MOLECULAR AND CELLULAR BIOLOGY, Mar. 1993, p. 1507–1515
0270-7306/93/031507-09$02.00/0

Vol. 13, No. 3

Retroposition in a Family of Carcinoma-Associated Antigen Genes

ALBAN J. LINNENBACH,[1]* BETH A. SENG,[1] SHUANG WU,[1] SHIRA ROBBINS,[1] MAUREEN SCOLLON,[1] JANIA J. PYRC,[1] TERESA DRUCK,[2] AND KAY HUEBNER[2]

The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104,[1] and The Jefferson Cancer Institute, Thomas Jefferson University, Philadelphia, Pennsylvania 19107[2]

Received 4 August 1992/Returned for modification 2 December 1992/Accepted 13 December 1992

The gene encoding the carcinoma-associated antigen defined by the monoclonal antibody GA733 is a member of a family of at least two type I membrane proteins. This study describes the mechanism of evolution of the GA733-1 and GA733-2 genes. A full-length cDNA clone for GA733-1 was obtained by screening a human placental library with a genomic DNA probe. Comparative analysis of the cDNA sequence with the previously determined genomic sequence confirmed that GA733-1 is an intronless gene. The GA733-2 gene encoding the monoclonal antibody-defined antigen was molecularly cloned with a cDNA probe and partially sequenced. Comparison of GA733-2 gene sequences with the previously established cDNA sequence revealed that this gene consists of nine exons. The putative promoter regions of the GA733-1 and GA733-2 genes are unrelated. These findings suggest that the GA733-1 gene was formed by the retroposition of the GA733-2 gene via an mRNA intermediate. Prior to retroposition, the GA733-2 gene had been affected by exon shuffling. Analysis of GA733-2 exons revealed that many delineate structural motifs. The GA733-1 retroposon was localized either to chromosome region 1p32-1p31 or to 1p13-1q12, and the GA733-2 founder gene was localized to chromosome 4q.

GA733 is a monoclonal antibody (MAb)-defined, 40-kDa cell surface glycoprotein that is associated with human carcinomas of various origins (9, 11). Full-length cDNA clones for this antigen have been obtained from human cell lines by several investigators (22, 24, 28, 29). MAbs recognizing the human GA733 antigen are being evaluated for the experimental diagnosis and passive immunotherapy of gastrointestinal tumors. As a first step toward experimental active immunotherapy of cancer patients, cDNA sequences for the human antigen have been transferred into the baculovirus expression system (26) in order to produce sufficient quantities of antigen for immunization.

Protein and DNA sequence analyses revealed that there are at least two related GA733 genes (15). Our initial cloning effort was based on the partial amino acid sequence of the native GA733 antigen that had been purified from the colorectal carcinoma cell line SW948. A 54-base oligonucleotide probe was designed and used to screen a human genomic library. The first genomic isolate (GA733-1) was found to encode a protein with a sequence similar to that of the native antigen derived from SW948 cells. This indicated that the native protein which we had partially sequenced is encoded by a second gene (GA733-2). Subsequently, we determined that clones for the GA733-2 gene were not represented in the genomic library used in our initial screening.

Analysis of the GA733-1 genomic DNA sequence revealed a long open reading frame for 323 amino acids. A 35.7-kDa protein with four potential N-linked glycosylation sites was predicted from the genomic sequence. The distribution of hydrophobic amino acids was characteristic of a type I membrane protein. A putative promoter region with an Sp1 binding site, an atypical CAAT box, and a TATA box was observed 5′ to the open reading frame. Polyadenylation signals were found 3′ to the open reading frame. The GA733-1 gene was found to be actively transcribed in pancreatic carcinoma cell lines (15).

To obtain GA733-2 cDNA clones, we utilized an independent cloning strategy based on immunoselection with MAb GA733 (29). A cDNA library was prepared from the human colorectal carcinoma cell line SW948, and cDNA clones were isolated by using the high-efficiency COS cell expression system. The amino acid sequence predicted by GA733-2 cDNA was identified as that of the native GA733-2 antigen. The GA733-2 and GA733-1 amino acid sequences were 49% identical. The distributions of hydrophobic and hydrophilic residues in the two sequences were strikingly similar.

The GA733-1 gene appeared not to be a defective retropseudogene but rather an apparently intronless gene with the potential to encode a protein with several structural features in common with the transmembrane glycoprotein GA733-2. Descriptions of expressed retroposons and their exon- and intron-containing founder genes are rare (4). More commonly, promoter acquisition does not occur and the retroposon becomes an apparently nonfunctional retropseudogene.

The results of this study establish the mechanism of evolution of the GA733-1 and GA733-2 genes. A cDNA clone for GA733-1 and genomic clones for GA733-2 were isolated and sequenced. This new information, along with previously established sequences, allowed a comparison of cDNA and genomic sequences for both the GA733-1 and GA733-2 genes. The chromosomal locations of the GA733-1 and GA733-2 genes were also determined.

MATERIALS AND METHODS

GA733-1 mRNA expression. Cytoplasmic RNA was isolated from cell lines by using the vanadium method (2), and total cellular RNA was isolated from placenta by using the guanidinium-cesium chloride method (17). RNA was fractionated by oligo(dT)-cellulose chromatography (1).

Aliquots (5 μg) of poly(A)$^+$ mRNA were electrophoresed on a 2.2 M formaldehyde–1% agarose gel (13) and transferred to nitrocellulose. The plasmid subclone Pst-B5.9 derived from GA733-1 genomic coding sequences (15) was radiolabeled by nick translation and hybridized to the filter in 50% deionized formamide–5× SSC (20× SSC is 3.0 M

* Corresponding author.

sodium chloride plus 0.3 M sodium citrate [pH 7.0])–1× Denhardt's solution (50× Denhardt's solution is 1% Ficoll 400, 1% polyvinylpyrrolidone, and 1% bovine serum albumin)–0.1% sodium dodecyl sulfate (SDS)–0.05 M $NaPO_4$ (pH 6.5)–sonicated denatured salmon sperm DNA (50 μg/ml)–poly(A), poly(I), and poly(C) (25 μg/ml each) for 18 h at 42°C. The filter was washed at high stringency in 0.1× SSC–0.1% SDS at 65°C.

Molecular cloning. GA733-1 cDNA clones were isolated from a human placental cDNA library (Clontech Laboratories, Inc.). λgt11 recombinants (10^6) were plated and transferred to nitrocellulose filters. Plaque hybridization was carried out as described previously (18) with radiolabeled Pst-B5.9 genomic insert DNA that had been gel purified by electroelution and Elutip-d chromatography (14). The filters were washed at high stringency in 0.1× SSC–0.1% SDS at 65°C. Plaque-purified clones were grown preparatively, banded on equilibrium CsCl gradients, and phenol extracted. cDNA inserts were sized by digestion with *Eco*RI and subcloned into the plasmid vector pBR322.

