J. Delbaere, L. A. Bauer, *J. Mol. Biol.* **144**, 43 (1980); W. A. Hendrickson and M. M. Teeter, *Nature* **290**, 107 (1981); J. Walter *et al.*, *Acta Crystallogr.* **B38**, 1462 (1982); T. A. Jones and S. Thirup, *EMBO J.* **5**, 819 (1986).
21. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981); G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer-Verlag, New York, 1979).
22. C. Chothia and A. V. Finkelstein, *Annu. Rev. Biochem.* **59**, 1007 (1990).
23. D. Eisenberg and A. D. McLachlan, *Nature* **319**, 199 (1986).
24. L. Cliche, L. M. Gregoret, F. E. Cohen, P. A. Kollman, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 3240 (1990).
25. S. Fahnestock, personal communication.
26. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
27. J. J. Langone, *Adv. Immunol.* **32**, 157 (1982); J. Deisenhofer, *Biochemistry* **20**, 2361 (1981); B. Nilsson *et al.*, *Protein Eng.* **1**, 107 (1987).
28. B. R. Brooks *et al.*, *J. Comput. Chem.* **4**, 187 (1983).
29. Supported by the AIDS Targeted Anti-Viral Program of the Office of the Director of the National Institutes of Health (G.M.C., A.M.G., and P.T.W.). We thank J. Richardson and R. Feldman for helpful discussions.

22 March 1991; accepted 10 May 1991

Identification of FAP Locus Genes from Chromosome 5q21

KENNETH W. KINZLER, MEF C. NILBERT, LI-KUO SU, BERT VOGELSTEIN,* TRACY M. BRYAN, DANIEL B. LEVY, KELLY J. SMITH, ANTONETTE C. PREISINGER, PHILIP HEDGE, DOUGLAS MCKECHNIE, RACHEL FINNIEAR, ALEX MARKHAM, JOHN GROFFEN, MARK S. BOGUSKI, STEPHEN F. ALTSCHUL, AKIRA HORII, HIROSHI ANDO, YASUO MIYOSHI, YOSHIO MIKI, ISAMU NISHISHO, YUSUKE NAKAMURA

Recent studies suggest that one or more genes on chromosome 5q21 are important for the development of colorectal cancers, particularly those associated with familial adenomatous polyposis (FAP). To facilitate the identification of genes from this locus, a portion of the region that is tightly linked to FAP was cloned. Six contiguous stretches of sequence (contigs) containing approximately 5.5 Mb of DNA were isolated. Subclones from these contigs were used to identify and position six genes, all of which were expressed in normal colonic mucosa. Two of these genes (APC and MCC) are likely to contribute to colorectal tumorigenesis. The MCC gene had previously been identified by virtue of its mutation in human colorectal tumors. The APC gene was identified in a contig initiated from the MCC gene and was found to encode an unusually large protein. These two closely spaced genes encode proteins predicted to contain coiled-coil regions. Both genes were also expressed in a wide variety of tissues. Further studies of MCC and APC and their potential interaction should prove useful for understanding colorectal neoplasia.

FAMILIAL ADENOMATOUS POLYPOSIS (FAP) is one of the most common autosomal dominant diseases leading to cancer predisposition, affecting nearly 0.01% of the American, British, and Japanese populations (*1*). Affected patients usually develop numerous benign colorectal tumors (polyps) that can progress to malignant forms. The first clue to the location of the gene responsible for FAP was provided by Herrera and colleagues, who demonstrated a constitutional deletion of chromosomal band 5q21 in an FAP patient (*2*). This cytogenetic observation stimulated linkage analyses that demonstrated that 5q21 chromosome markers were tightly linked to the development of polyps in numerous FAP kindreds (*3–5*). Other studies suggest that genes from the same region may be involved in tumorigenesis in kindreds with unusual forms of FAP (*1, 6*), as well as in patients with "sporadic" colorectal cancer (*7–9*). To facilitate the identification of the 5q21 gene or genes responsible for FAP and related disorders, we have cloned a relatively large region from 5q21 and identified several genes from within this region that are expressed in colorectal epithelium.

The cosmid markers YN5.64 and YN5.48 have previously been shown to delimit an 8-cM region containing the locus for FAP (*5*). Further linkage and pulse-field gel electrophoresis (PFGE) analysis with additional markers has shown that the FAP locus is contained within a 4-cM region bordered by cosmids EF5.44 and L5.99 (*10*). To isolate clones representing a significant portion of this locus, a yeast artificial chromosome (YAC) library was screened with 5q21 markers. Twenty-one YAC clones, distributed within six contigs and including 5.5 Mb from the region between YN5.64 and YN5.48, were obtained (Fig. 1A).

Three contigs encompassing approximately 4 Mb were contained within the central portion of this region (Fig. 1B). To initiate construction of each contig, the sequence of a genomic marker cloned from chromosome 5q21 was determined and used to design primers for amplification by the polymerase chain reaction (PCR) (*11*). PCR was then carried out on pools of YAC clones distributed in microtiter trays as described (*12*). Individual YAC clones from the positive pools were identified by further PCR or hybridization-based assays, and the YAC sizes were determined by PFGE. To extend the areas covered by the original YAC clones, "chromosomal walking" was done. YAC termini were isolated by a PCR-based method and sequenced (*13*). PCR primers based on these sequences were then used to rescreen the YAC library. Multipoint linkage analysis with the various markers used to define the contigs, combined with PFGE analysis, showed that contigs 1 and 2 were centromeric to contig 3.

Contig 1 contained the FER gene, which had previously been identified on the basis of its sequence similarity to the oncogene ABL (*14*). Linkage analysis and physical mapping with the YAC clones indicated that FER was tightly linked to previously defined polymorphic markers for the FAP locus (*15*). However, further analysis (*16*) did not indicate any FAP-specific mutations in this gene.

