Consommation et Corporations Canada
Consumer and Corporate Affairs Canada

Bureau des brevets
Patent Office

Ottawa, Canada
K1A 0C9

(21) (A1) 2,007,699
(22) 1990/01/12
(43) 1991/07/12
(52) 195-1.12
C.R. CL. 530-13.00
150-8.5

(51) INTL.CL.[5] C12N-15/00; C12N-5/00; C12P-21/00; C07K-13/00;

(19)(CA) **APPLICATION FOR CANADIAN PATENT** (12)

(54) Introns and Exons of the Cystic Fibrosis Gene and a Mutation at ▲I507 of the Gene

(72) Tsui, Lap-Chee - Canada ;
Rommers, Johanna M. - U.S.A. ;
Kerem, Bat-Sheva - U.S.A. ;

(73) HSC Research Development Corporation - Canada ;

(57) 41 Claims

Notice: The specification contained herein as filed

ABSTRACT OF THE DISCLOSURE

The cystic fibrosis gene and its gene product are described for the 507 mutant form. The genetic and protein information is used in developing DNA diagnosis, protein diagnosis, carrier and patient screening, cloning of the gene and manufacture of the protein, and development of cystic fibrosis affected animals.

INTRONS AND EXONS THE CYSTIC FIBROSIS GENE AND A MUTATION AT ΔI507 OF THE GENE

FIELD OF THE INVENTION

The present invention relates generally to the cystic fibrosis (CF) gene, and, more particularly to the identification, isolation and cloning of the DNA sequence corresponding to the normal and a mutant of the CF gene, as well as their transcripts, gene products and genetic information at exon/intron boundaries. The present invention also relates to methods of screening for and detection of CF carriers, CF diagnosis, prenatal CF screening and diagnosis, and gene therapy utilizing recombinant technologies and drug therapy using the information derived from the DNA, protein, and the metabolic function of the protein.

BACKGROUND OF THE INVENTION

Cystic fibrosis (CF) is the most common severe autosomal recessive genetic disorder in the Caucasian population. It affects approximately 1 in 2000 live births in North America [Boat et al, The Metabolic Basis of Inherited Disease, 6th ed, pp 2649-2680, McGraw Hill, NY (1989)]. Approximately 1 in 20 persons are carriers of the disease.

Although the disease was first described in the late 1930's, the basic defect remains unknown. The major symptoms of cystic fibrosis include chronic pulmonary disease, pancreatic exocrine insufficiency, and elevated sweat electrolyte levels. The symptoms are consistent with cystic fibrosis being an exocrine disorder. Although recent advances have been made in the analysis of ion transport across the apical membrane of the epithelium of CF patient cells, it is not clear that the abnormal regulation of chloride channels represents the primary defect in the disease. Given the lack of understanding of the molecular mechanism of the disease, an alternative approach has therefore been

taken in an attempt to understand the nature of the molecular defect through direct cloning of the responsible gene on the basis of its chromosomal location.

However, there is no clear phenotype that directs an approach to the exact nature of the genetic basis of the disease, or that allows for an identification of the cystic fibrosis gene. The nature of the CF defect in relation to the population genetics data has not been readily apparent. Both the prevalence of the disease and the clinical heterogeneity have been explained by several different mechanisms: high mutation rate, heterozygote advantage, genetic drift, multiple loci, and reproductive compensation.

Many of the hypotheses can not be tested due to the lack of knowledge of the basic defect. Therefore, alternative approaches to the determination and characterization of the CF gene have focussed on an attempt to identify the location of the gene by genetic analysis.

Linkage analysis of the CF gene to antigenic and protein markers was attempted in the 1950's, but no positive results were obtained [Steinberg et al Am. J. Hum. Genet. 8: 162-176, (1956); Steinberg and Morton Am. J. Hum. Genet 8: 177-189, (1956); Goodchild et al J. Med. Genet. 7: 417-419, 1976.

More recently, it has become possible to use RFLP's to facilitate linkage analysis. The first linkage of an RFLP marker to the CF gene was disclosed in 1985 [Tsui et al. Science 230: 1054-1057, 1985] in which linkage was found between the CF gene and an uncharacterized marker D0CRI-917. The association was found in an analysis of 39 families with affected CF children. This showed that although the chromosomal location had not been established, the location of the

disease gene had been narrowed to about 1% of the human genome, or about 30 million nucleotide base pairs.

The chromosomal location of the D0CRI-917 probe was established using rodent-human hybrid cell lines containing different human chromosome complements. It was shown that DOCR1-917 (and therefore the CF gene) maps to human chromosome 7.

Further physical and genetic linkage studies were pursued in an attempt to pinpoint the location of the CF gene. Zengerling et al [Am. J. Hum. Genet. 40: 228-236 (1987)] describe the use of human-mouse somatic cell hybrids to obtain a more detailed physical relationship between the CF gene and the markers known to be linked with it. This publication shows that the CF gene can be assigned to either the distal region of band q22 or the proximal region of band q31 on chromosome 7.

Rommens et al [Am. J. Hum. Genet. 43: 645-663, (1988)] give a detailed discussion of the isolation of many new 7q31 probes. The approach outlined led to the isolation of two new probes, D7S122 and D7S340, which are close to each other. Pulsed field gel electrophoresis mapping indicates that these two RFLP markers are between two markers known to flank the CF gene, MET [White, R., Woodward S., Leppert M., et al. Nature 318: 382-384, (1985)] and D7S8 [Wainwright, B. J., Scambler, P. J., and J. Schmidtke, Nature 318: 384-385 (1985)], therefore in the CF gene region. The discovery of these markers provides a starting point for chromosome walking and jumping.

Estivill et al, [Nature 326: 840-845(1987)] disclose that a candidate cDNA gene was located and partially characterized. This however, does not teach the correct location of the CF gene. The reference discloses a candidate cDNA gene downstream of a CpG island, which are undermethylated GC nucleotide-rich regions upstream of many vertebrate genes. The

chromosomal localization of the candidate locus is identified as the XV2C region. This region is described in European Patent Application 88303645.1. However, that actual region does not include the CF gene.

A major difficulty in identifying the CF gene has been the lack of cytologically detectable chromosome rearrangements or deletions, which greatly facilitated all previous successes in the cloning of human disease genes by knowledge of map position.

Such rearrangements and deletions could be observed cytologically and as a result, a physical location on a particular chromosome could be correlated with the particular disease. Further, this cytological location could be correlated with a molecular location based on known relationship between publicly available DNA probes and cytologically visible alterations in the chromosomes. Knowledge of the molecular location of the gene for a particular disease would allow cloning and sequencing of that gene by routine procedures, particularly when the gene product is known and cloning success can be confirmed by immunoassay of expression products of the cloned genes.

In contrast, neither the cytological location nor the gene product of the gene for cystic fibrosis was known in the prior art. With the recent identification of MET and D7S8, markers which flanked the CF gene but did not pinpoint its molecular location, the present inventors devised various novel gene cloning strategies to approach the CF gene in accordance with the present invention. The methods employed in these strategies include chromosome jumping from the flanking markers, cloning of DNA fragments from a defined physical region with the use of pulsed field gel electrophoresis, a combination of somatic cell hybrid and molecular cloning techniques designed to isolate DNA fragments from undermethylated CpG islands near CF, chromosome

microdissection and cloning, and saturation cloning of a large number of DNA markers from the 7q31 region. By means of these novel strategies, the present inventors were able to identify the gene responsible for cystic fibrosis where the prior art was uncertain or, even in one case, wrong.

The application of these genetic and molecular cloning strategies has allowed the isolation and cDNA cloning of the cystic fibrosis gene on the basis of its chromosomal location, without the benefit of genomic rearrangements to point the way. The identification of the normal and mutant forms of the CF gene and gene products has allowed for the development of screening and diagnostic tests for CF utilizing nucleic acid probes and antibodies to the gene product. Through interaction with the defective gene product and the pathway in which this gene product is involved, therapy through normal gene product supplementation and gene manipulation and delivery are now made possible.

The gene involved in the cystic fibrosis disease process, hereinafter the "CF gene" and its functional equivalents, has been identified, isolated and cDNA cloned, and its transcripts and gene products identified and sequenced. A three base pair deletion leading to the omission of a phenylalanine residue in the gene product has been determined to correspond to the mutations of the CF gene in approximately 70% of the patients affected with CF, with different mutations involved in most if not all the remaining cases. This subject matter is disclosed in co-pending United States patent application S.N. 396,894 filed August 22, 1989 and its related continuation-in-part applications S.N. 399,945 filed August 24, 1989 and S.N. 401,609 filed August 31, 1989.