Two human genomic libraries were utilized to isolate the GA733-2 gene. The LN87 B-cell lymphoma bacteriophage EMBL 3A library (31) was screened with a full-length GA733-2 cDNA insert (29) that had been gel purified and radiolabeled by nick translation. To isolate the 3′ end of the gene, a lambda DASH lymphocyte library (Stratagene) was screened with a DNA probe that had been prepared by the polymerase chain reaction (PCR) with primers designed to amplify the 3′-end 450 bases of GA733-2 cDNA. The PCR product was radiolabeled by the random-primed method (7). GA733-2 genomic clones were digested with various restriction enzymes and blotted. Exon-containing restriction fragments were identified by hybridization to radiolabeled full-length cDNA and then subcloned into the pBluescript vector (Stratagene).

DNA sequence analysis. To determine the sequence of GA733-1 cDNA, the 1.8-kb *Eco*RI insert of the pBR322 subclone 17-1-4 was gel purified, digested with various restriction enzymes, and subcloned into the M13mp18 and M13mp19 vectors. 17-1-4 DNA was sequenced by the dideoxynucleotide method (23) with T7 DNA polymerase (Pharmacia) and a vector-specific primer. To resolve compressions in GC-rich regions, two sets of sequencing reactions were carried out, one substituting 7-deaza-deoxyguanosine 5′-triphosphate for dGTP.

The structure of the GA733-2 gene was established by partial DNA sequence analysis. Double-stranded DNA templates were prepared from exon-containing pBluescript subclones and sequenced by using the dideoxynucleotide and specific-primer-directed methods (27). Initially, sequence data were obtained by using oligonucleotide primers designed from the cDNA sequence (29). Exon and intron junctions were identified, and reverse-sequencing primers were designed from intron sequences to confirm and extend the data. Some of these primers were useful in subsequent experiments to measure the sizes of small introns by PCR analysis.

DNA sequences were assembled by using the Staden programs (25). Relationships of GA733-2 coding sequences with known sequences were determined by searching release 70 of GenBank with the program TFASTA (21), release 30 of the NBRF protein data base with FASTP (16), and release 17 of the Swissprot data base with FASTA (21). Results of database searches were further analyzed for statistical significance by using ALIGN (6). Multiple sequence alignments were then performed with CLUSTAL V (12).

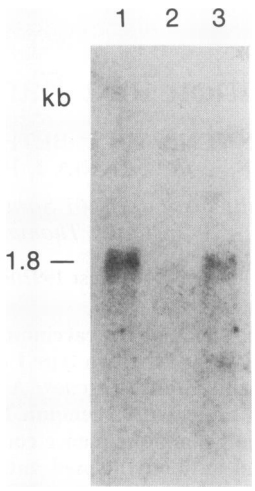

FIG. 1. Expression of GA733-1 mRNA in the placenta. Poly(A)$^+$ mRNA (5 μg) was electrophoresed on a 1% agarose gel containing 2.2 M formaldehyde and transferred to a nitrocellulose filter. The hybridization probe was a nick-translated 0.85-kb *Pst*I restriction fragment of the GA733-1 genomic clone (15). The filter was washed at high stringency with 0.1× SSC–0.1% SDS at 65°C. mRNAs were derived from SW948 colon carcinoma cells (lane 1), SKmel37 melanoma cells (lane 2), and placenta (lane 3).

Chromosome mapping. Isolation, propagation, and characterization of most somatic cell hybrids used in this study have been described elsewhere (5, 19). Hybrids 7299, 10115, 9142, 7300, and 10027 were obtained from the National Institute of General Medical Sciences Human Genetic Mutant Cell Repository (Coriell Institute, Camden, N.J.).

Hybrid and control DNAs were prepared by cell lysis, proteinase K digestion, phenol extraction, and ethanol precipitation. Cellular DNAs were digested with the appropriate restriction enzyme, fractionated on 0.8% agarose gels, transferred to a nylon membrane, and hybridized in 50% formamide–5× SSPE (1× SSPE is 0.15 M NaCl, 0.1 M Na_2PO_4, and 0.001 M EDTA [pH 7.5])–5× Denhardt's solution–0.1% SDS–sonicated denatured salmon sperm DNA (100 μg/ml) for 16 h at 42°C. GA733-1 and GA733-2 probes for filter hybridizations were radiolabeled by nick translation to a specific activity of 10^8 cpm/0.1 μg; 10^8 cpm was used for each filter hybridization.

Nucleotide sequence accession numbers. The GA733-1 cDNA sequence has been assigned GenBank data base accession number X13425. The GA733-1 gene has been assigned the symbol M1S1 by the Human Gene Mapping Nomenclature Committee. GA733-2 exon sequences have been assigned GenBank accession numbers M93029 through M93036.

RESULTS

Isolation of GA733-1 cDNA clones. A Northern (RNA) blot experiment was conducted to identify a cellular source for the isolation of GA733-1 cDNA clones. We have previously shown that a 1.8-kb message for GA733-1 was expressed abundantly in pancreatic carcinoma cell lines and to a lesser extent in the SW948 colorectal carcinoma line (15). The presence of GA733-1 mRNA in the human placenta was tested, given the commercial availability of a λgt11 library. A 1.8-kb message for GA733-1 was detected in placenta (Fig. 1,

FIG. 2. Sequence of the GA733-1 full-length cDNA clone 17-1-4, shown in relation to the 2.2-kb genomic sequence (15). The intronless nature of the GA733-1 gene is apparent from the colinearity of the two sequences. A putative Sp1 binding site, an atypical CAT box, and a TATA box are underlined. Amino acid residues predicted by the cDNA sequence which differ from those predicted from the genomic sequence are indicated in brackets. Nucleotide positions differing in the two sequences are marked with asterisks. Two possible polyadenylation signals in the cDNA sequence are underlined.

lane 3) and in positive-control SW948 cells (Fig. 1, lane 1) but not in the negative-control melanoma cell line SKmel37 (Fig. 1, lane 2). Under the stringent conditions of this experiment, no hybridization was observed for the 1.45-kb GA733-2 mRNA which is expressed in SW948 cells (29).

Six GA733-1 cDNA clones were plaque purified from the placental library by screening with a genomic DNA fragment containing GA733-1 coding sequences. *Eco*RI digestion indicated that two of the six clones contained full-length 1.8-kb inserts. To facilitate further analysis, a full-length cDNA clone was subcloned into the plasmid vector pBR322.