A cross-hybridization approach was used to identify TB1 in contig 2; in this procedure, potential exon sequences are identified by cross-hybridization between human and rodent DNAs (*17–19*). Subclones of all the cosmids shown in Fig. 1 were used to screen Southern blots containing rodent DNA samples. A subclone of cosmid N5.66 was shown to strongly hybridize to rodent DNA, and this clone was used to screen cDNA libraries derived from normal adult colon and fetal liver. The ends of the initial cDNA clones obtained in this screen were then used to extend the cDNA sequence. Sequence analysis of 11 overlapping cDNA clones revealed an open reading frame (ORF) that extended for 1303 bp starting from the most 5′ sequence data obtained (GenBank accession number M74089). The predicted product of this gene (TB1) contained two significant local similarities to a

K. W. Kinzler, M. C. Nilbert, L.-K. Su, B. Vogelstein, T. M. Bryan, D. B. Levy, K. J. Smith, A. C. Preisinger, Molecular Genetics Laboratory, The Johns Hopkins University School of Medicine, Baltimore, MD 21231.
P. Hedge, D. McKechnie, R. Finniear, A. Markham, ICI Pharmaceuticals, Cheshire, United Kingdom SK10 4TG.
J. Groffen, Department of Pathology, Children's Hospital of Los Angeles, Los Angeles, CA 90027.
M. S. Boguski and S. F. Altschul, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894.
A. Horii, H. Ando, Y. Miyoshi, Y. Miki, I. Nishisho, Y. Nakamura, Department of Biochemistry, Cancer Institute, Tokyo 170, Japan.

*To whom correspondence should be addressed.

family of ADP, ATP carrier/translocator proteins (*20, 21*). Other than its localization to the FAP region (Fig. 1), we found no other evidence linking TB1 to colorectal tumorigenesis in either sporadic or familial cases (*16*).

The MCC gene, which had also been discovered through a cross-hybridization approach, was considered a candidate for causing FAP by virtue of its tight genetic linkage to FAP susceptibility and its somatic mutation in sporadic colorectal carcinomas (*9*). However, mapping experiments suggested that the coding region of MCC was approximately 50 kb proximal to the centromeric end of a 200-kb deletion found in an FAP patient (*22*). MCC cDNA probes detected a major 10-kb mRNA transcript on Northern blot analysis (Fig. 2), of which 4151 bp, including the entire ORF, have been cloned (*9, 10*). Although the 3′ non-translated portion or an alternatively spliced form of MCC might have extended into this deletion, it was possible that the deletion did not affect the MCC gene product. We therefore used MCC sequences to initiate a YAC contig, and subsequently used the YAC clones to identify genes 50 to 250 kb distal to MCC that might be contained within the deletion.

In a first approach, the insert from YAC 24ED6 (contig 3, Fig. 1B) was radioactively labeled and hybridized to a cDNA library from normal colon (*23*). One of the cDNA clones (YS39) identified in this manner detected a 3.1-kb mRNA transcript when used as a probe for Northern blot hybridization (Fig. 2). Sequence analysis of YS39 and eight overlapping cDNA clones revealed that they encompassed 2322 nucleotides and contained an ORF that extended for 593 bp from the most 5′ sequence data obtained (GenBank accession number M74090). If the entire ORF were translated, it would encode 197 amino acids. Searches of nucleotide and protein databases revealed that the gene detected by YS39, TB2, was not identical to any previously reported sequences nor were there any striking similarities (*20*). In contrast to MCC, we found no evidence linking TB2 to colorectal tumorigenesis (*16*).

Another clone (YS11), identified through the YAC 24ED6 screen, appeared to contain portions of two distinct genes. Sequences from one end of YS11 were identical to at least 180 bp of the signal recognition particle protein SRP19 (*24*). A second ORF, from the opposite end of clone YS11, proved to be identical to 78 bp of a gene that was independently identified by a second YAC-based approach. For the latter, DNA from yeast cells containing YAC 14FH1 (Fig. 1B) was digested with Eco RI and subcloned into a plasmid vector. Plasmids that contained human DNA fragments were selected by colony hybridization with total human DNA as a probe. These clones were then used to search for cross-hybridizing sequences, and the cross-hybridizing clones were subsequently used to screen cDNA libraries. One of the cDNA clones discovered in this way (FH38) contained a long ORF (2496 bp), 78 bp of which were identical to sequences in YS11 (*25*). The ends of the FH38 cDNA clone were then used to initiate cDNA walking to extend the sequence. Eventually, 112 cDNA clones were isolated from normal colon, brain, and liver cDNA libraries and found to encompass 8972 nucleotides of contiguous tran-