SUMMARY OF THE INVENTION

According to this invention, another three base pair deletion leading to the omission of a isoleucine residue in the gene product has been determined. This three base pair deletion corresponds to a mutation of the CF gene in a minority of patients affected with CF. Although not accurately determined, it is believed that this three base pair deletion corresponds to a mutation of the CF gene in a small minority of the patients affected with CF. Furthermore, in accordance with this invention, considerable genetic information is provided at the exon/intron boundaries of the chromosomal CF gene.

With the identification and sequencing of the mutant gene and its gene product, nucleic acid probes and antibodies raised to the mutant gene product can be used in a variety of hybridization and immunological assays to screen for and detect the presence of either the defective CF gene or gene product. Assay kits for such screening and diagnosis can also be provided. The genetic information derived from the intron/exon boundaries is also very useful in various screening and diagnosis procedures.

Patient therapy through supplementation with the normal gene product, whose production can be amplified using genetic and recombinant techniques, or its functional equivalent, is now also possible. Correction or modification of the defective gene product through drug treatment means is now possible. In addition, cystic fibrosis can be cured or controlled through gene therapy by correcting the gene defect in situ or using recombinant or other vehicles to deliver a DNA sequence capable of expression of the normal gene product to the cells of the patient.

According to another aspect of the invention, a purified mutant CF gene comprises a DNA sequence

encoding an amino acid sequence for a protein where the protein, when expressed in cells of the human body, is associated with altered cell function which correlates with the genetic disease cystic fibrosis.

According to another aspect of the invention, a purified RNA molecule comprises an RNA sequence corresponding to the above DNA sequence.

According to another aspect of the invention, a DNA molecule comprises a cDNA molecule corresponding to the above DNA sequence.

According to another aspect of the invention, a DNA molecule comprises a DNA sequence encoding mutant CFTR polypeptide having the sequence according to the following Figure 1 for amino acid residue positions 1 to 1480. The sequence is further characterized by a three base pair mutation which results in the deletion of isoleucine from amino acid residue position 507.

According to another aspect of the invention, a DNA molecule comprises a cDNA molecule corresponding to the above DNA sequence.

According to another aspect of the invention, the cDNA molecule comprises a DNA sequence selected from the group consisting of:

(a) DNA sequences which correspond to the 507 mutant DNA sequence and which encode, on expression, for mutant CFTR polypeptide;

(b) DNA sequences which correspond to a fragment of the 507 mutant DNA sequence, including at least twenty nucleotides;

(c) DNA sequences which comprise at least twenty nucleotides and encode a fragment of the 507 mutant CFTR protein amino acid sequence;

(d) DNA sequences encoding an epitope encoded by at least eighteen sequential nucleotides in the 507 mutant DNA sequence.

According to another aspect of the invention, a DNA sequence selected from the group consisting of:

(a) DNA sequences which correspond to portions of DNA sequences of boundaries of exons/introns of the genomic CF gene;

(b) DNA sequences of at least eighteen sequential nucleotides at boundaries of exons/introns of the genomic CF gene depicted in Figure 1; and

(c) DNA sequences of at least eighteen sequential nucleotides of intron portions of the genomic CF gene of Figure 1.

According to another aspect of the invention, a purified nucleic acid probe comprises a DNA or RNA nucleotide sequence corresponding to the above noted selected DNA sequences of groups (a) to (d).

According to another aspect of the invention, purified RNA molecule comprising RNA sequence corresponds to the 507 mutant DNA sequence.

A purified nucleic acid probe comprising a DNA or RNA nucleotide sequence corresponding to the 507 mutant sequences as recited above.

According to another aspect of the invention, a recombinant cloning vector comprising the DNA sequences of the 507 mutant DNA and fragments thereof is provided. The vector, according to an aspect of this invention, is operatively linked to an expression control sequence in the recombinant DNA molecule so that the selected 507 mutant DNA sequences for the mutant CFTR polypeptide can be expressed. The expression control sequence is selected from the group consisting of sequences that control the expression of genes of prokaryotic or eukaryotic cells and their viruses and combinations thereof.

According to another aspect of the invention, a method for producing a 507 mutant CFTR polypeptide comprises the steps of:

(a) culturing a host cell transfected with the recombinant vector for the mutant DNA sequence in a medium and under conditions favorable for expression of the 507 mutant CFTR polypeptide; and

(b) isolating the expressed mutant CFTR polypeptide.

According to another aspect of the invention, a purified protein of human cell membrane origin comprises an amino sequence encoded by the 507 mutant DNA sequence where the protein, when present in human cell membrane, is associated with cell function which causes the genetic disease cystic fibrosis.

According to another aspect of the invention, a method is provided for screening a subject to determine if the subject is a CF carrier or a CF patient comprising the steps of providing a biological sample of the subject to be screened and providing an assay for detecting in the biological sample, the presence of at least a member from the group consisting of:

(a) 507 mutant CF gene;

(b) mutant CF gene products and mixtures thereof;

(c) DNA sequences which correspond to portions of DNA sequences of boundaries of exons/introns of the genomic CF gene;

(d) DNA sequences of at least eighteen sequential nucleotides at boundaries of exons/introns of the genomic CF gene depicted in Figure 1; and

(e) DNA sequences of at least eighteen sequential nucleotides of intron portions of the genomic CF gene of Figure 1.

According to another aspect of the invention, a kit for assaying for the presence of a CF gene by immunoassay techniques comprises:

(a) an antibody which specifically binds to a gene product of the 507 mutant DNA sequence;

(b) reagent means for detecting the binding of the antibody to the gene product; and

(c) the antibody and reagent means each being present in amounts effective to perform the immunoassay.

According to another aspect of the invention, a kit for assaying for the presence of a 507 mutant CF gene by hybridization technique comprises:

(a) an oligonucleotide probe which specifically binds to the 507 mutant CF gene;

(b) reagent means for detecting the hybridization of the oligonucleotide probe to the 507 mutant CF gene; and

(c) the probe and reagent means each being present in amounts effective to perform the hybridization assay.

According to another aspect of the invention, an animal comprises an heterologous cell system. The cell system includes a recombinant cloning vector which includes the recombinant DNA sequence corresponding to the 507 mutant DNA sequence which induces cystic fibrosis symptoms in the animal.

According to another aspect of the invention, in a polymerase chain reaction to amplify a selected exon of a cDNA sequence of Figure 1, the use of oligonucleotide primers from intron portions near the 5′ and 3′ boundaries of the selected exon of Figure 18.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is the nucleotide sequence of the CF gene and the amino acid sequence of the CFTR protein amino acid sequence with Δ indicating mutations at the 507 and 508 protein positions.

Figure 2 is a restriction map of the CF gene and the schematic strategy used to chromosome walk and jump to the gene.

Figure 3 is a pulsed-field-gel electrophoresis map of the region including and surrounding the CF gene.

Figures 4A, 4B and 4C show the detection of conserved nucleotide sequences by cross-species hybridization.

Figure 4D is a restriction map of overlapping segments of probes E4.3 and H1.6.

Figure 5 is an RNA blot hybridization analysis, using genomic and cDNA probes. Hybridization to fibroblast, trachea (normal and CF), pancreas, liver, HL60, T84, and brain RNA is shown.

Figure 6 is the methylation status of the E4.3 cloned region at the 5′ end of the CF gene.

Figure 7 is a restriction map of the CFTR cDNA showing alignment of the cDNA to the genomic DNA fragments.

Figure 8 is an RNA gel blot analysis depicting hybridization by a portion of the CFTR cDNA (clone 10-1) to a 6.5 kb mRNA transcript in various human tissues.

Figure 9 is a DNA blot hybridization analysis depicting hybridization by the CFTR cDNA clones to genomic DNA digested with EcoRI and Hind III.

Figure 10 is a primer extension experiment characterizing the 5′ and 3′ ends of the CFTR cDNA.

Figure 11 is a hydropathy profile and shows predicted secondary structures of CFTR.

Figure 12 is a dot matrix analysis of internal homologies in the predicted CFTR polypeptide.

Figure 13 is a schematic model of the predicted CFTR protein.

Figure 14 is a schematic diagram of the restriction fragment length polymorphisms (RFLP's) closely linked to the CF gene where the inverted triangle indicates the location of the F508 3 base pair deletion.

Figure 15 represents alignment of the most conserved segments of the extended NBFs of CFTR with comparable regions of other proteins.

Figure 16 is the DNA sequence around the F508 deletion.