Sequence comparison between the GA733-1 cDNA clone

```
5'-                                   Sp1
GATCCCTAACGCCGCCATGGAGACGAAGCACCTGGGGCGGGGCGGAGCGGGGCGCGCGGGCCCACACCTGTGGAGAGGGCCGCGCCCCAACTGCAGCGCCGGGGCTGGGGGAGGGGAGCC   120
                                  AP-1
TACTCACTCCCCCAACTCCCGGGCGGTGACTCATCAACGAGCACCAGCGGCCAGAGGTGAGCAGTCCCGGGAAGGGGCCGAGAGGCGGGGCCGCCAGGTCGGGCAGGTGTGCGCTCCGCC   240
                                                                                                   Exon 1
CCGCCGCGCGCACAGAGCGCTAGTCCTTCGGCGAGCGAGCACCTTCGACGCGGTCCGGGGACCCCCTCGTCGCTGTCCTCCCGACGCGGACCCGCGTGCCCCAGGCCTCGCGCTGCCCGG   360
                                  ➥ cDNA 5'-end
                                                                                                MetAlaProProGlnValLeuAlaPheGlyLeuL
CCGGCTCCTCGTGTCCCACTCCCGGCGCACGCCCTCCCGCGAGTCCCGGGCCCCTCCCGCGCCCCTCTTCTCGGCGCGCGCGCAGCATGGCGCCCCCGCAGGTCCTCGCGTTCGGGCTTC   480
                                  | insertion →                |
euLeuAlaAlaAlaThrAlaThrPheAlaAlaAlaGlnGluG                                                          ...intron 1 (3.8 kb)...
TGCTTGCCGCGGCGACGGCGACTTTTGCCGCAGCTCAGGAAGGTGAGGCGCGGATTGGAGCAGAGTTGTGGAGCTGGGCTGGGCTGGGGGGCA TTATTATTACAATATAACTTAGCTGGG   600
                                                                                                Exon 2          Δ
                                                                                luCysValCysGluAsnTyrLysLeuAlaValAsnCysPheValAsnAsnAsnArgGlnCysGln
ACATGAGAGTTAATAGATCCACATTTTAAAGTAGATTTTTTTTTTAATTTTCTAGAATGTGTCTGTGAAAACTACAAGCTGGCCGTAAACTGCTTTGTGAATAATAATCGTCAATGCCAG   720

CysThrSerValGlyAlaGlnAsnThrValIleCysSerLysL                                                         intron 2 (0.237 kb)
TGTACTTCAGTTGGTGCACAAAATACTGTCATTTGCTCAAAGCGTGAGTAAAATATCCTAATTACCTGTAAGCTTTATTTTGACTTAATACTTCTTTAATTGATGTGCCTTGAGTTGGAA   840

AGAGTTTTATTGGCTTAAATCTGAATCATGTTACAAAGTAAGTGTGGGAACACATAAATTTCAAATAATCTTTGACCCTGGAACTTTAGAGTTAATTTTTTTTTCCCGTAATCATGAAA   960

                                                euAlaAlaLysCysLeuValMetLysAlaGluMetAsnGlySerLysLeuGlyArgArgAlaLysProGluGlyAlaLeu
TCAGTTATTTTTCAGTTTGGCATTAAGGTTTCTTTTTCAGTGGCTGCCAAATGTTTGGTGATGAAGGCAGAAATGAATGGCTCAAAACTTGGGAGAAGAGCAAAACCTGAAGGGGCCCTC   1080
                                                                Exon 3
GlnAsnAsnAspGlyLeuTyrAspProAspCysAspGluSerGlyLeuPheLysAlaLysGlnCysAsnGlyThrSerThrCysTrpCysValAsnThrAlaGlyValArgArgThrAsp
CAGAACAATGATGGGCTTTATGATCCTGACTGCGATGAGAGCGGGCTCTTTAAGGCCAAGCAGTGCAACGGCACCTCCACGTGCTGGTGTGTGAACACTGCTGGGGTCAGAAGAACAGAC   1200

LysAspThrGluIleThrCysSerGluArgValArgThrTy
AAGGACACTGAAATAACCTGCTCTGAGCGAGTGAGAACCTAGTGAGTGGGGCTGCCTATACTACTTGTTTTCATGCTGTTCAGATTCATTTAATTAAATTTATTTTTGATTATGTAATAT   1320
                                                                                                Exon 4
        ...intron 3 (1.1 kb)...                                                 rTrpIleIleIleGluLeuLysHisLysAlaArgGluLysProTyrAspSerLysSerLeuArgTh
GATTTCATGGTT AACTGAACATTGTCTTTTTAACTTTATAGCTGGATCATCATTGAACTAAAACACAAAGCAAGAGAAAAACCTTATGATAGTAAAAGTTTGCGGACGTAAGTGCAATTA   1440
            Δ
AATGCATCATATTCTTGCACAGTTGGTGGCTCAAATCTTCCATCCTACACCATTAGAAAAAGCCAAGTCTAAATGCTTTTTTATATTTCTGAAAAATAAAGTTACTTGAAATAGAGTTGC   1560
AAGAATAGCACAGAGATTCTGGGAATACACTTCACTCAGATTCACCAATTAACATTTTGGCACATTTGCTTTTTATATGTGTATGTGTGGATGAATATGTGTGTGTGCTTTACATCAGTG   1680
    ...intron 4 (1.7 kb)...
ATCTATG TCCAGCCTGAATCAATTATTATTATGATAGTTGCAAAATGGCAGAAAAATTTTAACTTTAATGACAGTTTTAGACCCTGAGCTGTCTGCTTAAAGAGTAGTGCTTCTTACTGT   1800
       Δ                                                                                        Exon 5
                                                                                rAlaLeuGlnLysGluIleThrThrArgTyrGlnLeuAspProLysPheIleThrSerIleLeu
TGTGTGGTACAAACATTTTTTTTTAATACAGATTTTAAATTCTTTACAGTGCACTTCAGAAGGAGATCACAACGCGTTATCAACTGGATCCAAAATTTATCACGAGTATTTTGGTATGAT   1920

TTTTTAATAAGTGAGCTTTAGCAGACAGTTGGTGAGACAGTATGTTTTGAGTATAAGGACAGCCAGTGATTTAAGTGGTGGTTAAATGCACTTACTGGAGCAACAGTTTCGGATCTGGGT   2040
            ...intron 5 (1.9 kb)...
ACTTAATGTGAATT AAAATACTTTCATATCTTATTATTTAAAAATCTTTTCAATATAAGAAAATTCTCTTAGAAAAATTGTACATTGTAATTATGTTTGGGTTGCATGGCTGTCTTATTT   2160
              Δ
CCCTTTGATAGATTTAGAGACCTCCCAAAGATTTCTTGATTAGTGATAAACTTAGTTATCCACTAATGGAAAGGAACAGTGATGCATGTAGATTATAGAAAATCAAACACTGAATATTCT   2280
                                                                                                Exon 6
                                                                                TyrGluAsnAsnValIleThrIleAspLeuValGlnAsnSerS
GATTCTCAATTAATGTTATTTTCAAATGATTTTGATTATATTAGTATTAATTTGTATTATTCAATTTTTTTCCCCAGTATGAGAATAATGTTATCACTATTGATCTGGTTCAAAATTCTT   2400

erGlnLysThrGlnAsnAspValAspIleAlaAspValAlaTyrTyrPheGluLysAsp
CTCAAAAAACTCAGAATGATGTGGACATAGCTGATGTGGCTTATTATTTTGAAAAAGATGTGAGTATCATCTTCTTTATTCCTGTGTTCAGGAATGTAGTCTATCATGCCTCAATGAATT   2520

AAATATATTTCATCACCTTTTTATCCACTTACAGATCAACCAAATGGTTCGCTGCTGCCGTTAATTTTGTCCTCCCTGTCACTCACATGCATCTTGCTTGTTTGTATATTTATGCCTCTT   2640
            ...intron 6 (0.7 kb)...
ATCAAATTGTAAGT AAGAGAGGCTATGTGTTGTTGGCCTTGCATCTGGGTGATAGGTACATGGGCATCATTGCACTACTCTTTCTACTTTCGTGTATGTTGAAAGGTTCCTGTAATAAAC   2760
              Δ
AGTTTTTTAAAGTTCCAATAAATTAGATTGTTATCACTAAAACCATAAAGATTCTTGGCAGCGGTTCTTTTGGCATACAATTTGTATGTAATTATATGTGGCCATGGTTGGTTTCCTTAA   2880
                                                                                                Exon 7
                                                                        ValLysGlyGluSerLeuPheHisSerLysLysMetAspLeuThrValAsnGlyGluGlnLeuAspLeuAspProGlyGlnThrLe
ATATTTTTAATTCCTTTTCTCCTTTTCAATACAGGTTAAAGGTGAATCCTTGTTTCATTCTAAGAAAATGGACCTGACAGTAAATGGGGAACAACTGGATCTGGATCCTGGTCAAACTTT   3000

uIleTyrTyrValAspGluLysAlaProGluPheSerMetGlnGlyLeuLysAlaGlyValIleAlaValIleValValValValIleAlaValValAlaGlyIleValValLeu
AATTTATTATGTTGATGAAAAAGCACCTGAATTCTCAATGCAGGGTCTAAAAGCTGGTGTTATTGCTGTTATTGTGGTTGTGGTGATAGCAGTTGTTGCTGGAATTGTTGTGCTGGTGAG   3120
                                                                                                ...intron 7 (>1.4 kb)...
TACAGAACAAGTAAAATTTCATTTAAGGGTATATTTTTTCAAGAAAAAGTAATAGTGGCTGGGCGCGGTGGCTCACCACACCTGTTATCCCTACACTT TCAAAATTATGTCTATTAAAAG   3240
                                                                                                Δ
                                                                                                Exon 8
                                                                                ValIleSerArgLysLysArgMetAlaLysTyrGluLysAlaGlu
CATATATGTCTGTTTAGATAATCTTTTTTTGAATAGCAGTCCTAAAACAATAGTTGTCTTTCTTCCACTCAGGTTATTTCCAGAAAGAAGAGAATGGCAAAGTATGAGAAGGCTGAGGTA   3360

AATGGATTACTTACCTAAATAGAAAGGCCCTGTTGAATCTCTTACTCCTAATCACTCTACCTTCCTACACACTGATGCATTTCAGTTATACTGGAGTCCCTTTATACTGTTGTCTTTAGG   3480
                                                                                                ...intron 8 (1.2 kb)...
GTCTTAGGGACAGTCTTAGAATGTACTCTTACCTAAATATTCTTGCGTTGAGTTTCCATGGCAGATCACCATCT AATATAACTTTTTCTTTTTTTATTTAAATAAGTCTTATAAATGTGG   3600
                                                                                                Δ
                                                                                                                        IleLysG
GAAAAAATTATCTTGTGTTCCTTTAATTTCATTTTTATTTAATACTATTTTCAGAATGAACAAAAGATTGAAAAATTATTTAGAATTTTTTTCTGTGCTTTTTCCTGTTTCAGATAAAGG   3720
                Exon 9
luMetGlyGluMetHisArgGluLeuAsnAlaEnd
AGATGGGTGAGATGCATAGGGAACTCAATGCATAACTATATAATTTGAAGATTATAGAAGAAGGGAAATAGCAAATGGACACAAATTACAAATGTGTGTGCGTGGGACGAAGACATCTTT   3840

GAAGGTCATGAGTTTGTTAGTTTAACATCATATATTTGTAATAGTGAAACCTGTACTCAAAATATAAGCAGCTTGAAACTGGCTTTACCAATCTTGAAATTTGACCACAAGTGTCTTATA   3960
TATGCAGATCTAATGTAAAATCCAGAACTTGGACTCCATCGTTAAAATTATTTATGTGTAACATTCAAATGTGTGCATTAAATATGCTTCCACAGTAAAATCTGAAAAACTGATTTGTGA   4080
                                                                                                                        ➥ cDNA 3'-end
TTGAAAGCTGCCTTTCTATTTACTTGAGTCTTGTACATACATACTTTTTTATGAGCTATGAAATAAAACATTTTAAACTGAATTTCTTAACTTTGACATTTCAAATTTCTTCTTCTTTTT   4200
CTTTTCTTTTTTTTTTTTTTTTGAGA-3'  4226                                                    ➥ cDNA 3'-end
```