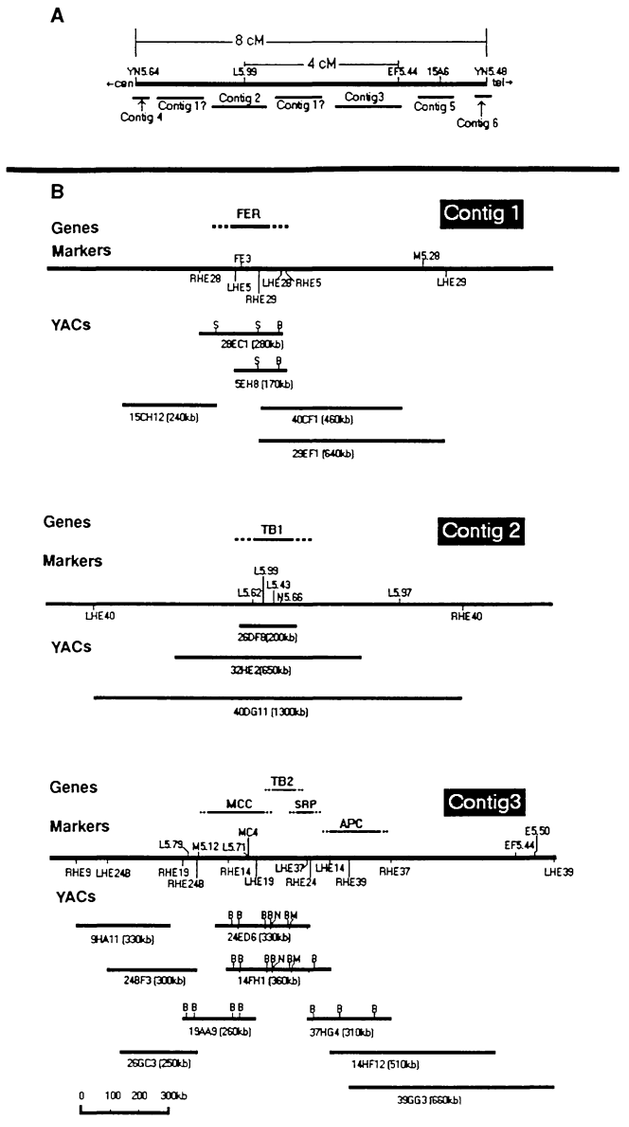

Fig. 1. (**A**) Overview of YAC contigs. Genetic distances between selected RFLP markers from within the contigs are shown in centiMorgans. (**B**) Detailed map of the three central contigs. The YAC contigs were obtained as described in the text, by means of the indicated markers. The positions of the six identified genes from within the FAP region are shown; the 5′ and 3′ ends of the transcripts from these genes have, in general, not yet been exactly determined, as indicated by the string of dots surrounding the bars denoting the gene positions. Contigs 1, 2, and 3 were initiated from sequences derived from FER, cosmid N5.66, and MCC, respectively. The markers indicated by RHE and LHE were derived from YAC termini, as described (*12*). The other markers were derived from sequencing cosmid subclones, except for MC4 and FE3, which were derived from genomic phage subclones of the MCC and FER genes, respectively (*11*). Linkage and PFGE analysis demonstrated that contig 3 was telomeric to contigs 1 and 2 and that contig 3 was oriented with its left and right ends (as drawn) closest to the centromere and telomere of chromosome 5, respectively. The relative orientations of contigs 1 and 2 with respect to each other and to the centromere and telomere could not be determined with certainty, as indicated by question marks. In contig 2, the four cosmid markers L5.62, N5.66, L5.43, and L5.99 were shown by cross-hybridization to lie within 100 kb of one another; however, the illustrated order of these four cosmids and the position of the three contig 2 YACs with respect to them is only approximate. B, Bss H2; S, Sst II; M, Mlu I; N, Nru I.