Figure 17 is a representation of the nucleotide sequencing gel showing the DNA sequence at the F508 deletion.

Figure 18 is the nucleotide sequence of the portions of introns and complete exons of the genomic CF gene for exons 4 and 6 to 24 of cDNA sequence of Figure 1;

Figure 19 shows the results of amplification of genomic DNA using intron oligonucleotides bounding exon 10;

Figure 20 shows the separation by gel electrophoresis of the amplified genomic DNA products of a CF family.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

1. DEFINITIONS

In order to facilitate review of the various embodiments of the invention and an understanding of various elements and constituents used in making the invention and using same, the following definition of terms used in the invention description is as follows:

CF - cystic fibrosis

CF carrier - a person in apparent health whose chromosomes contain a mutant CF gene that may be transmitted to that person's offspring.

CF patient - a person who carries a mutant CF gene on each chromosome, such that they exhibit the clinical symptoms of cystic fibrosis.

CF gene - the gene whose mutant forms are associated with the disease cystic fibrosis. This definition is understood to include the various sequence polymorphisms that exist, wherein nucleotide substitutions in the gene sequence do not affect the essential function of the gene product. This term primarily relates to an isolated coding sequence, but

can also include some or all of the flanking regulatory elements and/or introns.

Genomic CF gene - the CF gene which includes flanking regulatory elements and/or introns at boundaries of exons of the CF gene.

CF - PI - cystic fibrosis pancreatic insufficient, the major clinical subgroup of cystic fibrosis patients, characterized by insufficient pancreatic exocrine function.

CF - PS - cystic fibrosis pancreatic sufficient, a clinical subgroup of cystic fibrosis patients with sufficient pancreatic exocrine function for normal digestion of food.

CFTR - cystic fibrosis transmembrane conductance regulator protein, encoded by the CF gene. This definition includes the protein as isolated from human or animal sources, as produced by recombinant organisms, and as chemically or enzymatically synthesized. This definition is understood to include the various polymorphic forms of the protein wherein amino acid substitutions in the variable regions of the sequence does not affect the essential functioning of the protein, or its hydropathic profile or secondary or tertiary structure.

DNA - standard nomenclature is used to identify the bases.

Intronless DNA - a piece of DNA lacking internal non-coding segments, for example, cDNA.

IRP locus sequence - (protooncogene int-1 related), a gene located near the CF gene.

Mutant CFTR - a protein that is highly analagous to CFTR in terms of primary, secondary, and tertiary structure, but wherein a small number of amino acid substitutions and/or deletions and/or insertions result in impairment of its essential function, so that organisms whose epithelial cells express mutant CFTR

rather than CFTR demonstrate the symptoms of cystic fibrosis.

mCF - a mouse gene orthologous to the human CF gene

NBFs - nucleotide (ATP) binding folds

ORF - open reading frame

PCR - polymerase chain reaction

Protein - standard single letter nomenclature is used to identify the amino acids

R-domain - a highly charged cytoplasmic domain of the CFTR protein

RSV - Rous Sarcoma Virus

SAP - surfactant protein

RFLP - restriction fragment length polymorphism

507 mutant CF gene - the CF gene which includes a DNA base pair mutation at the 506 or 507 protein position of the cDNA of the CF gene

507 mutant DNA sequence - equivalent meaning to the 507 mutant CF gene

507 mutant CFTR protein or mutant CFTR protein amino acid sequence, or mutant CFTR polypeptide - the mutant CFTR protein wherein an amino acid deletion occurs at the isoleucine 506 or 507 protein position of the CFTR.

2. ISOLATING THE CF GENE

Using chromosome walking, jumping, and cDNA hybridization, DNA sequences encompassing > 500 kilobase pairs (kb) have been isolated from a region on the long arm of human chromosome 7 containing the cystic fibrosis (CF) gene. This technique is disclosed in detail in the aforemention co-pending United States patent applications. For purposes of convenience in understanding and isolating the CF gene and identifying the 507 mutation, the technique is reiterated here. Several transcribed sequences and conserved segments have been identified in this region. One of these corresponds to the CF gene and spans approximately 250

kb of genomic DNA. Overlapping complementary DNA (cDNA) clones have been isolated from epithelial cell libraries with a genomic DNA segment containing a portion of the cystic fibrosis gene. The nucleotide sequence of the isolated cDNA is shown in Figure 1. In each row of the respective sequences the lower row is a list by standard nomenclature of the nucleotide sequence. The upper row in each respective row of sequences is standard single letter nomenclature for the amino acid corresponding to the respective codon.

Accordingly, the isolation of the CF gene provided a cDNA molecule comprising a DNA sequence selected from the group consisting of:

(a) DNA sequences which correspond to the DNA sequence of Figure 1 from amino acid residue position 1 to position 1480;

(b) DNA sequences encoding normal CFTR polypeptide having the sequence according to Figure 1 for amino acid residue positions from 1 to 1480;

(c) DNA sequences which correspond to a fragment of the sequence of Figure 1 including at least 16 sequential nucleotides between amino acid residue positions 1 and 1480;

(d) DNA sequences which comprise at least 16 nucleotides and encode a fragement of the amino acid sequence of Figure 1; and

(e) DNA sequences encoding an epitope encoded by at least 18 sequential nucleotides in the sequence of Figure 1 between amino acid residue positions 1 and 1480.

According to this invention, the isolation of another mutation in the CF gene also provides a cDNA molecule comprising a DNA sequence selected from the group consisting of:

a) DNA sequences which correspond to the DNA sequence encoding mutant CFTR polypeptide characterized

by cystic fibrosis-associated activity in human epithelial cells, or the DNA sequence of Figure 1 for the amino acid residue positions 1 to 1480 yet further characterized by a three base pair mutation which results in the deletion of isoleucine from amino acid residue position 507;

b) DNA sequences which correspond to fragments of the mutant portion of the sequence of paragraph a) and which include at least sixteen nucleotides;

c) DNA sequences which comprise at least sixteen nucleotides and encode a fragment of the amino acid sequence encoded for by the mutant portion of the DNA sequence of paragraph a); and

d) DNA sequences encoding an epitope encoded by at least 18 sequential nucleotides in the mutant portion of the sequence of the DNA of paragraph a).

Transcripts of approximately 6,500 nucleotides in size are detectable in tissues affected in patients with CF. Based upon the isolated nucleotide sequence, the predicted protein consists of two similar regions, each containing a first domain having properties consistent with membrane association and a second domain believed to be involved in ATP binding.

A 3 bp deletion which results in the omission of a phenylalanine residue at the center of the first predicted nucleotide binding domain (amino acid position 508 of the CF gene product) was detected in CF patients. This mutation in the normal DNA sequence of Figure 1 corresponds to approximately 70% of the mutations in cystic fibrosis patients. Extended haplotype data based on DNA markers closely linked to the putative disease gene suggest that the remainder of the CF mutant gene pool consists of multiple, different mutations. This is now exemplified by this invention at the 506 or 507 protein position. A small set of these latter mutant alleles (approximately 8%) may confer

residual pancreatic exocrine function in a subgroup of patients who are pancreatic sufficient.

2.1 CHROMOSOME WALKING AND JUMPING

Large amounts of the DNA surrounding the D7S122 and D75340 linkage regions of Rommens et al supra were searched for candidate gene sequences. In addition to conventional chromosome walking methods, chromosome jumping techniques were employed to accelerate the search process. From each jump endpoint a new bidirectional walk could be initiated. Sequential walks halted by "unclonable" regions often encountered in the mammalian genome could be circumvented by chromosome jumping.

The chromosome jumping library used has been described previously [Collins et al, Science 235, 1046 (1987); Ianuzzi et al, Am. J. Hum. Genet. 44, 695 (1989)]. The original library was prepared from a preparative pulsed field gel, and was intended to contain partial EcoR1 fragments of 70 - 130 kb; subsequent experience with this library indicates that smaller fragments were also represented, and jumpsizes of 25 - 110 kb have been found. The library was plated on sup^- host MC1061 and screened by standard techniques, [Maniatis et al]. Positive clones were subcloned into pBRΔ23Ava and the beginning and end of the jump identified by EcoR1 and Ava 1 digestion, as described in Collins, Genome analysis: A practical approach (IRL, London, 1988), pp. 73-94) . For each clone, a fragment from the end of the jump was checked to confirm its location on chromosome 7. The contiguous chromosome region covered by chromosome walking and jumping was about 250 kb. Direction of the jumps was biased by careful choice of probes, as described by Collins et al and Ianuzzi et al, supra. The entire region cloned, including the sequences isolated with the use of the CF gene cDNA, is approximately 500 kb.