FIG. 3. Organization of the GA733-2 gene. Putative binding sites for the Sp1 and AP-1 and transcription factors are underlined. The position corresponding to the 5′ end of the cDNA clone with the longest reported 5′-untranslated region (22) is underlined. Vertical bars delineate the 20-base insertion observed in the 5′-untranslated region of one of three cDNA clones, and the associated 10-base repeats are overlined. The first and last 2 bases of the introns conform to the GT-AG rule. Three poly(A) addition signals observed in cDNA clones are underlined.

```
                  | Exon 2:EGF-like repeat                          | Exon 3:Thyro-like repeat

GA733-2           ECVCENYKLAVNCFVNNNR-QCQCTSVGAQNTVIC---SKLAAKCLVMKAEMNGSKLGRR
GA733-1           NCTCPTNKMTVCSPDGPGG-RCQCRALGSGMAVDC---STLTSKCLLLKARMSAPKNART
NIDOGEN           ECQPSRCHPDAFCYNTPGSFTCQCKPGYQGDGFRCVPGEVEKTRCQHEREHILGAAGATD
IGFBP2            ----------------------------------------RTPCQQELDQVLERISTMR
IGFBP4            -------------------------AGVQ---------DTEMGPCRRHLDSVLQQLQTEV
IGFBP3            ------------------------------------------GPCRREMEDTLNHLKFLN
THYRO             -C-PTKCEVERF---TATSFG-------------------------------------
HDRI              V--------------------------------------LTKCQEEVSHI--------


GA733-2           -AKP-EGALQNNDGLYDPDCDESGLFKAKQCNGTST-----CWCVNTAGVRRTD---K-D
GA733-1           LVRPSEHALVDNDGLYDPDCDPEGRFKARQCNQTSV-----CWCVNSVGVRRTD---KGD
NIDOGEN           PQRPI-PP-----GLFVPECDAHGHYAPTQCHGST----GYCWCVDRDGREVEGTRTRPG
IGFBP2            LPDER-GPLEHLYSLHIPNCDKHGLYNLKQCKMSLNGQRGECWCVNPNTGKLIQGAPTIR
IGFBP4            YRGAQ-T-------LYVPNCDHRGFYRKRQCRSSQGQRRGPCWCVDRMGKSLPG-SPDGN
IGFBP3            VLSPR-G-------VHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQPLPGYTTKGK
THYRO             ------HP-------YVPSCRRNGDYQAVQCQ-TE----GPCWCVDAQGKEMHGTR-QQG
HDRI              --PAV-HP-----GSFRPKCDENGNYLPLQCYGSI----GYCWCVFPNGTEVPNTRSRG-
                                     * *   * .   **  .       ****


                                  |
GA733-2           TEITCSERVRTY
GA733-1           LSLRCDELVRT-
NIDOGEN           MTPPCL----ST
IGFBP2            GDPECH-L---F
IGFBP4            GSSSCP-TGSSG
IGFBP3            EDVHC----YSM
THYRO             EPPSC------A
HDRI              -HHNC-----SE
                     *
```

