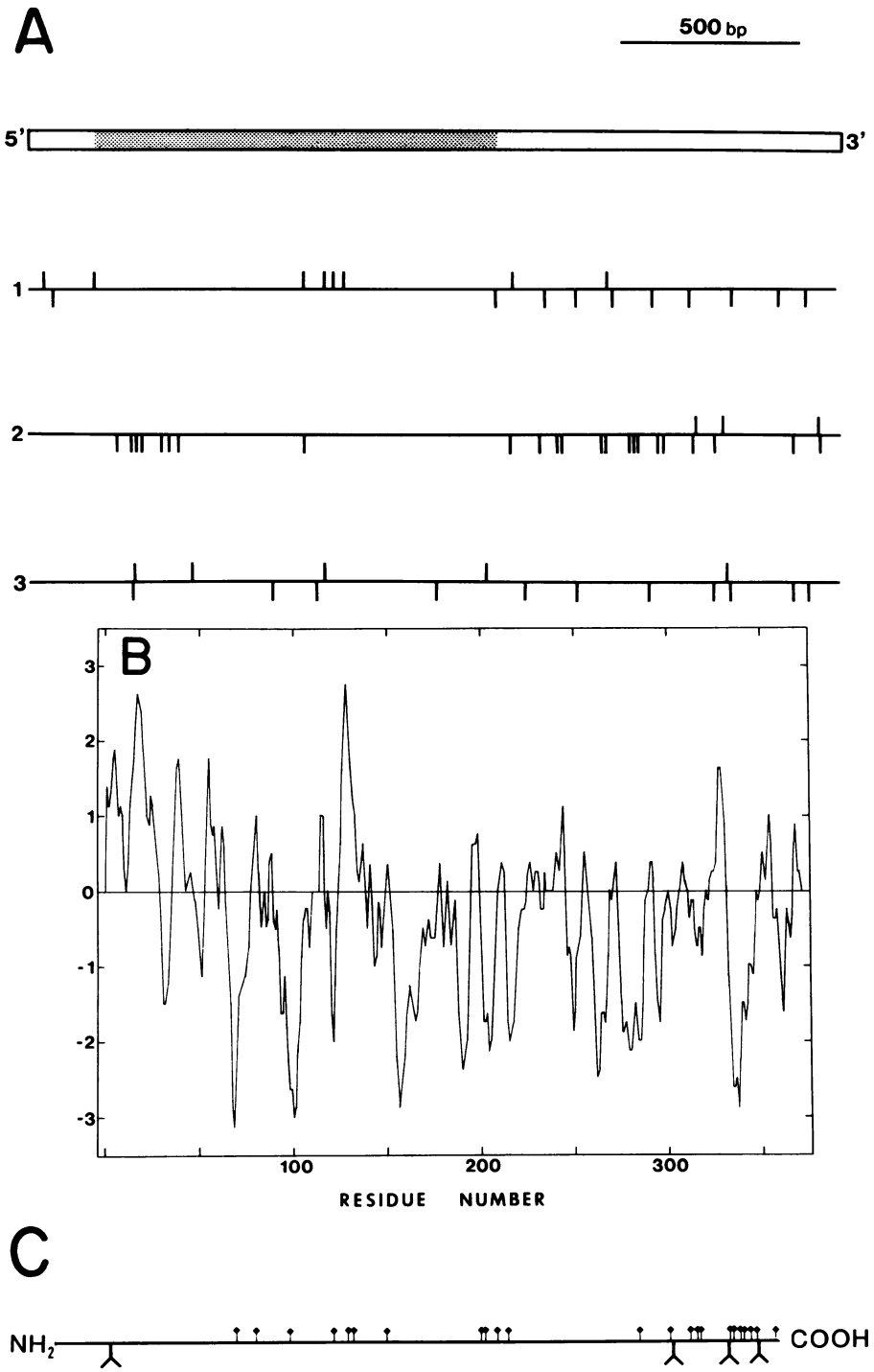


FIG. 4. The coding potential of *int*-1. (A) Initiation and termination codons in the three reading frames of *int*-1 cDNA. The box at the top represents cDNA clone 2, the proposed coding region of which is shaded. The three reading frames are displayed with their respective initiation codons (vertical bars above the line) and termination codons (vertical bars below the line). Reading frame 1 contains the proposed *int*-1 coding region, and reading frames 2 and 3 represent the +1 and −1 reading frames, respectively. (B) Hydropathicity plot of the deduced *int*-1 protein. Hydropathicity values were calculated as described by Kyte and Doolittle (15) with the Bionet computer program. Hydrophobic values are positive. (C) Positions of cysteine residues (◆) and potential N-linked glycosylation sites (forks below line) in the predicted *int*-1 protein.