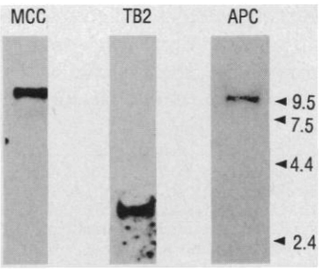

Fig. 2. Northern blot analysis of contig 3 genes. Total RNA (10 μg) was separated by electrophoresis through a denaturing gel and transferred to nylon membranes. The RNA on the Northern blots was then hybridized with cDNA probes for the indicated genes (*42*). The sizes of the transcripts identified were determined by the position of co-electrophoresed markers (arrowheads; kilobases).

```
MAAASYDQLL KQVEALKMEN SNLRQELEDN SNHLTKLETE ASNMKEVLKQ LQGSIEDEAM   60
ASSGQIDLLE RLKELNLDSS NFPGVKLRSK MSLRSYGSRE GSVSSRSGEC SPVPMGSFPR  120
RGFVNGSRES TGYLEELEKE RSLLLADLDK EEKEKDWYYA QLQNLTKRID SLPLTENFSL  180
QTDMTRRQLE YEARQIRVAM EEQLGTCQDM EKRAQRRIAR IQQIEKDILR IRQLLQSQAT  240
EAERSSQNKH ETGSHDAERQ NEGQGVGEIN MATSGNGQGS TTRMDHETAS VLSSSSTHSA  300
PRRLTSHLGT KVEMVYSLLS MLGTHDKDDM SRTLLAMSSS QDSCISMRQS GCLPLLIQLL  360
HGNDKDSVLL GNSRGSKEAR ARASAALHNI IHSQPDDKRG RREIRVLHLL EQIRAYCETC  420
WEWQEAHEPG MDQDKNPMPA PVEHQICPAV CVLMKLSFDE EHRHAMNELG GLQAIAELLQ  480
VDCEMYGLTN DHYSITLRRY AGMALTNLTF GDVANKATLC SMKGCMRALV AQLKSESEDL  540
QQVIASVLRN LSWRADVNSK KTLREVGSVK ALMECALEVK KESTLKSVLS ALWNLSAHCT  600
ENKADICAVD GALAFLVGTL TYRSQTNTLA IIESGGGILR NVSSLIATNE DHRQILRENN  660
CLQTLLQHLK SHSLTIVSNA CGTLWNLSAR NPKDQEALWD MGAVSMLKNL IHSKHKMIAM  720
GSAAALRNLM ANRPAKYKDA NIMSPGSSLP SLHVRKQKAL EAELDAQHLS ETFDNIDNLS  780
PKASHRSKQR HKQSLYGDYV FDTNRHDDNR SDNFNTGNMT VLSPYLNTTV LPSSSSSRGS  840
LDSSRSEKDR SLERERGIGL GNYHPATENP GTSSKRGLQI STTAAQIAKV MEEVSAIHTS  900
QEDRSSGSTT ELHCVTDERN ALRRSSAAHT HSNTYNFTKS ENSNRTCSMP YAKLEYKRSS  960
NDSLNSVSSS DGYGKRGQMK PSIESYSEDD ESKFCSYGQY PADLAHKIHS ANHMDDNDGE 1020
LDTPINYSLK YSDEQLNSGR QSPSQNERWA RPKHIIEDEI KQSEQRQSRN QSTTYPVYTE 1080
STDDKHLKFQ PHFGQQECVS PYRSRGANGS ETNRVGSNHG INQNVSQSLC QEDDYEDDKP 1140
TNYSERYSEE EQHEEEERPT NYSIKYNEEK RHVDQPIDYS LKYATDIPSS QKQSFSFSKS 1200
SSGQSSKTEH MSSSSENTST PSSNAKRQNQ LHPSSAQSRS GQPQKAATCK VSSINQETIQ 1260
TYCVEDTPIC FSRCSSLSSL SSAEDEIGCN QTTQEADSAN TLQIAEIKEK IGTRSAEDPV 1320
SEVPAVSQHP RTKSSRLQGS SLSSESARHK AVEFSSGAKS PSKSGAQTPK SPPEHYVQET 1380
PLNFSRCTSV SSLDSFESRS IASSVQSEPC SGMVSGIISP SDLPDSPGQT MPPSRSKTPP 1440
PPPQTAQTKR EVPKNKAPTA EKRESGPKQA AVNAAVQRVQ VLPDADTLLH FATESTPDGF 1500
SCSSSLSALS LDEPFIQKDV ELRIMPPVQE NDNGNETESE QPKESNENQE KEAEKTIDSE 1560
KDLLDDSDDD DIEILEECII SAMPTKSSRK AKKPAQTASK LPPPVARKPS QLPVYKLLPS 1620
QNRLQPQKHV SFTPGDDMPR VYCVEGTPIN FSTATSLSDL TIESPPNELA AGEGVRGGAQ 1680
SGEFEKRDTI PTEGRSTDEA QGGKTSSVTI PELDDNKAEE GDILAECINS AMPKGKSHKP 1740
FRVKKIMDQV QQASASSSAP NKNQLDGKKK KPTSPVKPIP QNTEYRTRVR KNADSKNNLN 1800
AERVFSDNKD SKKQNLKNNS KDFNDKLPNN EDRVRGSFAF DSPHHYTPIE GTPYCFSRND 1860
SLSSLDFDDD DVDLSREKAE LRKAKENKES EAKVTSHTEL TSNQQSANKT QAIAKQPINR 1920
GQPKPILQKQ STFPQSSKDI PDRGAATDEK LQNFAIENTP VCFSHNSSLS SLSDIDQENN 1980
NKENEPIKET EPPDSQGEPS KPQASGYAPK SFHVEDTPVC FSRNSSLSSL SIDSEDDLLQ 2040
ECISSAMPKK KKPSRLKGDN EKHSPRNMGG ILGEDLTLDL KDIQRPDSEH GLSPDSENFD 2100
WKAIQEGANS IVSSLHQAAA AACLSRQASS DSDSILSLKS GISLGSPFHL TPDQEEKPFT 2160
SNKGPRILKP GEKSTLETKK IESESKGIKG GKKVYKSLIT GKVRSNSEIS GQMKQPLQAN 2220
MPSISRGRTM IHIPGVRNSS SSTSPVSKKG PPLKTPASKS PSEGQTATTS PRGAKPSVKS 2280
ELSPVARQTS QIGGSSKAPS RSGSRDSTPS RPAQQPLSRP IQSPGRNSIS PGRNGISPPN 2340
KLSQLPRTSS PSTASTKSSG SGKMSYTSPG RQMSQQNLTK QTGLSKNASS IPRSESASKG 2400
LNQMNNGNGA NKKVELSRMS STKSSGGSESD RSERPVLVRQ STFIKEAPSP TLRRKLEESA 2460
SFESLSPSSR PASPTRSQAQ TPVLSPSLPD MSLSTHSSVQ AGGWRKLPPN LSPTIEYNDG 2520
RPAKRHDIAR SHSESPSRLP INRSGTWKRE HSKHSSSLPR VSTWRRTGSS SSILSASSES 2580
SEKAKSEDEK HVNSISGTKQ SKENQVSAKG TWRKIKENEF SPTNSTSQTV SSGATNGAES 2640
KTLIYQMAPA VSKTEDVWVR IEDCPINNPR SGRSPTGNTP PVIDSVSEKA NPNIKDSKDN 2700
QAKQNVGNGS VPMRTVGLEN RLNSFIQVDA PDQKGTEIKP GQNNPVPVSE TNESSIVERT 2760
PFSSSSSSKH SSPSGTVAAR VTPFNYNPSP RKSSADSTSA RPSQIPTPVN NNTKKRDSKT 2820
DSTESSGTQS PKRHSGSYLV TSV                                          2843
```

Fig. 3. Predicted amino acid sequence of the APC gene. The cDNA sequence was determined through the analysis of 112 cDNA clones derived from colon, liver, and brain. A total of 8972 bp were contained within overlapping cDNA clones, defining an ORF of 2843 amino acids. The nucleotide sequence has been deposited with GenBank accession number M74088.

```
A
APC      204  LGTCQDMEKRAQRRIARIQQIEKDILRIRQL  234
              |   :: || ||||||:|   |   |
RAL2     576  LTGAKGLQLRALRRIARIEQGGTAISPTSPL  606

B
APC      454  MKLSFDEEHRHAMNELGGLQAIAELLQVD  482
              :    | :  ||:|||| :  : :
M3 MACHR 249  LYWRIYKETEKRTKELAGLQASGTEAETE  277
              || : | :  ||||||
MCC      220  LYPNLAEERSRWEKELAGLREENESLTAM  248
              :   ||: :  ||:||   | |
APC      454  MKLSFDEEHRHAMNELGGLQAIAELLQVD  482
```

Fig. 4. (**A**) Local similarity between APC and *ral2*. (**B**) Local similarity among the APC and MCC genes and the m3 mAChR. The connecting lines indicate identities; dots indicate related amino acid residues.