FIG. 4. Common ancestry of GA733-2 exons with other sequences. Portions of otherwise unrelated sequences are aligned to sequences encoded by exons 2 and 3 of the GA733-2 gene (delineated by vertical bars). Multiple sequence alignment was performed with CLUSTAL V by using a fixed gap penalty of 10, a floating gap penalty of 10, and the PAM 250 protein weight matrix. Amino acid positions identical (*) and chemically conserved (.) in all eight sequences are indicated. Cysteine positions are shown in boldface. IGFBP2, -3, and -4, insulin-like growth factor binding proteins 2, 3, and 4, respectively; THYRO, thyroglobulin (exon 8); HDRI, HLA-DR antigen-associated invariant chain (exon 6b).

and the previously sequenced GA733-1 gene confirmed that this gene contained no introns (Fig. 2). Single base differences were found at 7 of 1,793 possible positions. Three base substitutions occurred in the coding region. Two of these substitutions resulted in very conserved exchanges between Asp and Glu (codons 147 and 216), and the third substitution is silent (codon 276). These variations may represent polymorphisms, although it cannot be ruled out that reverse transcriptase errors had occurred during the construction of the cDNA library.

Isolation and partial DNA sequence analysis of the GA733-2 gene. Two λ clones for the GA733-2 gene were characterized in detail. Clone λ21726, spanning 13.2 kb, was isolated from the LN87 library by using a full-length cDNA probe. Clone λ19A12S66, spanning 19.0 kb, was isolated from the lymphocyte library by using a 3′-end DNA probe.

The organization of the GA733-2 gene was established by DNA sequence analysis. Comparison between genomic and cDNA sequence data (29) revealed that the GA733-2 gene consists of nine exons with consensus exon and intron junction sequences (Fig. 3). The λ21726 clone contained a portion of the putative promoter, exons 1 through 7, and a portion of intron 7. The λ19A12S66 isolate contained a nonoverlapping portion of intron 7, exon 8, and exon 9.

The full characterization of the GA733-2 promoter will require the isolation and analysis of additional genomic clones and a definitive establishment of the RNA start sites. On the basis of limited data, the GA733-2 and GA733-1 genes have different promoters. Putative promoter sequences of GA733-2 had no homology with sequences 5′ to the GA733-1 coding region. GA733-2 was found not to have a TATA or CAAT box (Fig. 3), as was observed in the putative GA733-1 promoter region (15) (Fig. 2). The consensus binding site for the Sp1 transcription factor observed in GA733-2 differed from the Sp1 site found in GA733-1 (15) (Fig. 2). A potential AP-1 transcription factor binding site was identified in the GA733-2 sequence (Fig. 3), whereas none was observed in GA733-1 (15) (Fig. 2).

We had previously observed a 20-base insertion in the 5′-untranslated region of the GA733-2-2 cDNA clone (29) but not in independently isolated cDNA clones (22, 28). Analysis of GA733-2 genomic sequences revealed that this difference was not due to alternative splicing. The entire 107-base 5′-untranslated region of GA733-2-2 was contained within exon 1. Ten bases of flanking DNA are repeated in the 20-base insert. This phenomenon may represent a polymorphism; however, the Mendelian inheritance of the insertion has not been investigated.

The boundaries of the GA733-2 exons corresponded to the predicted structural domains of this type I membrane protein (29). Amino acid sequences encoded by exon 1 correspond to the putative signal peptide sequence. The putative extracellular domain is encoded by exons 2 to 6. Exon 7 encodes the hydrophobic transmembrane domain. Exon 8 encodes a portion of the cytoplasmic domain, including a clustering of six positively charged amino acids. Exon 9 contains the final 13 residues of the cytoplasmic domain, the termination codon, and all of the 3′-untranslated region.

Introns with sizes of ≤1.7 kb were successfully sized by PCR amplification of cloned genomic DNA templates with primers from adjacent exons. The identities of PCR products were verified by Southern blotting with nested oligonucle-

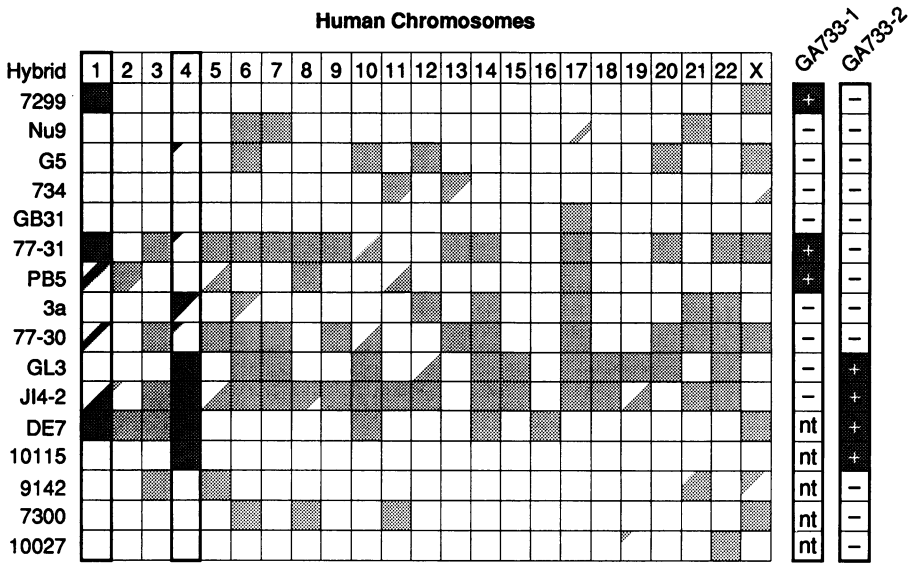

FIG. 5. Chromosomal mapping studies of the GA733-2 gene family. A completely stippled box indicates that the hybrid named in the left column contains the chromosome indicated in the upper row; lower-right stippling indicates the presence of the long arm (or part of the long arm, indicated by a smaller fraction of stippling) of the chromosome shown above the column; upper-left stippling indicates the presence of the short arm (or partial short arm) of the chromosome listed above the column. The columns for chromosomes 1 and 4 are boldly outlined and stippled to highlight the correlation of the presence of these chromosomes (or regions of chromosomes) with the presence of the GA733-1 and GA733-2 genes, respectively. The pattern of retention of the GA733-1 and GA733-2 genes in the panel is shown to the right of the figure where the presence of the gene in a hybrid is indicated by a stippled box with a plus sign and absence of the gene is indicated by an open box enclosing a minus sign. nt, not tested.