(13) is preceded by a short open reading frame containing an initiation codon (Fig. 3A). Thus, *int*-1 mRNA joins the list of exceptions to the proposal that translation of eucaryotic mRNAs initiates at the AUG nearest the 5′ end (11, 12).

Although nothing is known about the normal function of the *int*-1 gene, the sequence and organization of the gene are well conserved. The four exons discussed here have been displayed in the human homolog of *int*-1 by heteroduplex

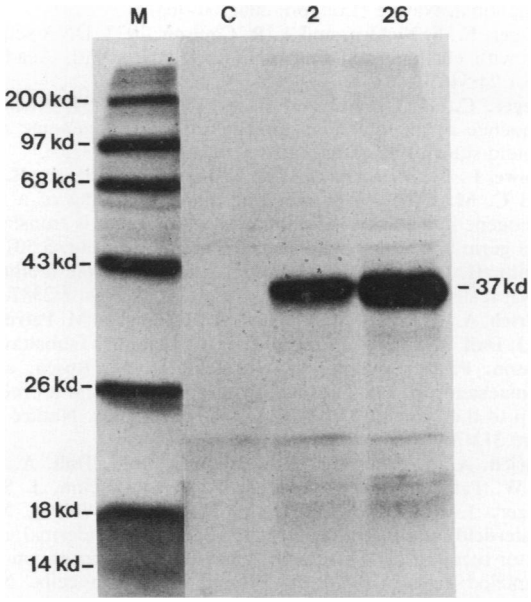


FIG. 5. In vitro synthesis of *int*-1 protein. *int*-1 cDNA clones in pSP65 were linearized by digestion with *Mlu*I and transcribed in vitro with SP6 polymerase, and the resulting RNAs were translated in rabbit reticulocyte lysates in the presence of [^{35}S]methionine. The labeled proteins were electrophoresed on a 12% polyacrylamide gel and revealed by fluorography. Lanes 2 and 26, Translation products of RNAs transcribed from *int*-1 cDNAs 2 and 26, respectively. The position of the major band at M_r 37,000 is indicated. Lane C (control), Translation of RNA derived from a pSP65 clone containing *int*-1 cDNA 26 in reverse orientation relative to the SP6 promoter. Lane M, Molecular mass markers.

mapping (32). More strikingly, the open reading frame of the human *int*-1 gene encodes a protein that differs from the mouse *int*-1 protein at only 4 of 370 residues (R. Nusse, personal communication). Sequences related to the mouse *int*-1 gene have been detected by hybridization in genomes of other mammals, birds, and insects (22); however, further work will be required to determine the similarity of the hybridizing sequences to the *int*-1 locus examined here.

Activation of the *int*-1 proto-oncogene. The *int*-1 gene was discovered during a search for cellular genes affected by MMTV insertion mutations in virus-induced mammary cancers (21). The primary effect of the insertions appears to be conversion of the *int*-1 locus from a silent gene in normal mammary tissue to one expressed at low levels (ca. 10 copies of mRNA per cell). The influence of *int*-1 insertion mutations presumably requires synthesis of *int*-1 protein, since van Ooyen and Nusse (31) have shown that the insertions spare the *int*-1 coding sequence, even when they are within the first or fourth coding exon. These observations are consistent with the proposal that the oncogenic action of *int*-1 is due solely to quantitative effects upon expression of the gene, but they do not exclude the possibility that alterations in the coding sequence of *int*-1 might assist—or even be required for—its postulated role as an oncogene. c-*myc* alleles affected by proviral insertions or translocations to immunoglobulin domains sometimes exhibit nucleotide substitutions that dictate amino acid changes of uncertain functional significance in c-*myc* protein (24, 27, 34). However, the cDNA sequence reported here shows that the *int*-1 gene in tumor 103 is not altered in its exonic regions. Assuming that the transcriptionally activated gene is playing an oncogenic role in this tumor, we conclude that amino acid substitutions or nucleotide changes in noncoding portions of the mRNA are not required for conversion of the normal *int*-1 gene to an oncogene. A functional assay for the neoplastic effects of *int*-1 will be required to evaluate these conclusions more thoroughly.