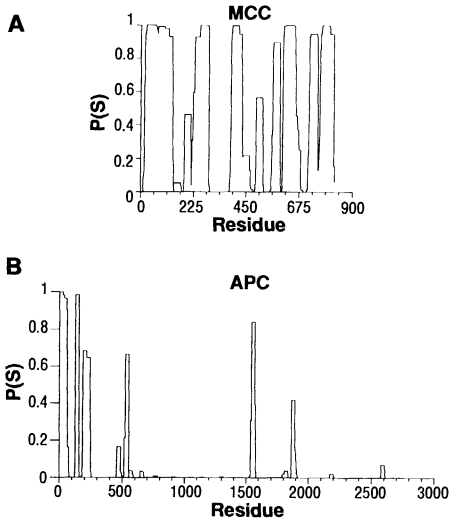

Fig. 5. Probabilities of forming coiled-coils in MCC and APC gene products. The program of Lupas and colleagues (*39*) with a window of 28 residues was used to calculate the probabilities of coiled-coil conformation. Values on the *Y* and *X* axes represent probabilities and codon numbers, respectively. The predicted MCC gene product is depicted in (**A**) and that of the APC gene product in (**B**).

script (GenBank accession number M74088). As demonstrated in the accompanying paper (*16*), mutations of the gene corresponding to this transcript, APC (adenomatous polyposis coli), were found in the germ line of FAP patients as well as in sporadic colorectal cancers. When used as probes for Northern blot analysis, APC cDNA clones hybridized to a single transcript of approximately 9.5 kb (Fig. 2), suggesting that the great majority of the gene product was represented in the cDNA clones obtained. Sequences from the 5′ end of the APC gene were found in YAC 37HG4 but not in YAC 14FH1. However, the 3′ end of the APC gene was found in both YACs. Analogously, the 5′ end of the MCC coding region was found in YAC clones 19AA9 and 26GC3 but not in YAC clones 24ED6 or 14FH1, while the 3′ end displayed the opposite pattern. Thus, MCC and APC transcription units point in opposite directions, with the direction of transcription going from centromeric to telomeric in the case of MCC and telomeric to centromeric in the case of APC. PFGE analysis of YAC DNA digested with various restriction endonucleases showed that TB2 and SRP were between MCC and APC, and that the 3′ ends of the coding regions of MCC and APC were separated by approximately 150 kb (Fig. 1B).

Sequence analysis of the APC cDNA clones revealed an ORF of 8538 nucleotides. The 5′ end of the ORF contained a methionine codon (codon 1) that was preceded by an in-frame stop codon 9 bp upstream, and the 3′ end was followed by several in-frame stop codons. The protein produced by initiation at codon 1 would contain 2843 amino acids (Fig. 3). The results of database searches (*20*) with the APC gene product were quite complex because of the presence of large segments with locally biased amino acid compositions. In spite of this, APC could be roughly divided into two domains. The NH_2-terminal 25% of the protein had a high content of leucine residues (12%) and showed local sequence similarities to myosins, intermediate filament proteins (such as desmin, vimentin, and neurofilaments) and *Drosophila* armadillo protein (human plakoglobin) (*26, 27*). The COOH-terminal 75% of APC (residues 731 to 2832) was 17% serine by composition with serine residues more or less uniformly distributed. This large domain also contains local concentrations of charged (mostly acidic) and proline residues (*28*). There was no indication of potential signal peptides, transmembrane regions, or nuclear targeting signals in APC, which suggests a cytoplasmic localization.

To detect short similarities to APC, a database search was done with the PAM-40 matrix (*29*). Potentially interesting matches to several proteins were found (*30*). The most suggestive of these involved the *ral2* gene product of yeast, which is implicated in the regulation of *ras* activity (*31*); the local alignment is shown in Fig. 4A. Little is known about how *ral2* might interact with ras but the positively charged character of this region is interesting in the context of the negatively charged GAP interaction region of *ras* (*32, 33*).

Because of the proximity of the MCC and APC genes, and the fact that both were implicated in colorectal tumorigenesis (*9, 16*), we searched for similarities between the two predicted proteins. MCC shares a short similarity with the region of the m3 muscarinic acetylcholine receptor (mAChR) known to regulate specificity of G protein coupling (*9, 34*). The APC gene also contained local similarity to the region of the m3 mAChR that overlapped with the MCC similarity (Fig. 4B). Although the similarities to *ral2* and m3 mAChR were not statistically significant, they were intriguing in light of previous observations relating G proteins to neoplasia (*35–37*). Additionally, MCC has the potential to form α-helical coiled-coils (*38*). We used a program that predicts coiled-coil potential from primary

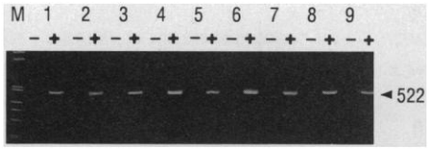

Fig. 6. Exon connection analysis of the APC gene. For each RNA sample, cDNA synthesis was performed in the presence (+) or absence (−) of reverse transcriptase (*41*). The cDNA was generated from normal colon (lanes 1); several colorectal cancer (CRC) cell lines (*43*) (lanes 2, Difi; lanes 3, SW948; lanes 4, LoVo; lanes 5, HT29; lanes 6, HCT116; lanes 7, SW480); the Hut 82 lung tumor cell line (lanes 8) and the SV-HUC bladder tumor cell line (*44*) (lanes 9). The PCR products were separated by nondenaturing polyacrylamide gel electrophoresis and stained with ethidium bromide. Size is shown in bases pairs.

sequence data (*39*) to analyze both MCC and APC. Analysis of MCC indicated a discontinuous pattern of coiled-coil domains separated by putative "hinge" or "spacer" regions similar to those seen in laminin and other intermediate filament (IF) proteins (Fig. 5A). Analysis of the APC sequence revealed two regions in the NH_2-terminal domain that had strong coiled-coil potential (Fig. 5B), and these regions corresponded to those that showed local similarities with myosin and IF proteins on database searching. In addition, one other putative coiled-coil region was identified in the central portion of APC (Fig. 5B). The potential for both APC and MCC to form coiled-coils is interesting in that such structures often mediate homo- and hetero-oligomerization (*40*).