The schematic representation of the chromosome walking and jumping strategy is illustrated in Figure 2. CF gene exons are indicated by Roman numerals in this Figure. Horizontal lines above the map indicate walk steps whereas the arcs above the map indicate jump steps. The Figure proceeds from left to right in each of six tiers with the direction of ends toward 7cen and 7qter as indicated. The restriction map for the enzymes EcoRI, HindIII, and BamHI is shown above the solid line, spanning the entire cloned region. Restriction sites indicated with arrows rather than vertical lines indicate sites which have not been unequivocally positioned. Additional restriction sites for other enzymes are shown below the line. Gaps in the cloned region are indicated by ||. These occur only in the portion detected by cDNA clones of the CF transcript. These gaps are unlikely to be large based on pulsed field mapping of the region. The walking clones, as indicated by horizontal arrows above the map, have the direction of the arrow indicating the walking progress obtained with each clone. Cosmid clones begin with the letter c; all other clones are phage. Cosmid CF26 proved to be a chimera; the dashed portion is derived from a different genomic fragment on another chromosome. Roman numerals I through XXIV indicate the location of exons of the CF gene. The horizontal boxes shown above the line are probes used during the experiments. Three of the probes represent independent subcloning of fragments previously identified to detect polymorphisms in this region: H2.3A corresponds to probe XV2C (X. Estivill et al, Nature, 326: 840 (1987)), probe E1 corresponds to KM19 (Estivill, supra), and probe E4.1 corresponds to Mp6d.9 (X. Estivill et al. Am. J. Hum. Genet. 44, 704 (1989)). G-2 is a subfragment of E6 which detects a transcribed sequence. R161, R159, and R160 are synthetic oligonucleotides constructed from

parts of the IRP locus sequence [B. J. Wainwright et al, EMBO J., 7: 1743 (1988)], indicating the location of this transcript on the genomic map.

As the two independently isolated DNA markers, D7S122 (pH131) and D7S340 (TM58), were only approximately 10 kb apart (Figure 2), the walks and jumps were essentially initiated from a single point. The direction of walking and jumping with respect to MET and D7S8 was then established with the crossing of several rare-cutting restriction endonuclease recognition sites (such as those for Xho I, Nru I and Not I, see Figure 2) and with reference to the long range physical map of J. M. Rommens et al. Am. J. Hum. Genet., in press; A. M. Poustka, et al, Genomics 2, 337 (1988); M. L. Drumm et al. Genomics 2, 346 (1988). The pulsed field mapping data also revealed that the Not I site identified by the inventors of the present invention (see Figure 2, position 113 kb) corresponded to the one previously found associated with the IRP locus (Estivill et al 1987, supra). Since subsequent genetic studies showed that CF was most likely located between IRP and D7S8 [M. Farrall et al, Am. J. Hum. Genet. 43, 471 (1988), B.S. Kerem et al. Am. J. Hum. Genet. 44, 827 (1989)], the walking and jumping effort was continued exclusively towards cloning of this interval. It is appreciated, however, that other coding regions, as identified in Figure 2, for example, G-2, CF14 and CF16, were located and extensively investigated. Such extensive investigations of these other regions revealed that they were not the CF gene based on genetic data and sequence analysis. Given the lack of knowledge of the location of the CF gene and its characteristics, the extensive and time consuming examination of the nearby presumptive coding regions did not advance the direction of search for the CF gene. However, these investigations were necessary in order to

rule out the possibility of the CF gene being in those regions.

Three regions in the 280 kb segment were found not to be readily recoverable in the amplified genomic libraries initially used. These less clonable regions were located near the DNA segments H2.3A and X.6, and just beyond cosmid cW44, at positions 75-100 kb, 205-225 kb, and 275-285 kb in Figure 2, respectively. The recombinant clones near H2.3A were found to be very unstable with dramatic rearrangements after only a few passages of bacterial culture. To fill in the resulting gaps, primary walking libraries were constructed using special host-vector systems which have been reported to allow propagation of unstable sequences [A. R. Wyman, L. B. Wolfe, D. Botstein, Proc. Nat. Acad. Sci. U. S. A. 82, 2880 (1985); K. F. Wertman, A. R. Wyman, D. Botstein, Gene 49, 253 (1986); A. R. Wyman, K. F. Wertman, D. Barker, C. Helms, W. H. Petri, Gene, 49, 263 (1986)]. Although the region near cosmid cW44 remains to be recovered, the region near X.6 was successfully rescued with these libraries.

2.2 CONSTRUCTION OF GENOMIC LIBRARIES

Genomic libraries were constructed after procedures described in Manatis, et al, Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1982) and are listed in Table 1. This includes eight phage libraries, one of which was provided by T. Maniatis [Fritsch et al, Cell, 19:959 (1980)]; the rest were constructed as part of this work according to procedures described in Maniatis et al, supra. Four phage libraries were cloned in λDASH (commercially available from Stratagene) and three in λFIX (commercially available from Stratagene), with vector arms provided by the manufacturer. One λDASH library was constructed from Sau 3A-partially digested DNA from a human-hamster

hybrid containing human chromosome 7 (4AF/102/K015) [Rommens et al Am. J. Hum. Genet 43, 4 (1988)], and other libraries from partial Sau3A, total BamHI, or total EcoRI digestion of human peripheral blood or lymphoblastoid DNA. To avoid loss of unstable sequences, five of the phage libraries were propagated on the recombination-deficient hosts DB1316 ($recD^-$), CES 200 ($recBC^-$) [Wyman et al, supra , Wertman et al supra, Wyman et al supra]; or TAP90 [Patterson et al Nucleic Acids Res. 15:6298 (1987)]. Three cosmid libraries were then constructed. In one the vector pCV108 [Lau et al Proc. Natl. Acad. Sci USA 80:5225 (1983)] was used to clone partially digested (Sau 3A) DNA from 4AF/102/K015 [Rommens et al Am.J. Hum. Genet. 43:4 (1988)]. A second cosmid library was prepared by cloning partially digested (Mbo I) human lymphoblastoid DNA into the vector pWE-IL2R, prepared by inserting the RSV (Rous Sarcoma Virus) promoter-driven cDNA for the interleukin-2 receptor α-chain (supplied by M. Fordis and B. Howard) in place of the neo-resistance gene of pWE15 [Wahl et al Proc. Natl. Acad. Sci. USA 84:2160 (1987)]. An additional partial Mbo I cosmid library was prepared in the vector pWE-IL2-Sal, created by inserting a Sal I linker into the Bam HI cloning site of pWE-EL2R (M. Drumm, unpublished data); this allows the use of the partial fill-in technique to ligate Sal I and Mbo I ends, preventing tandem insertions [Zabarovsky et al Gene 42:19 (1986)]. Cosmid libraries were propagated in E. coli host strains DH1 or 490A [M. Steinmetz, A. Winoto, K. Minard, L. Hood, Cell 28, 489(1982)].

TABLE 1

GENOMIC LIBRARIES

Vector	Source of human DNA	Host	Complexity	Ref
λ Charon 4A	HaeII/AluI-partially digested total human liver DNA	LE392	1×10^6 (amplified)	Lawn et al 1980
pCV108	Sau3a-partially digested DNA from 4AF/KO15	DK1	3×10^6 (amplified)	
λdash	Sau3A-partially digested DNA from 4AF/KO15	LE392	1×10^6 (amplified)	
λdash	Sau3A-partially digested total human peripheral blood DNA	DB1316	1.5×10^6	
λdash	BamHI-digested total human peripheral blood DNA	DB1316	1.5×10^6	
λdash	EcoRI-partially digested total human peripheral blood DNA	DB1316	8×10^6	
λFIX	MboI-partially digested human lymphoblastoid DNA	LE392	1.5×10^6	
λFIX	MboI-partially digested human lymphoblastoid DNA	CE200	1.2×10^6	
λFIX	MboI-partially digested human lymphoblastoid DNA	TAP90	1.3×10^6	
pWE-IL2R	MboI-partially digested human lymphoblastoid DNA	490A	5×10^5	
pWE-IL2R-Sal	MboI-partially digested human lymphoblastoid DNA	490A	1.2×10^6	
λCh3A Δlac (jumping)	EcoRI-partially digested (24-110 kb) human lymphoblastoid DNA	MC1061	3×10^6	Collins et al supra and Iannuzzi et al supra

Three of the phage libraries were propagated and amplified in E. coli bacterial strain LE392. Four subsequent libraries were plated on the recombination-deficient hosts DB1316 (recD$^-$) or CES200 (rec BC$^-$) [Wyman 1985, supra; Wertman 1986, supra; and Wyman 1986, supra] or in one case TAP90 [T.A. Patterson and M. Dean, Nucleic Acids Research 15, 6298 (1987)].