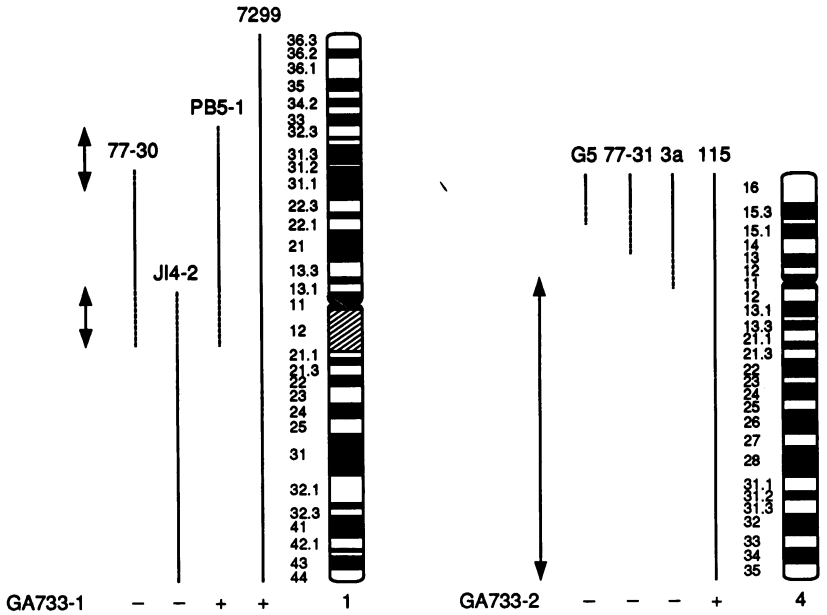

FIG. 6. Regional localization of the GA733-1 and GA733-2 genes. DNA from rodent-human hybrids retaining portions of chromosome 1 or 4 was tested for retention of the respective genes as shown in Fig. 5. Most hybrids retaining partial chromosome 1 have been described previously (19). The exact breaks in hybrids 77-30, JI4-2, and PB5-1 are not known, as indicated by the dotted regions. Because of this uncertainty, there are two regions on chromosome 1 where GA733-1 could map, either 1p32-1p31 or 1p13-1q12. For chromosome 4, hybrids G5, 77-31, and 3a carry a portion of chromosome 4p. Hybrid GM10115 carries only chromosome 4. The presence of the GA733-2 gene in hybrid GM10115 and its absence in hybrids G5, 77-31, and 3a localize the GA733-2 gene to chromosome region 4q. The arrows indicate the regions where the GA733-1 or GA733-2 genes are located.

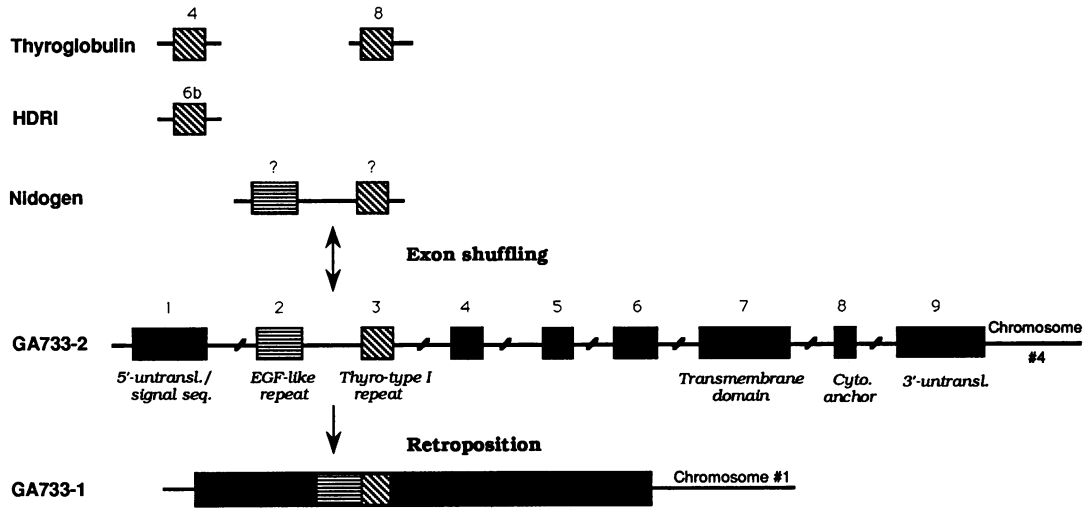

FIG. 7. Evolution of the GA733-1 and GA733-2 genes. Exon-shuffling events have occurred among the genes for the HLA-DR-associated invariant chain (HDRI), thyroglobulin, and GA733-2. Nidogen gene structure is unknown but is likely to consist of exons and to have participated in the shuffling of a pair of exons, since two closely linked exons of GA733-2 are homologous to sequences juxtaposed in nidogen. The GA733-2 gene consists of a total of nine exons, some encoding structural motifs of the transmembrane protein. The retroposition of the GA733-2 gene occurred after the exon-shuffling events.

otides (data not shown). Larger introns had to be sized by restriction enzyme mapping and Southern blotting with exon-specific oligonucleotide probes. Intron 2 was determined by DNA sequencing. Restriction enzyme analysis of a λ21726 subclone indicated that intron 7 is at least 1.4 kb. The full extent of this intron could not be determined, despite the analysis of two additional λ clones. This intron is apparently quite large, and its isolation and characterization will probably require cosmid cloning. On the basis of the analyses of the 21726 and 19A12S66 λ clones, a minimum estimate of 13.9 kb can be obtained for the size of the GA733-2 gene.

Sequences homologous to GA733 genes. Computer analysis of sequence data bases revealed homologies characteristic of exon shuffling. A multiple sequence alignment was performed with amino acid sequences encoded by exons 2 and 3 of GA733-2, a portion of GA733-1, and portions of six human sequences otherwise unrelated to the GA733 genes (Fig. 4).