Each of the six genes described above was expressed in normal colon, as indicated by their representation in colon cDNA libraries. The MCC gene has already been shown to be expressed in a variety of tissue types (*9*). Reverse transcription–polymerase chain reaction (RT-PCR) assays (*41*) were used to demonstrate that APC was expressed in normal colon mucosa, in eight cell lines derived from sporadic colorectal cancer patients, and in two FAP colorectal cancer cell lines (examples in Fig. 6). Similarly, APC was found to be expressed in human fetal muscle, liver, and skin; in adult peripheral white blood cells; lymphoblasts immortalized by Epstein-Barr virus, and SV40-immortalized glial cells; and in cell lines derived from normal fibroblasts (WI38), a myeloid leukemia (HL60), an osteosarcoma (HOS), a fibrosarcoma (HT1080), a teratocarcinoma (Tera-2), a lung carcinoma (H82), a bladder carcinoma (SV-HUC), and a thyroid carcinoma (TT) (examples in Fig. 6).

In summary, we have cloned a small part of the human genome corresponding to the locus linked to FAP. Concurrently, we identified several genes that map within this region and showed them to be expressed in normal colon. The hypothesis underlying these cloning, mapping, and expression studies was that one or more of these genes would be responsible for the inheritance of FAP and other conditions associated with the predisposition to colorectal neoplasia. The identification of these genes allowed testing of this hypothesis, as described in the accompanying paper (*16*).

REFERENCES AND NOTES

1. J. Utsunomiya and H. T. Lynch, Eds., *Hereditary Colorectal Cancer* (Springer-Verlag, New York, 1990).
2. L. Herrera *et al.*, *Am. J. Med. Genet.* **25**, 473 (1986).
3. W. F. Bodmer *et al.*, *Nature* **328**, 614 (1987); M. Leppert *et al.*, *Science* **238**, 1411 (1987).
4. M. G. Dunlop, C. M. Steel, A. H. Wyllie, C. C. Bird, H. J. Evans, *Genomics* **5**, 350 (1989); P. Meera Khan *et al.*, *Hum. Genet.* **79**, 183 (1989).
5. Y. Nakamura *et al.*, *Am. J. Hum. Genet.* **43**, 638 (1988).
6. M. Leppert *et al.*, *N. Engl. J. Med.* **322**, 904 (1990).
7. E. R. Fearon and B. Vogelstein, *Cell* **61**, 759 (1990).
8. E. Solomon *et al.*, *Nature* **328**, 616 (1987); M. Sasaki *et al.*, *Cancer Res.* **49**, 4402 (1989); P. Delattre *et al.*, *Lancet* **ii**, 353 (1989); P. G. Ashton-Rickardt *et al.*, *Oncogene* **4**, 1169 (1989); B. Vogelstein *et al.*, *N. Engl. J. Med.* **319**, 525 (1988).
9. K. W. Kinzler *et al.*, *Science* **251**, 1366 (1991).
10. K. W. Kinzler *et al.*, unpublished data.
11. The sequence of PCR primers used to obtain the YACs shown in Fig. 1 are available from the authors upon request or electronically from the NCBI data repository by anonymous ftp from the pub/apc directory at ncbi.nim.nih.gov (numerical address 130.14.20.1).
12. R. Anand, J. H. Riley, P. Butler, J. C. Smith, A. F. Markham, *Nucleic Acids Res.* **18**, 1951 (1990).
13. J. H. Riley *et al.*, *ibid.*, p. 2887.
14. Q. L. Hao, N. Heisterkamp, J. Groffen, *Mol. Cell. Biol.* **9**, 1587 (1989); C. Morris, N. Heisterkamp, Q. L. Hao, J. Groffen, *Cytogenet. Cell. Genet.* **53**, 4 (1990).
15. A genomic probe specific for the 3′ end of FER (pMN2.3) was used to define a restriction fragment length polymorphism (RFLP) by screening human genomic DNA samples cleaved with restriction endonucleases. Msp I was found to provide a two-allele polymorphism with alleles of 4.2 kb and 1.9 kb at frequencies of 0.6 and 0.4, plus a constant band at 5.8 kb. Linkage analysis was done on three-generation reference families from the CEPH panel and analyzed by the LINKAGE program. This analysis suggested that FER was between YN5.64 (Lod score of 4.8 at $\theta = 0.056$) and YN5.48 (Lod score of 4.0 at $\theta = 0.11$), and this location was confirmed by PFGE analysis.
16. I. Nishisho *et al.*, *Science* **253**, 665 (1991).
17. E. R. Fearon *et al.*, *ibid.* **247**, 49 (1990).
18. A. P. Monoco *et al.*, *Nature* **323**, 646 (1986).
19. J. M. Rommens *et al.*, *Science* **245**, 1059 (1989).
20. The National Center for Biotechnology Information (NCBI) maintains a composite database consisting of all nonidentical sequences from the following databases: NBRF/PIR (Release 28.0), SWISS-PROT (Release 18.0), GenPept (translated GenBank, Release 64.3), GenPept (daily update), and NCBI's GenInfo Backbone (daily update). At the time of submission, this composite database contained 14,020,527 residues in 51,948 sequences and was used for all searches described in the present work. Unless otherwise specified, all data searches were performed with the BLAST family of programs [S. F. Altschul *et al.*, *J. Mol. Biol.* **215**, 403 (1990)] and the PAM-120 scoring matrix [M. O. Dayhoff, Ed., *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, D.C., 1978); S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991).
21. The TB1 amino acid sequence showed significant ($P < 0.05$) local similarities to the following sequences in the NBRF/PIR database: S03894, B28116, A32446, A24363, A29278, C28116, and A24849.
22. G. Joslyn *et al.*, *Cell*, in press.
23. P. Elvin *et al.*, *Nucleic Acids Res.* **18**, 3913 (1990); M. R. Wallace *et al.*, *Science* **249**, 181 (1990).
24. K. Lingelbach *et al.*, *Nucleic Acids Res.* **16**, 9431 (1988).
25. The complex structure of the YS11 clone appeared to result from an abnormal splicing event between two adjacent genes; cDNA library construction artifacts would not be expected to result in the joining of transcripts from two adjacent genes into a single cDNA. The SRP 19 sequences and APC sequences of clone YS11 were independently amplified by PCR and both shown to reside within contig 3. Moreover, the 78 nucleotides of YS11 contained within APC diverged from the APC cDNA sequence at a known splice site (*16*). The sequences separating APC and SRP19 in YS11 were of unknown derivation. Unusual splicing patterns joining widely spaced exons from the same chromosomal segment have been found previously [J. Nigro *et al.*, *Cell* **64**, 607 (1991)].
26. The regons of APC showing the strongest similarities to intermediate filament (IF) and myosin-like proteins included codons 7 to 72 and, to a lesser extent, codons 185 to 227.
27. The region of APC showing the strongest similarity to the *armadillo* protein and plakoglobin was contained within codons 661 to 706. This latter similarity was less significant in a statistical sense but was of potential biological relevance because intermediate filament proteins interact with desmosome components [W. Franke *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4027 (1989); M. Peifer and E. Wieschaus, *Cell* **63**, 1167 (1990)].
28. Serine-rich regions of APC gave multiple spurious matches to vitellogenins and other serine-rich proteins. Similarly, acidic stretches of APC yielded pseudo-homologies with nucleolin, prothymosin α, and other proteins with similarly biased compositions. Rigorous methods for studying charge clusters in proteins identified two acidic runs (residues 1131 to 1156 and 1558 to 1577) and a mixed negative/positive charge cluster (residues 1866 to 1893) in APC [S. Karlin and V. Brendel, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5698 (1989)].
29. S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991).
30. The three highest scoring matches were to a *Caenorhabditis elegans* protein involved in embryogenesis [P. W. Carter, J. M. Roos, K. J. Kemphues, *Mol. Gen. Genet.* **221**, 72 (1990)], a wheat, sequence-specific, DNA binding protein [S. T. Tabata *et al.*, *Science* **245**, 967 (1989)], and *ral2* (*31*).
31. Y. Fukui, S. Miyake, M. Satoh, M. Yamamoto, *Mol. Cell. Biol.* **9**, 5617 (1989).
32. H. R. Bourne, D. A. Sanders, F. McCormick, *Nature* **349**, 117 (1990).
33. Y. Wang, M. S. Boguski, M. Riggs, L. Rodgers, M. Wigler, *Cell Regul.* **2**, 453 (1991).
34. J. Lechleiter *et al.*, *EMBO J.* **9**, 4381 (1990).
35. M. Barbacid, *Annu. Rev. Biochem.* **56**, 779 (1987); H. R. Bourne, D. A. Sanders, F. McCormick, *Nature* **348**, 125 (1990); M. H. Noda *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 162 (1989).
36. C. Landis *et al.*, *Nature* **340**, 692 (1989); J. Lyons *et al.*, *Science* **249**, 655 (1990).
37. G. Tu *et al.*, *Cell* **63**, 835 (1990); G. A. Martin *et al.*, *ibid.*, p. 843; R. Ballester *et al.*, *ibid.*, p. 851.
38. H. R. Bourne, *Nature* **351**, 188 (1991).
39. A. Lupas, M. Van dyke, J. Stock, *Science* **252**, 1162 (1991).
40. C. Cohen and D. A. D. Parry, *Proteins Struct. Funct. Genet.* **7**, 1 (1990).
41. RNA was purified as described [P. Chomczynski and N. Sacchi, *Anal. Biochem.* **162**, 156 (1987)]. cDNA was generated as described [E. Noonan and I. B. Roninson, *Nucleic Acids Res.* **162**, 10366 (1988)]. The cDNA was used in PCR as described (*S. J. Baker et al.*, *Cancer Res.* **50**, 7717 (1990)], with the following primers: 5′-TGGCACTCT-TACTTACCGG-3′ and 5′-GTCTCTGCTTAC-TACGATG-3′, corresponding to nt 1851 to 2372 of APC.