Single copy DNA segments (free of repetitive elements) near the ends of each phage or cosmid insert were purified and used as probes for library screening to isolate overlapping DNA fragments by standard procedures. (Maniatis, et al, supra).

1-2 x 10^6 phage clones were plated on 25-30 150 mm petri dishes with the appropriate indicator bacterial host and incubated at 37°C for 10-16 hr. Duplicate "lifts" were prepared for each plate with nitrocellulose or nylon membranes, prehybridized and hybridized under conditions described [Rommens et al, 1988, supra]. Probes were labelled with ^{32}P to a specific activity of >5 x 10^8 cpm/μg using the random priming procedure [A.P. Feinberg and B. Vogelstein, Anal. Biochem. 132, 6 (1983)]. The cosmid library was spread on ampicillin-containing plates and screened in a similar manner.

DNA probes which gave high background signals could often be used more successfully by preannealing the boiled probe with 250 μg/ml sheared denatured placental DNA for 60 minutes prior to adding the probe to the hybridization bag.

For each walk step, the identity of the cloned DNA fragment was determined by hybridization with a somatic cell hybrid panel to confirm its chromosomal location, and by restriction mapping and Southern blot analysis to confirm its colinearity with the genome.

The total combined cloned region of the genomic DNA sequences isolated and the overlapping cDNA clones,

extended >500 kb. To ensure that the DNA segments isolated by the chromosome walking and jumping procedures were colinear with the genomic sequence, each segment was examined by:

(a) hybridization analysis with human-rodent somatic hybrid cell lines to confirm chromosome 7 localization,

(b) pulsed field gel electrophoresis, and

(c) comparison of the restriction map of the cloned DNA to that of the genomic DNA.

Accordingly, single copy human DNA sequences were isolated from each recombinant phage and cosmid clone and used as probes in each of these hybridization analyses as performed by the procedure of Maniatis, et al supra.

While the majority of phage and cosmid isolates represented correct walk and jump clones, a few resulted from cloning artifacts or cross-hybridizing sequences from other regions in the human genome, or from the hamster genome in cases where the libraries were derived from a human-hamster hybrid cell line. Confirmation of correct localization was particularly important for clones isolated by chromosome jumping. Many jump clones were considered and resulted in non-conclusive information leading the direction of investigation away from the gene.

2.3 CONFIRMATION OF THE RESTRICTION MAP

Further confirmation of the overall physical map of the overlapping clones was obtained by long range restriction mapping analysis with the use of pulsed field gel electrophoresis (J. M. Rommens, et al. Am. J. Hum. Genet, in press, A. M. Poustka et al, 1988, supra M.L. Drumm et al, 1988 supra).

Figures 3A to 3E illustrates the findings of the long range restriction mapping study, where a schematic representation of the region is given in Panel E. DNA

from the human-hamster cell line 4AF/102/K015 was digested with the enzymes (A) Sal I, (B) Xho I, (C) Sfi I and (D) Nae I, separated by pulsed field gel electrophoresis, and transferred to Zetaprobe™(BioRad). For each enzyme a single blot was sequentially hybridized with the probes indicated below each of the panels of Figure A to D, with stripping of the blot between hybridizations. The symbols for each enzyme of Figure 3E are: A, Nae I; B, Bss HII; F. Sfi I; L, Sal I; M, Mlu I; N, Not I; R, Nru I; and X, Xho 1. C corresponds to the compression zone region of the gel. DNA preparations, restriction digestion, and crossed field gel electrophoresis methods have been described (Rommens et al, in press, supra). The gels in Figure 3 were run in 0.5X TBE at 7 volts/cm for 20 hours with switching linearly ramped from 10-40 seconds for (A), (B), and (C), and at 8 volts/cm for 20 hours with switching ramped linearly from 50-150 seconds for (D). Schematic interpretations of the hybridization pattern are given below each panel. Fragment lengths are in kilobases and were sized by comparison to oligomerized bacteriophage λDNA and Saccharomyces cerevisiae chromosomes.

H4.0, J44, EG1.4 are genomic probes generated from the walking and jumping experiments (see Figure 2). J30 has been isolated by four consecutive jumps from D7S8 (Collins et al, 1987, supra; Ianuzzi et al, 1989, supra; M. Dean, et al, submitted for publication). 10-1, B.75, and CE1.5/1.0 are cDNA probes which cover different regions of the CF transcript: 10-1 contains exons I - VI, B.75 contains exons V - XII, and CE1.5/1.0 contains exons XII - XXIV. Shown in Figure 3E is a composite map of the entire MET - D7S8 interval. The open boxed region indicates the segment cloned by walking and jumping, and the closed arrow portion indicates the region covered by the CF transcript. The

CpG-rich region associated with the D7S23 locus (Estivill et al, 1987, supra) is at the Not I site shown in parentheses. This and other sites shown in parentheses or square brackets do not cut in 4AF/102/K015, but have been observed in human lymphoblast cell lines.

2.4 IDENTIFICATION OF CF GENE

Based on the findings of long range restriction mapping detailed above it was determined that the entire CF gene is contained on a 380 kb Sal I fragment. Alignment of the restriction sites derived from pulsed field gel analysis to those identified in the partially overlapping genomic DNA clones revealed that the size of the CF gene was approximately 250 kb.

The most informative restriction enzyme that served to align the map of the cloned DNA fragments and the long range restriction map was Xho I; all of the 9 Xho 1 sites identified with the recombinant DNA clones appeared to be susceptible to at least partial cleavage in genomic DNA (compare maps in Figures 1 and 2). Furthermore, hybridization analysis with probes derived from the 3' end of the CF gene identified 2 SfiI sites and confirmed the position of an anticipated Nae I site.

These findings further supported the conclusion that the DNA segments isolated by the chromosome walking and jumping procedures were colinear with the genuine sequence.

2.5 CRITERIA FOR IDENTIFICATION

A positive result based on one or more of the following criteria suggested that a cloned DNA segment may contain candidate gene sequences:

(a) detection of cross-hybridizing sequences in other species (as many genes show evolutionary conservation),

(b) identification of CpG islands, which often mark the 5' end of vertebrate genes [A. P. Bird, Nature, 321,

209 (1986); M. Gardiner-Garden and M. Frommer, J. Mol. Biol. 196, 261 (1987)],

(c) examination of possible mRNA transcripts in tissues affected in CF patients,

(d) isolation of corresponding cDNA sequences,

(e) identification of open reading frames by direct sequencing of cloned DNA segments.

Cross-species hybridization showed strong sequence conservation between human and bovine DNA when CF14, E4.3 and H1.6 were used as probes, the results of which are shown in Figures 4A, 4B and 4C.

Human, bovine, mouse, hamster, and chicken genomic DNAs were digested with Eco RI (R), Hind III (H), and Pst I (P), electrophoresed, and blotted to Zetabind™ (BioRad). The hybridization procedures of Rommens et al, 1988, supra, were used with the most stringent wash at 55°C, 0.2X SSC, and 0.1% SDS. The probes used for hybridization, in Figure 4, included: (A) entire cosmid CF14, (B) E4.3, (C) H1.6. In the schematic of Figure (D), the shaded region indicates the area of cross-species conservation.

The fact that different subsets of bands were detected in bovine DNA with these two overlapping DNA segments (H1.6 and E4.3) suggested that the conserved sequences were located at the boundaries of the overlapped region (Figure 4(D)). When these DNA segments were used to detect RNA transcripts from a variety of tissues, no hybridization signal was detected. In an attempt to understand the cross-hybridizing region and to identify possible open reading frames, the DNA sequences of the entire H1.6 and part of the E4.3 fragment were determined. The results showed that, except for a long stretch of CG-rich sequence containing the recognition sites for two restriction enzymes (Bss HII and Sac II), often found associated with undermethylated CpG islands, there were only short

open reading frames which could not easily explain the strong cross-species hybridization signals.