All exons of the GA733-2 gene are represented in the intronless GA733-1 gene. The sequence motifs encoded by exons 2 and 3 are also juxtaposed in the cell matrix protein nidogen. Exon 2 of GA733-2 encodes a cysteine-rich sequence that is homologous to the epidermal growth factor-like repeat IId of nidogen (20, 24). Exon 3 of GA733-2 encodes a cysteine-rich thyroglobulin type I repeat, as does a segment of nidogen, exon 6b of the HLA-DR-associated invariant chain gene, and segments of several insulin-like growth factor binding proteins. Overall, 10 of 117 possible positions were identical, and 2 were conserved in all eight sequences. The positions of cysteine residues were particularly well conserved.

Southern blot hybridizations with GA733-1 and GA733-2 DNA probes under various stringencies indicated that sequences homologous to both genes were present in DNA derived from monkey COS cells, murine LTK$^-$ cells, hamster CHO cells, and chicken erythrocytes (data not shown). These studies gave no indication of recent duplications of either gene.

Chromosome mapping of the GA733 gene family. Southern blot analysis of a panel of DNAs derived from rodent-human hybrids (5, 10) was used to map the GA733-1 and GA733-2 genes. Chromosome regions retained by each hybrid are diagrammed in Fig. 5.

By using the pBR322 subclone 17-1-4 as a probe, the gene for GA733-1 was found to be located on chromosome 1 (summarized in Fig. 5). Hybrid 77-30 retains a portion of chromosome 1 between 1p31 and 1q12 and is negative for GA733-1. Hybrid J14-2 retains a portion of 1q distal to 1q12 and is negative for the GA733-1 gene. Hybrid PB5-1 retains a portion of chromosome 1 between 1p32 and 1q12 and is positive for GA733-1. As diagrammed in Fig. 6, the GA733-1 locus maps to chromosome region 1p32-1p31 or to the pericentric region of chromosome 1 (where the presence or absence of overlap among these three hybrids is unknown).

The GA733-2 locus was present only in hybrids retaining chromosome 4 and was absent in hybrids which did not contain chromosome 4 (summarized in Fig. 5). Hybrids G5, 77-31, 3a, and 77-30 retain portions of 4p but are negative for sequences which map to 4q (not shown). These hybrids are negative for GA733-2 sequences, suggesting that the GA733-2 locus maps to 4q (Fig. 6).

Evolution of the GA733 gene family. The combined results of molecular cloning, DNA sequencing, amino acid sequence comparisons, and chromosome mapping studies indicate that both exon shuffling and retroposition have been factors in the evolution of the GA733 gene family (Fig. 7).

Exon 3 of the GA733-2 gene exhibits common ancestry with the thyroglobulin type 1 repeat encoded by exon 8 of the thyroglobulin gene. Exon 6b of HLA-DR-associated invariant chain is homologous to exon 3 of GA733-2, by virtue of its homology with exon 4 of thyroglobulin, which also encodes a type 1 repeat. The nidogen and GA733-2 genes may have acquired a closely linked pair of exons for an epidermal growth factor-like repeat and a thyroglobulin type I repeat.

The chronological order of these complex exon-shuffling

events is unknown. However, these exon-shuffling events must have preceded retroposition, since all GA733-2 exons are represented in the GA733-1 sequence. The retroposition event in this gene family occurred between ancestral chromosome regions that correspond to present-day regions of chromosomes 4 and 1.

DISCUSSION

To allow for immunological, physical, and biochemical studies of MAb-defined tumor-associated antigens, we have molecularly cloned antigen sequences (15, 29, 30) to produce large quantities of recombinant antigen (26). During the course of a genomic cloning experiment designed to isolate the gene for the antigen defined by MAb GA733, we isolated the related GA733-1 gene with a GA733-2 oligonucleotide probe (15). Serendipitously, the oligonucleotide probe corresponded to the region with the highest degree of homology between the two sequences (29). The unusual structure of the transcriptionally active GA733-1 gene led us to investigate the evolution of the GA733 gene family by studying how the GA733-2 gene was organized.

For this study, a full-length GA733-1 cDNA clone and genomic clones for GA733-2 were isolated, analyzed by DNA sequencing, and chromosomally mapped. By comparing GA733-1 cDNA and genomic sequences, we have confirmed that the actively transcribed GA733-1 gene is intronless. By comparing GA733-2 cDNA and genomic sequences, we have determined that the GA733-2 gene consists of nine exons with consensus exon and intron junction sequences. Chromosome mapping studies showed that the GA733-2 gene is located on chromosome 4 and that the GA733-1 gene is located on chromosome 1. Thus, gene duplication had occurred by a retroposition event whereby the GA733-2 founder gene gave rise to the GA733-1 retroposon.

The GA733-1 promoter is necessarily distinct from the promoter of its founder gene, since GA733-1 arose via an mRNA intermediate. This is supported by the lack of sequence homology between genomic sequences 5′ to the GA733-1 and GA733-2 coding regions described here and by the previous observation of various levels of activity of these promoters in a given cell type (15, 29).

Portions of GA733-2 sequences were found to be homologous to portions of several genes. A comparison of the exon organization of these genes with that of GA733-2 revealed that the phenomenon of exon shuffling (8) could account for these homologies. The epidermal growth factor-like repeat encoded by exon 2 and the thyroglobulin-like repeat encoded by exon 3 of GA733-2 are present in several otherwise unrelated genes. It is interesting that these sequence motifs are also juxtaposed in the predicted amino acid sequence of the cell matrix adhesion protein nidogen. The exon organization of the nidogen gene is unknown; however, given the observation that intron 2 of GA733-2 has a relatively short length of 237 bases, it is likely that these exons were shuffled as one unit. The nidogen homology is the most extensive homology observed between GA733 and other sequences. On the basis of this relationship, it has been suggested that the GA733 antigens function in cell-cell or cell-matrix interactions (24).

Three other GA733-2 exons encode important structural features of the GA733-2 type I membrane protein. Exons 1, 7, and 8 encode the signal peptide, the transmembrane domain, and the cytoplasmic anchor, respectively. It is likely that the acquisition of exon 8 was a critical event in the evolution of this transmembrane protein, as it encodes a clustering of positively charged amino acids which probably anchor the protein in the membrane.

Sequences encoded by present-day GA733-2 exons are all represented in the GA733-1 gene. Thus, the exon-shuffling events predated the retroposition event, and no additional shuffling events have occurred since retroposition. Moreover, the degree of homology among shuffled exons is less than the degree homology observed between the GA733-1 and GA733-2 genes.

Evidence for the approximate age of the GA733-2 gene family was obtained by Southern blot assay for the presence of homologous sequences in other species. The monkey, mouse, hamster, and chicken were found to have sequences homologous to those of both the GA733-1 and GA733-2 genes. Direct evidence for a functional GA733-2 gene in the mouse has been provided by an analysis of a cDNA clone derived from a murine plasmocytoma cell line (3). The coding region of this murine cDNA was 80% homologous to human GA733-2 coding sequences and 56% homologous to human GA733-1 coding sequences. GA733 genes have not been molecularly cloned from other species. On the basis of Southern blot data, the retroposition event which gave rise to GA733-1 preceded the divergence of avian and mammalian lineages about 300 million years ago.