42. Northern blots were performed as described in K. W. Kinzler, J. M. Ruppert, S. H. Bigner, B. Vogelstein, *Nature* **332**, 371 (1988).
43. The two CRC cell lines from FAP patients were JW [C. Paraskeva, B. G. Buckle, D. Sheer, C. B. Wigley, *Int. J. Cancer* **34**, 49 (1984)] and DiFi [M. E. Gross *et al.*, *Cancer Res.* **51**, 1452 (1991)]. The eight sporadic CRC cell lines examined were SW948, LoVo, HT29, HCT116, SW480, SW1116, SW403 (obtained from the American Type Culture Collection, Rockville, MD), and KS (obtained from C. Paraskeva).
44. S.-Q. Wu *et al.*, *Cancer Res.* **51**, 3323 (1991).
45. Supported in part by grants from the Minister of Education, Culture and Science, the Princess Takamatsu Cancer Research Fund, the Uehara Memorial Foundation, the Kato Memorial Bioscience Foundation, the Damon Runyon–Walter Winchell Cancer Research Fund, the Clayton Fund, The McAshan Fund, and NIH grants CA35494, CA06973, CA44688, CA47527, CA41183, CA47456, GM07309, and GM07184. The authors thank T. Gwiazda for expert preparation of the manuscript; J. M. Jessup, J. Willson, M. Brattain, C. Paraskeva, J. Trent, C. Reznikoff, B. Boman, and M. Schwab for providing tumors and cell lines; A. Lupas for providing his program for analyzing coiled-coil proteins; and R. White and colleagues for helpful discussions and for exchanging information prior to publication.