To examine the methylation status of this highly CpG-rich region revealed by sequencing, genomic DNA samples prepared from fibroblasts and lymphoblasts were digested with the restriction enzymes Hpa II and Msp I and analyzed by gel blot hybridization. The enzyme Hpa II cuts the DNA sequence 5'-CCGG-3' only when the second cytosine is unmethylated, whereas Msp I cuts this sequence regardless of the state of methylation. Small DNA fragments were generated by both enzymes, indicating that this CpG-rich region is indeed undermethylated in genomic DNA. The gel-blot hybridization with the E4.3 segment (Figure 6) reveals very small hybridizing fragments with both enzymes, indicating the presence of a hypomethylated CpG island.

The above results strongly suggest the presence of a coding region at this locus. Two DNA segments (E4.3 and H1.6) which detected cross-species hybridization signals from this area were used as probes to screen cDNA libraries made from several tissues and cell types.

cDNA libraries from cultured epithelial cells were prepared as follows. Sweat gland cells derived from a non-CF individual and from a CF patient were grown to first passage as described [G. Collie et al, In Vitro Cell. Dev. Biol. 21, 592,1985]. The presence of outwardly rectifying channels was confirmed in these cells (J.A. Tabcharani, T.J. Jensen, J.R. Riordan, J.W. Hanrahan, J. Memb. Biol., in press) but the CF cells were insensitive to activation by cyclic AMP (T.J. Jensen, J.W. Hanrahan, J.A. Tabcharani, M. Buchwald and J.R. Riordan, Pediatric Pulmonology, Supplement 2, 100, 1988). RNA was isolated from them by the method of J.M. Chirgwin et al (Biochemistry 18, 5294, 1979). Poly A+RNA was selected (H. Aviv and P. Leder, Proc. Natl. Acad. Sci. USA 69, 1408, 1972) and used as template for

the synthesis of cDNA with oligo (dT) 12-18 as a primer. The second strand was synthesized according to Gubler and Hoffman (Gene 25, 263, 1983). This was methylated with Eco RI methylase and ends were made flush with T4 DNA polymerase. Phosphorylated Eco RI linkers were ligated to the cDNA and restricted with Eco RI. Removal of excess linkers and partial size fractionation was achieved by Biogel A-50 chromatography. The cDNAs were then ligated into the Eco RI site of the commercialy available lamdba ZAP. Recombinants were packaged and propagated in E. coli BB4. Portions of the packaging mixes were amplified and the remainder retained for screening prior to amplification. The same procedures were used to construct a library from RNA isolated from preconfluent cultures of the T-84 colonic carcinoma cell line (Dharmsathaphorn, K. et al. Am. J. Physiol. 246, G204, 1984). The numbers of independent recombinants in the three libraries were: 2×10^6 for the non-CF sweat gland cells, 4.5×10^6 for the CF sweat gland cells and 3.2×10^6 from T-84 cells. These phages were plated at 50,000 per 15 cm plate and plaque lifts made using nylon membranes (Biodyne) and probed with DNA fragments labelled with ^{32}P using DNA polymerase I and a random mixture of oligonucleotides as primer. Hybridization conditions were according to G.M. Wahl and S.L. Berger (Meth. Enzymol. 152,415, 1987). Bluescript™ plasmids were rescued from plaque purified clones by excision with M13 helper phage. The lung and pancreas libraries were purchased from Clontech Lab Inc. with reported sizes of 1.4×10^6 and 1.7×10^6 independent clones.

After screening 7 different libraries each containing 1×10^5 - 5×10^6 independent clones, 1 single clone (identified as 10-1) was isolated with H1.6 from a cDNA library made from the cultured sweat gland epithelial cells of an unaffected (non-CF) individual.

DNA sequencing analysis showed that probe 10-1 contained an insert of 920 bp in size and one potential, long open reading frame (ORF). Since one end of the sequence shared perfect sequence identity with H1.6, it was concluded that the cDNA clone was probably derived from this region. The DNA sequence in common was, however, only 113 bp long (see Figures 1 and 7). As detailed below, this sequence in fact corresponded to the 5'-most exon of the putative CF gene. The short sequence overlap thus explained the weak hybridization signals in library screening and inability to detect transcripts in RNA gel-blot analysis. In addition, the orientation of the transcription unit was tentatively established on the basis of alignment of the genomic DNA sequence with the presumptive ORF of 10-1.

Since the corresponding transcript was estimated to be approximately 6500 nucleotides in length by RNA gel-blot hybridization experiments, further cDNA library screening was required in order to clone the remainder of the coding region. As a result of several successive screenings with cDNA libraries generated from the colonic carcinoma cell line T84, normal and CF sweat gland cells, pancreas and adult lungs, 18 additional clones were isolated (Figure 7, as subsequently discussed in greater detail). DNA sequence analysis revealed that none of these cDNA clones corresponded to the length of the observed transcript, but it was possible to derive a consensus sequence based on overlapping regions. Additional cDNA clones corresponding to the 5' and 3' ends of the transcript were derived from 5' and 3' primer-extension experiments. Together, these clones span a total of about 6.1 kb and contain an ORF capable of encoding a polypeptide of 1480 amino acid residues (Figure 1).

It was unusual to observe that most of the cDNA clones isolated here contained sequence insertions at

various locations of the restriction map of Figure 7. The map details the genomic structure of the CF gene. Exon/intron boundaries are given where all cDNA clones isolated are schematically represented on the upper half of the figure. Many of these extra sequences clearly corresponded to intron regions reversely transcribed during the construction of the cDNA, as revealed upon alignment with genomic DNA sequences.

Since the number of recombinant cDNA clones for the CF gene detected in the library screening was much less than would have been expected from the abundance of transcript estimated from RNA hybridization experiments, it seemed probable that the clones that contained aberrant structures were preferentially retained while the proper clones were lost during propagation. Consistent with this interpretation, poor growth was observed for the majority of the recombinant clones isolated in this study, regardless of the vector used.

The procedures used to obtain the 5′ and 3′ ends of the cDNA were similar to those described (M. Frohman et al, Proc. Nat. Acad. Sci, USA, 85, 8998-9002, 1988). For the 5′ end clones, total pancreas and T84 poly A + RNA samples were reverse transcribed using a primer, (10b), which is specific to exon 2 similarly as has been described for the primer extension reaction except that radioactive tracer was included in the reaction. The fractions collected from an agarose bead column of the first strand synthesis were assayed by polymerase chain reaction (PCR) of eluted fractions. The oligonucleotides used were within the 10-1 sequence (145 nucleotides apart) just 5′ of the extension primer. The earliest fractions yielding PCR product were pooled and concentrated by evaporation and subsequently tailed with terminal deoxynucleotidyl transferase (BRL Labs.) and dATP as recommended by the supplier (BRL Labs). A second strand synthesis was then carried out with Taq

Polymerase (Cetus, AmpliTaq™) using an oligonucleotide containing a tailed linker sequence 5'CGGAATTCTCGAGATC$(T)_{12}$3'.

Amplification by an anchored (PCR) experiment using the linker sequence and a primer just internal to the extension primer which possessed the Eco RI restriction site at its 5' end was then carried out. Following restriction with the enzymes Eco RI and Bgl II and agarose gel purification size selected products were cloned into the plasmid Bluescript KS available from Stratagene by standard procedures (Maniatis et al, supra). Essentially all of the recovered clones contained inserts of less than 350 nucleotides. To obtain the 3' end clones, first strand cDNA was prepared with reverse transcription of 2 μg T84 poly A + RNA using the tailed linker oligonucleotide previously described with conditions similar to those of the primer extension. Amplification by PCR was then carried out with the linker oligonucleotide and three different oligonucleotides corresponding to known sequences of clone T16-4.5. A preparative scale reaction (2 x 100 ul) was carried out with one of these oligonucleotides with the sequence 5'ATGAAGTCCAAGGATTTAG3'.

This oligonucleotide is approximately 70 nucleotides upstream of a Hind III site within the known sequence of T16-4.5. Restriction of the PCR product with Hind III and Xho 1 was followed by agarose gel purification to size select a band at 1.0-1.4 kb. This product was then cloned into the plasmid Bluescript KS available from Stratagene. Approximately 20% of the obtained clones hybridized to the 3' end portion of T16-4.5. 10/10 of plasmids isolated from these clones had identical restriction maps with insert sizes of approx. 1.2 kb. All of the PCR reactions were carried out for 30 cycles in buffer suggested by an enzyme supplier.