The *int*-1 protein. The deduced amino acid sequence of *int*-1 protein shows that it has a primary length of 370 amino acids, is cysteine rich near its carboxy terminus and hydrophobic at its amino terminus, and bears four potential glycosylation sites. Translation of in vitro transcripts of the cDNA clone in a cell-free system generates protein that moves in conventional sodium dodecyl sulfate-polyacrylamide gels with a mobility close to that expected for a 41-kilodalton protein. These deductions and observations do not, however, provide much insight into the location and function of the protein. The hydrophobicity and potential glycosylation sites may indicate membrane association or secretion, and the cysteine clustering is reminiscent of ligand-binding sites recently described for transmembrane receptors (6, 29, 30, 35). Assessment of these possibilities now requires detection of *int*-1 protein produced in cells that contain an activated *int*-1 gene.

ACKNOWLEDGMENTS

We thank Christoph Seeger for instruction in nucleotide sequencing techniques and for partial analysis of some of the initial cDNA clones, S. Nandi for tumor-bearing BALB/cfC3H mice, Roel Nusse and Op van Ooyen for discussions and communication of results before publication, and J. Marinos for help with the manuscript.

Y.K.T.F. and G.M.S. were supported by the Damon Runyon-Walter Winchell Cancer Fund, A.M.C.B. is a Special Fellow of the Leukemia Society, and H.E.V. is an American Cancer Society Research Professor. This work was supported by grants from the National Institutes of Health and American Cancer Society.

LITERATURE CITED

1. **Aviv, H., and P. Leder.** 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. Proc. Natl. Acad. Sci. USA **69**:1408–1412.
2. **Benton, W. D., and R. W. Davis.** 1977. Screening λgt10 recombinant clones by hybridization to single plaques in situ. Science. **196**:180–182.
3. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
4. **Bishop, J. M.** 1983. Cellular oncogenes and retroviruses. Annu. Rev. Biochem. **52**:301–354.
5. **Bos, J. L., L. J. Polder, R. Bernards, P. I. Schrier, P. J. van den Elsen, A. J. van der Eb, and H. van Ormond.** 1981. The 2.2 kb E1b mRNA of human Ad12 and Ad5 codes for two tumor antigens starting at different AUG triplets. Cell **27**:121–131.
6. **Ebina, Y., L. Ellis, K. Jarnagin, M. Edery, L. Graf, E. Clauser, J.-H. Ou, F. Masiarz, Y. W. Kan, I. D. Goldfine, R. A. Roth, and W. J. Rutter.** 1985. The human insulin receptor cDNA: the structural basis for hormone-activated transmembrane signalling. Cell **40**:747–758.
7. **Edmonds, M., M. H. Vaughn, Jr., and H. Nakazato.** 1971. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. Proc. Natl. Acad. Sci. USA **68**:1336–1344.
8. **Fung, Y.-K. T., A. M. Fadly, L. B. Crittenden, and H.-J. Kung.** 1981. On the mechanism of retrovirus-induced avian lymphoid leukosis: deletion and integration of the proviruses. Proc. Natl. Acad. Sci. USA **78**:3418–3422.
9. **Hughes, S., K. Mellstrom, E. Kosik, F. Tamanoi, and J. Brugge.** 1984. Mutation of a termination codon affects *src* initiation.