13 June 1991; accepted 3 July 1991

Mutations of Chromosome 5q21 Genes in FAP and Colorectal Cancer Patients

ISAMU NISHISHO, YUSUKE NAKAMURA,* YASUO MIYOSHI, YOSHIO MIKI, HIROSHI ANDO, AKIRA HORII, KUMIKO KOYAMA, JOJI UTSUNOMIYA, SHOZO BABA, PHILIP HEDGE, ALEX MARKHAM, ANNE J. KRUSH, GLORIA PETERSEN, STANLEY R. HAMILTON, MEF C. NILBERT, DANIEL B. LEVY, TRACY M. BRYAN, ANTONETTE C. PREISINGER, KELLY J. SMITH, LI-KUO SU, KENNETH W. KINZLER, BERT VOGELSTEIN

Previous studies suggested that one or more genes on chromosome 5q21 are responsible for the inheritance of familial adenomatous polyposis (FAP) and Gardner's syndrome (GS), and contribute to tumor development in patients with noninherited forms of colorectal cancer. Two genes on 5q21 that are tightly linked to FAP (MCC and APC) were found to be somatically altered in tumors from sporadic colorectal cancer patients. One of the genes (APC) was also found to be altered by point mutation in the germ line of FAP and GS patients. These data suggest that more than one gene on chromosome 5q21 may contribute to colorectal neoplasia, and that mutations of the APC gene can cause both FAP and GS. The identification of these genes should aid in understanding the pathogenesis of colorectal neoplasia and in the diagnosis and counseling of patients with inherited predispositions to colorectal cancer.

RECENT STUDIES HAVE REVEALED that the accumulation of several genetic changes is associated with colorectal tumorigenesis (*1*). Because most colorectal cancers (CRCs) arise from benign adenomatous polyps (*2*), it is of great importance to identify the genes responsible for adenoma formation. Familial adenomatous polyposis (FAP) is one of the most common autosomal-dominant diseases leading to cancer predisposition, affecting 1 in 5,000 and 1 in 17,000 of the American and Japanese populations, respectively (*3*). Affected individuals usually develop hundreds to thousands of adenomatous polyps of the colon and rectum, a small fraction of which will progress to carcinoma if not surgically treated. Gardner's syndrome (GS) is a variant of FAP in which desmoid tumors, osteomas, and other neoplasms occur together with multiple adenomas of the colon and rectum.

The gene (or genes) responsible for FAP has been assigned to chromosome 5q21 by cytogenetic and linkage analysis (*4–7*). Although FAP is a relatively rare cause of colorectal cancer (*3*), the importance of 5q21 genes has been accentuated by the finding that 5q21 alleles are often lost from the tumors of sporadic colorectal cancer patients (that is, those without obvious inherited predispositions) (*1, 8*). Moreover, losses of 5q21 alleles are the earliest genetic alterations yet identified in sporadic colorectal neoplasms, present in adenomas as small as 5 mm in diameter (*9*). Such losses are generally thought to indicate the presence of a tumor suppressor gene in the deleted region (*10, 11*).

We have used probes from within the region tightly linked to FAP as tools to search for genes expressed in normal colonic mucosa. As described in the accompanying paper (*12*), six such genes were identified by means of cosmid and YAC clones. We considered the gene FER as a candidate because of its proximity to the FAP locus as judged by physical and genetic criteria (*12, 13*), and its homology to known tyrosine kinases with oncogenic potential (*14*). Primers were designed to amplify the complete coding sequence of FER from the RNA of two colorectal cancer cell lines derived from FAP patients (*15*). The resultant 2554-bp fragments were cloned and sequenced in their entirety (*16*). Only a single conservative amino acid change (GTG → CTG, creating a valine to leucine substitution at codon 439) was observed. The region surrounding this codon was then amplified from the DNA of individuals without FAP and this substitution was found to be a common polymorphism, not specifically associated with FAP (*17*). On the basis of these results, we considered it unlikely (though still possible) that the FER gene was responsible for FAP.

We next turned to the analysis of the four genes (MCC, TB2, SRP, and APC) in contig 3 (*12*). These genes were considered as

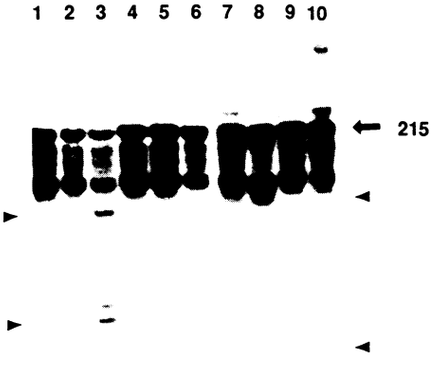

Fig. 1. PCR and RNase protection analysis of the APC gene in FAP patients. RNase protection analysis was performed on PCR products as described (*21*) and the resulting cleavage products separated by denaturing gel electrophoresis. The amplified genomic fragments containing APC exon nt 835 to 933 were 215 bp in length and contained codons 278 to 311. Lanes 1 to 10 show the results obtained from constitutional DNA of ten FAP patients. Lanes 3 (P24) and 8 (P93) show abnormal RNase cleavage products (arrowheads). Subsequent sequence analysis revealed that the abnormal patterns resulted from a C to T transition in P24 and a C to G transition in P93 (Table 1).

I. Nishisho, Y. Nakamura, Y. Miyoshi, Y. Miki, H. Ando, A. Horii, K. Koyama, Department of Biochemistry, Cancer Institute, Tokyo 170, Japan.
J. Utsunomiya, The Second Department of Surgery, Hyogo Medical College, Hyogo 663, Japan.
S. Baba, The Second Department of Surgery, Hamamatsu Medical College, Shizuoka 431-31, Japan.
P. Hedge and A. Markham, ICI Pharmaceuticals, Cheshire, United Kingdom SK10 4TG.
A. J. Krush, Center for Medical Genetics, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21231.
G. Petersen, The Oncology Center and Department of Epidemiology, The Johns Hopkins University Schools of Medicine, Hygiene, and Public Health.
S. R. Hamilton, Department of Pathology and the Oncology Center, The Johns Hopkins University School of Medicine, Baltimore, MD 21231.
M. C. Nilbert, D. B. Levy, T. M. Bryan, A. C. Preisinger, K. J. Smith, L.-K. Su, K. W. Kinzler, B. Vogelstein, Molecular Genetics Laboratory, The Johns Hopkins University School of Medicine, Baltimore, MD 21231.

*To whom correspondence should be addressed.