An extension primer positioned 157 nt from the 5'end of 10-1 clone was used to identify the start point of the putative CF transcript. The primer was end labeled with $\gamma[^{32}P]ATP$ at 5000 Curies/mmole and T4 polynucleotide kinase and purified by spun column gel filtration. The radiolabeled primer was then annealed with 4-5 ug poly A + RNA prepared from T-84 colonic carcinoma cells in 2X reverse transcriptase buffer for 2 hrs. at 60°C. Following dilution and addition of AMV reverse transcriptase (Life Sciences, Inc.) incubation at 41°C proceeded for 1 hour. The sample was then adjusted to 0.4M NaOH and 20 mM EDTA, and finally neutralized, with NH_4OAc, pH 4.6, phenol extracted, ethanol precipitated, redissolved in buffer with formamide, and analyzed on a polyacrylamide sequencing gel. Details of these methods have been described (Meth. Enzymol. 152, 1987, Ed. S.L. Berger, A.R. Kimmel, Academic Press, N.Y.).

Results of the primer extension experiment using an extension oligonucleotide primer starting 157 nucleotides from the 5' end of 10-1 is shown in Panel A of Figure 10. End labeled ϕX174 bacteriophage digested with Hae III (BRL Labs) is used as size marker. Two major products are observed at 216 and 100 nucleotides. The sequence corresponding to 100 nucleotides in 10-1 corresponds to a very GC rich sequence (11/12) suggesting that this could be a reverse transcriptase pause site. The 5' anchored PCR results are shown in panel B of Figure 10. The 1.4% agarose gel shown on the left was blotted and transferred to Zetaprobe™ membrane (Bio-Rad Lab). DNA gel blot hybridization with radiolabeled 10-1 is shown on the right. The 5' extension products are seen to vary in size from 170-280 nt with the major product at about 200 nucleotides. The PCR control lane shows a fragment of 145 nucleotides. It was obtained by using the test oligomers within the

10-1 sequence. The size markers shown correspond to sizes of 154, 220/210, 298, 344, 394 nucleotides (1kb ladder purchased from BRL Lab).

The schematic shown below Panel B of Figure 10 outlines the procedure to obtain double stranded cDNA used for the amplification and cloning to generate the clones PA3-5 and TB2-7 shown in Figure 7. The anchored PCR experiments to characterize the 3'end are shown in panel C. As depicted in the schematic below Figure 10C, three primers whose relative position to each other were known were used for amplification with reversed transcribed T84 RNA as described. These products were separated on a 1% agarose gel and blotted onto nylon membrane as described above. DNA-blot hybridization with the 3' portion of the T16-4.5 clone yielded bands of sizes that corresponded to the distance between the specific oligomer used and the 3'end of the transcript. These bands in lanes 1, 2a and 3 are shown schematically below Panel C in Figure 10. The band in lane 3 is weak as only 60 nucleotides of this segment overlaps with the probe used. Also indicated in the schematic and as shown in the lane 2b is the product generated by restriction of the anchored PCR product to facilitate cloning to generate the THZ-4 clone shown in Figure 7.

DNA-blot hybridization analysis of genomic DNA digested with EcoRI and HindIII enzymes probed with portions of cDNAs spanning the entire transcript suggest that the gene contains at least 26 exons numbered as Roman numerals I through XXVI (see Figure 9). These correspond to the numbers 1 through 26 shown in Figure 7. The size of each band is given in kb.

In Figure 7, open boxes indicate approximate positions of the 24 exons which have been identified by the isolation of >22 clones from the screening of cDNA libraries and from anchored PCR experiments designed to clone the 5' and 3' ends. The lengths in kb of the Eco

RI genomic fragments detected by each exon is also indicated. The hatched boxes in Figure 7 indicate the presence of intron sequences and the stippled boxes indicate other sequences. Depicted in the lower left by the closed box is the relative position of the clone H1.6 used to detect the first cDNA clone 10-1 from among 10^6 phage of the normal sweat gland library. As shown in Figures 4(D) and 7, the genomic clone H1.6 partially overlaps with an EcoRI fragment of 4.3 kb. All of the cDNA clones shown were hybridized to genomic DNA and/or were fine restriction mapped. Examples of the restriction sites occurring within the cDNAs and in the corresponding genomic fragments are indicated.

With reference to Figure 9, the hybridization analysis includes probes; i.e., cDNA clones 10-1 for panel A, T16-1 (3′ portion) for panel B, T16-4.5 (central portion) for panel C and T16-4.5 (3′ end portion) for panel D. In panel A of Figure 9, the cDNA probe 10-1 detects the genomic bands for exons I through VI. The 3′ portion of T16-1 generated by NruI restriction detects exons IV through XIII as shown in Panel B. This probe partially overlaps with 10-1. Panels C and D, respectively, show genomic bands detected by the central and 3′ end EcoRI fragments of the clone T16-4.5. Two EcoRI sites occur within the cDNA sequence and split exons XIII and XIX. As indicated by the exons in parentheses, two genomic EcoRI bands correspond to each of these exons. Cross hybridization to other genomic fragments was observed. These bands, indicated by N, are not of chromosome 7 origin as they did not appear in human-hamster hybrids containing human chromosome 7. The faint band in panel D indicated by XI in brackets is believed to be caused by the cross-hybridization of sequences due to internal homology with the cDNA.

Since 10-1 detected a strong band on gel blot hybridization of RNA from the T-84 colonic carcinoma cell line, this cDNA was used to screen the library constructed from that source. Fifteen positives were obtained from which clones T6, T6/20, T11, T16-1 and T13-1 were purified and sequenced. Rescreening of the same library with a 0.75 kb Bam HI-Eco RI fragment from the 3' end of T16-1 yielded T16-4.5. A 1.8kb EcoRI fragment from the 3' end of T16-4.5 yielded T8-B3 and T12a, the latter of which contained a polyadenylation signal and tail. Simultaneously a human lung cDNA library was screened; many clones were isolated including those shown here with the prefix 'CDL'. A pancreas library was also screened, yielding clone CDPJ5.

To obtain copies of this transcript from a CF patient, a cDNA library from RNA of sweat gland epithelial cells from a patient was screened with the 0.75 kb Bam HI - Eco RI fragment from the 3' end of T16-1 and clones C16-1 and C1-1/5, which covered all but exon 1, were isolated. These two clones both exhibit a 3 bp deletion in exon 10 which is not present in any other clone containing that exon. Several clones, including CDLS26-1 from the lung library and T6/20 and T13-1 isolated from T84 were derived from partially processed transcripts. This was confirmed by genomic hybridization and by sequencing across the exon-intron boundaries for each clone. T11 also contained additional sequence at each end. T16-4.5 contained a small insertion near the boundary between exons 10 and 11 that did not correspond to intron sequence. Clones CDLS16A, 11a and 13a from the lung library also contained extraneous sequences of unknown origin. The clone C16-1 also contained a short insertion corresponding to a portion of the γ-transposon of E. coli; this element was not detected in the other clones.

The 5′ clones PA3-5, generated from pancreas RNA and TB2-7 generated from T84 RNA using the anchored PCR technique have identical sequences except for a single nucleotide difference in length at the 5′ end as shown in Figure 1. The 3′ clone, THZ-4 obtained from T84 RNA contains the 3′ sequence of the transcript in concordance with the genomic sequence of this region.

A combined sequence representing the presumptive coding region of the CF gene was generated from overlapping cDNA clones. Since most of the cDNA clones were apparently derived from unprocessed transcripts, further studies were performed to ensure the authenticity of the combined sequence. Each cDNA clone was first tested for localization to chromosome 7 by hybridization analysis with a human-hamster somatic cell hybrid containing a single human chromosome 7 and by pulsed field gel electrophoresis. Fine restriction enzyme mapping was also performed for each clone. While overlapping regions were clearly identifiable for most of the clones, many contained regions of unique restriction patterns.

To further characterize these cDNA clones, they were used as probes in gel hybridization experiments with EcoRI-or HindIII-digested human genomic DNA. As shown in Figure 9, five to six different restriction fragments could be detected with the 10-1 cDNA and a similar number of fragments with other cDNA clones, suggesting the presence of multiple exons for the putative CF gene. The hybridization studies also identified those cDNA clones with unprocessed intron sequences as they showed preferential hybridization to a subset of genomic DNA fragments. For the confirmed cDNA clones, their corresponding genomic DNA segments were isolated and the exons and exon/intron boundaries sequenced. As indicated in Figure 7, at least 28 exons have been identified which includes split exons 6a, 6b,

6c, 14a, 14b and 17a, 17b. Based on this information and the results of physical mapping experiments, the gene locus was estimated to span 250 kb on chromosome 7.