Mol. Cell. Biol. **4**:1738–1746.
10. **Jay, G., S. Nomura, C. W. Anderson, and G. Khoury.** 1981. Identification of the SV40 agnoprotein: a DNA binding protein. Nature (London) **291**:346–349.
11. **Kozak, M.** 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. **47**:1–45.
12. **Kozak, M.** 1984. Selection of initiation sites by eukaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. Nucleic Acids Res. **12**:3873–3893.
13. **Kozak, M.** 1984. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin *in vivo*. Nature **308**:241–246.
14. **Krieg, P. A., and D. A. Melton.** 1984. Functional messenger RNAs are produced by SP6 in vitro transcription of cloned cDNAs. Nucleic Acids Res. **12**:7057–7070.
15. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
16. **Liu, C. C., C. C. Simonsen, and A. D. Levinson.** 1984. Initiation of translation at internal AUG codons in mammalian cells. Nature (London) **309**:82–85.
17. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
18. **Mardon, G., and H. E. Varmus.** 1983. Frameshift and intragenic suppressor mutations in a Rous sarcoma provirus suggest *src* encodes two proteins. Cell **32**:871–879.
19. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
20. **Mount, S.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10**:459–472.
21. **Nusse, R., and H. E. Varmus.** 1982. Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. Cell **31**:99–109.
22. **Nusse, R., A. Van Ooyen, D. Cox, Y. K. Fung, and H. E. Varmus.** 1984. Mode of proviral activation of a putative mammary oncogene (*int*-1) on mouse chromosome 15. Nature (London) **307**:131–136.
23. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3′ non-coding region sequences in eukaryotic mRNA. Nature (London) **263**:211–214.
24. **Rabbitts, T. H., P. H. Hamlyn, and R. Baer.** 1983. Altered nucleotide sequences of a translocated c-*myc* gene in Burkitt lymphoma. Nature (London) **306**:760–765.
25. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
26. **Seeger, C., D. Ganem, and H. E. Varmus.** 1984. Nucleotide sequence of an infectious molecularly cloned genome of the ground squirrel hepatitis virus. J. Virol. **51**:367–375.
27. **Showe, L. C., M. Ballantine, K. Nishikura, J. Erikson, H. Kaji, and C. M. Croce.** 1985. Cloning and sequencing of a c-*myc* oncogene in a Burkitt's lymphoma cell line that is translocated to a germ line alpha switch region. Mol. Cell. Biol. **5**:501–509.
28. **Smith, H. O., and M. L. Birnstiel.** 1976. A simple method for DNA restriction site mapping. Nucleic Acids Res. **3**:2387–2395.
29. **Ullrich, A., J. R. Bell, E. Y. Chen, R. Herrera, L. M. Petruzzelli, T. J. Dull, A. Gray, L. Coussens, Y.-C. Liao, M. Tsubokawa, A. Mason, P. H. Seeburg, C. Grunfeld, O. M. Rosen, and J. Ramachandran.** 1985. Human insulin receptor and its relationship to the tyrosine kinase family of oncogenes. Nature (London) **313**:756–761.
30. **Ullrich, A., L. Coussens, J. S. Hayflick, T. J. Dull, A. Gray, A. W. Tam, J. Lee, Y. Yarden, T. A. Libermann, J. Schlessinger, J. Downward, E. L. V. Mayes, N. Whittle, M. D. Waterfield, and P. H. Seeburg.** 1984. Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. Nature (London) **309**:418–425.
31. **van Ooyen, A., and R. Nusse.** 1984. Structure and nucleotide sequence of the putative mammary oncogene *int*-1: proviral insertions leave the protein-encoding domain intact. Cell **39**:233–240.
32. **van't Veer, L. J., A. G. van Kessel, H. van Heerikhuizen, A. van Ooyen, and R. Nusse.** 1984. Molecular cloning and chromosomal assignment of the human homolog of *int*-1, a mouse gene implicated in mammary tumorigenesis. Mol. Cell. Biol. **4**:2532–2534.
33. **Varmus, H. E.** 1984. The molecular genetics of cellular oncogenes. Annu. Rev. Genet. **18**:553–612.
34. **Westaway, D., G. Payne, and H. E. Varmus.** 1984. Deletions and base substitutions in provirally mutated c-*myc* alleles may contribute to the progression of B-cell tumors. Proc. Natl. Acad. Sci. USA **81**:843–847.
35. **Yamamato, T., G. C. Davis, M. S. Brown, W. J. Schneider, M. L. Casey, J. L. Goldstein, and D. W. Russell.** 1984. The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. Cell **39**:27–38.