We have recently transferred a GA733-1 cDNA clone into a eukaryotic viral expression vector and have observed that recombinant virus directs the expression of a novel protein with the predicted molecular weight in infected cells. The production and purification of recombinant GA733-1 protein will enable the production of polyclonal and monoclonal antibodies. This will then allow the study of the expression of the native GA733-1 antigen in normal and tumor tissues.

ACKNOWLEDGMENTS

Oligonucleotide primers were synthesized by The Wistar Institute DNA Synthesis Facility.

This work was supported by NIH grants CA10815, CA21124, and CA39860.

REFERENCES

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. **Berger, S. L., and C. S. Berkenmeier.** 1979. Inhibition of intractable nucleases with ribonucleoside-vanadyl complexes: isolation of messenger ribonucleic acid from resting lymphocytes. Biochemistry **18:**5143–5149.
3. **Bergsagel, P. L., C. Victor-Kobrin, C. R. Timblin, J. Trepel, and W. M. Kuehl.** 1992. A murine cDNA encodes a pan-epithelial glycoprotein that is also expressed on plasma cells. J. Immunol. **148:**590–596.
4. **Brosius, J.** 1991. Retroposons—seeds of evolution. Science **251:**753.
5. **Cyr, C., K. Huebner, T. Druck, and R. Kris.** 1991. Cloning and chromosomal location of a human endothelin ETA receptor. Biochem. Biophys. Res. Commun. **181:**184–190.
6. **Dayhoff, M. O., W. C. Barker, and T. L. Hunt.** 1983. Establishing homologies in protein sequences. Methods Enzymol. **91:** 524–545.
7. **Feinberg, A. P., and B. Vogelstein.** 1984. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **137:**266–267.
8. **Gilbert, W.** 1978. Why genes in pieces? Nature (London) **271:**501.
9. **Göttlinger, H. G., I. Funke, J. P. Johnson, J. M. Gokel, and G. Riethmüller.** 1986. The epithelial cell surface antigen 17-1A, a target for antibody-mediated tumor therapy: its biochemical nature, tissue distribution and recognition by different monoclonal antibodies. Int. J. Cancer **38:**47–53.

10. **Haluska, F. H., K. Huebner, M. Isobe, T. Nishimura, C. M. Croce, and P. K. Vogt.** 1988. Localization of the human JUN protooncogene to chromosome region 1p31-32. Proc. Natl. Acad. Sci. USA **85:**2215–2219.
11. **Herlyn, D., M. Herlyn, A. H. Ross, C. Ernst, B. Atkinson, and H. Koprowski.** 1984. Efficient selection of human tumor growth-inhibiting monoclonal antibodies. J. Immunol. Methods **73:**157–167.
12. **Higgins, D. G., A. J. Bleasby, and R. Fuchs.** 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Appl. Biosci. **8:**189–191.
13. **Lehrach, H., D. Diamond, J. M. Wozney, and H. Boedtker.** 1977. RNA molecular weight determinations by gel electrophoresis under denaturing conditions, a critical reexamination. Biochemistry **16:**4743–4751.
14. **Linnenbach, A. J., D. S. Speicher, V. T. Marchesi, and B. G. Forget.** 1986. Cloning of a portion of the chromosomal gene for human erythrocyte α-spectrin by using a synthetic gene fragment. Proc. Natl. Acad. Sci. USA **83:**2397–2401.
15. **Linnenbach, A. J., J. Wojcierowski, S. Wu, J. J. Pyrc, A. Ross, B. Dietzschold, D. Speicher, and H. Koprowski.** 1989. Sequence investigation of the major gastrointestinal tumor-associated antigen gene family, GA733. Proc. Natl. Acad. Sci. USA **86:**27–31.
16. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227:**1435–1441.
17. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual, p. 196. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
18. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual, p. 326–328. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
19. **Morris, S., T. Druck, and K. Huebner.** Reassignment of the *ARH*H9 gene to chromosome region 1p. Blood, in press.
20. **Nagayoshi, T., D. Sanborn, N. J. Hickok, D. R. Olsen, M. J. Fazio, M.-L. Chu, R. Knowlton, K. Mann, R. Deutzman, R. Timpl, and J. Uitto.** 1989. Human nidogen: complete amino acid sequence and structural domains deduced from cDNAs, and evidence for polymorphism of the gene. DNA **8:**581–594.
21. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
22. **Perez, M. E., and L. E. Walker.** 1989. Isolation and characterization of a cDNA encoding the KS1/4 epithelial carcinoma marker. J. Immunol. **142:**3662–3667.
23. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
24. **Simon, B., D. K. Podolsky, G. Moldenhauer, K. J. Isselbacher, S. Gattoni-Celli, and S. T. Brand.** 1990. Epithelial glycoprotein is a member of a family of epithelial cell surface antigens homologous to nidogen, a matrix adhesion protein. Proc. Natl. Acad. Sci. USA **87:**2755–2759.
25. **Staden, R.** 1986. The current status and portability of our sequence handling software. Nucleic Acids Res. **14:**217–231.
26. **Strassburg, C. P., Y. Kasai, B. A. Seng, P. Miniou, J. Zaloudik, D. Herlyn, H. Koprowski, and A. J. Linnenbach.** 1992. Baculovirus recombinant expressing a secreted form of a transmembrane carcinoma-associated antigen. Cancer Res. **52:**815–821.
27. **Strauss, E. C., J. A. Kobori, G. Siu, and L. E. Hood.** 1986. Specific-primer-directed DNA sequencing. Anal. Biochem. **154:**353–360.
28. **Strnad, J., A. E. Hamilton, L. S. Beavers, G. C. Gamboa, L. D. Apelgren, L. D. Taber, J. R. Sportsman, T. F. Bumol, J. D. Sharp, and R. A. Gadski.** 1989. Molecular cloning and characterization of a human/epithelial cell surface antigen complementary DNA. Cancer Res. **49:**314–317.
29. **Szala, S., M. Froehlich, M. Scollon, Y. Kasai, Z. Steplewski, H. Koprowski, and A. J. Linnenbach.** 1990. Molecular cloning of cDNA for the carcinoma-associated antigen GA733-2. Proc. Natl. Acad. Sci. USA **87:**3542–3546.
30. **Szala, S., Y. Kasai, Z. Steplewski, U. Rodeck, H. Koprowski, and A. J. Linnenbach.** 1990. Molecular cloning of cDNA for the human tumor-associated antigen CO-029 and identification of related transmembrane antigens. Proc. Natl. Acad. Sci. USA **87:**6833–6837.
31. **Tsujimoto, Y., E. Jaffe, J. Cossman, J. Garham, P. C. Nowell, and C. M. Croce.** 1985. Clustering of breakpoints on chromosome 11 in human B-cell neoplasms with the t(11; 14) chromosome translocation. Nature (London) **315:**340–343.