2.6 THE SEQUENCE

Figure 1 shows the nucleotide sequence of the cloned cDNA encoding CFTR together with the deduced amino acid sequence. The first base position corresponds to the first nucleotide in the 5′ extension clone PA3-5 which is one nucleotide longer than TB2-7. Arrows indicate position of transcription initiation site by primer extension analysis. Nucleotide 6129 is followed by a poly(dA) tract. Positions of exon junctions are indicated by vertical lines. Potential membrane-spanning segments were ascertained using the algorithm of Eisenberg et al J. Mol. Biol. 179:125 (1984). Potential membrane-spanning segments as analyzed and shown in Figure 11 are enclosed in boxes of Figure 1. In Figure 11, the mean hydropathy index [Kyte and Doolittle, J. Molec. Biol. 157: 105, (1982)] of 9 residue peptides is plotted against the amino acid number. The corresponding positions of features of secondary structure predicted according to Garnier et al, [J. Molec. Biol. 157, 165 (1982)] are indicated in the lower panel. Amino acids comprising putative ATP-binding folds are underlined in Figure 1. Possible sites of phosphorylation by protein kinases A (PKA) or C (PKC) are indicated by open and closed circles, respectively. The open triangle is over the 3bp (CTT) which are deleted in CF (see discussion below). The cDNA clones in Figure 1 were sequenced by the dideoxy chain termination method employing ^{35}S labelled nucleotides by the Dupont Genesis 2000™ automatic DNA sequencer.

The combined cDNA sequence spans 6129 base pairs excluding the poly(A) tail at the end of the 3′ untranslated region and it contains an ORF capable of

encoding a polypeptide of 1480 amino acids (Figure 1). An ATG (AUG) triplet is present at the beginning of this ORF (base position 133-135). Since the nucleotide sequence surrounding this codon (5'-AGACCAUGCA-3') has the proposed features of the consensus sequence (CC) A/GCCAUGG(G) of an eukaryotic translation initiation site with a highly conserved A at the -3 position, it is highly probable that this AUG corresponds to the first methionine codon for the putative polypeptide.

To obtain the sequence corresponding to the 5' end of the transcript, a primer-extension experiment was performed, as described earlier. As shown in Figure 10A, a primer extension product of approximately 216 nucleotides could be observed suggesting that the 5' end of the transcript initiated approximately 60 nucleotides upstream of the end of cDNA clone 10-1. A modified polymerase chain reaction (anchored PCR) was then used to facilitate cloning of the 5'-end sequences (Figure 10b). Two independent 5'-extension clones, one from pancreas and the other from T84 RNA, were characterized by DNA sequencing and were found to differ by only 1 base in length, indicating the most probable initiation site for the transcript as shown in Figure 1.

Since most of the initial cDNA clones did not contain a polyA tail indicative of the end of a mRNA, anchored PCR was also applied to the 3' end of the transcript (Frohman et al, 1988, supra). Three 3'-extension oligonucleotides were made to the terminal portion of the cDNA clone T16-4.5. As shown in Figure 10c, 3 PCR products of different sizes were obtained. All were consistent with the interpretation that the end of the transcript was approximately 1.2 kb downstream of the HindIII site at nucleotide position 5027 (see Figure 1). The DNA sequence derived from representative clones was in agreement with that of the T84 cDNA clone

T12a (see Figure 1 and 7) and the sequence of the corresponding 2.3 kb EcoRI genomic fragment.

3.0 MOLECULAR GENETICS OF CF

3.1 SITES OF EXPRESSION

To visualize the transcript for the putative CF gene, RNA gel blot hybridization experiments were performed with the 10-1 cDNA as probe. The RNA hybridization results are shown in Figure 8.

RNA samples were prepared from tissue samples obtained from surgical pathology or at autopsy according to methods previously described (A.M. Kimmel, S.L. Berger, eds. Meth. Enzymol. 152, 1987). Formaldehyde gels were transferred onto nylon membranes (Zetaprobe ™; BioRad Lab). The membranes were then hybridized with DNA probes labeled to high specific activity by the random priming method (A.P. Feinberg and B. Vogelstein, Anal. Biochem. 132, 6, 1983) according to previously published procedures (J. Rommens et al, Am. J. Hum. Genet. 43, 645-663, 1988). Figure 8 shows hybridization by the cDNA clone 10-1 to a 6.5kb transcript in the tissues indicated. Total RNA (10 μg) of each tissue, and Poly A+ RNA (1 μg) of the T84 colonic carcinoma cell line were separated on a 1% formaldehyde gel. The positions of the 28S and 18S rRNA bands are indicated. Arrows indicate the position of transcripts. Sizing was established by comparison to standard RNA markers (BRL Labs). HL60 is a human promyelocytic leukemia cell line, and T84 is a human colon cancer cell line.

Analysis reveals a prominent band of approximately 6.5 kb in size in T84 cells. Similar, strong hybridization signals were also detected in pancreas and primary cultures of cells from nasal polyps, suggesting that the mature mRNA of the putative CF gene is approximately 6.5 kb. Minor hybridization signals, probably representing degradation products, were detected at the lower size ranges but they varied

between different experiments. Identical results were obtained with other cDNA clones as probes. Based on the hybridization band intensity and comparison with those detected for other transcripts under identical experimental conditions, it was estimated that the putative CF transcripts constituted approximately 0.01% of total mRNA in T84 cells.

A number of other tissues were also surveyed by RNA gel blot hybridization analysis in an attempt to correlate the expression pattern of the 10-1 gene and the pathology of CF. As shown in Figure 8, transcripts, all of identical size, were found in lung, colon, sweat glands (cultured epithelial cells), placenta, liver, and parotid gland but the signal intensities in these tissues varied among different preparations and were generally weaker than that detected in the pancreas and nasal polyps. Intensity varied among different preparations, for example, hybridization in kidney was not detected in the preparation shown in Figure 8, but can be discerned in subsequent repeated assays. No hybridization signals could be discerned in the brain or adrenal gland (Figure 8), nor in skin fibroblast and lymphoblast cell lines.

In summary, expression of the CF gene appeared to occur in many of the tissues examined, with higher levels in those tissues severely affected in CF. While this epithelial tissue-specific expression pattern is in good agreement with the disease pathology, no significant difference has been detected in the amount or size of transcripts from CF and control tissues, consistent with the assumption that CF mutations are subtle changes at the nucleotide level.

3.2 THE MAJOR CF MUTATION

Figure 16 shows the DNA sequence at the F508 deletion. On the left, the reverse complement of the sequence from base position 1649-1664 of the normal sequence (as derived from the cDNA clone T16). The nucleotide sequence is displayed as the output (in arbitrary fluorescence intensity units, y-axis) plotted against time (x-axis) for each of the 2 photomultiplier tubes (PMT#1 and #2) of a Dupont Genesis 2000^{TM} DNA analysis system. The corresponding nucleotide sequence is shown underneath. On the right is the same region from a mutant sequence (as derived from the cDNA clone C16). Double-stranded plasmid DNA templates were prepared by the alkaline lysis procedure. Five μg of plasmid DNA and 75 ng of oligonucleotide primer were used in each sequencing reaction according to the protocol recommended by Dupont except that the annealing was done at 45°C for 30 min and that the elongation/termination step was for 10 min at 42°C. The unincorporated fluorescent nucleotides were removed by precipitation of the DNA sequencing reaction product with ethanol in the presence of 2.5 M ammonium acetate at pH 7.0 and rinsed one time with 70% ethanol. The primer used for the T16-1 sequencing was a specific oligonucleotide 5'GTTGGCATGCTTTGATGACGCTTC3' spanning base position 1708 - 1731 and that for C16-1 was the universal primer SK for the Bluescript vector (Stratagene).

Figure 17 also shows the DNA sequence around the F508 deletion, as determined by manual sequencing. The normal sequence from base position 1726-1651 (from cDNA T16-1) is shown beside the CF sequence (from cDNA C16-1). The left panel shows the sequences from the coding strands obtained with the B primer (5'GTTTTCCTGGATTATGCCTGGCAC3') and the right panel those from the opposite strand with the D primer

(5'GTTGGCATGCTTTGATGACGCTTC3'). The brackets indicate the three nucleotides in the normal that are absent in CF (arrowheads). Sequencing was performed as described in F. Sanger, S. Nicklen, A. R. Coulsen, Proc. Nat. Acad. Sci. U. S. A. 74: 5463 (1977).

The extensive genetic and physical mapping data have directed molecular cloning studies to focus on a small segment of DNA on chromosome 7. Because of the lack of chromosome deletions and rearrangements in CF and the lack of a well-developed functional assay for the CF gene product, the identification of the CF gene required a detailed characterization of the locus itself and comparison between the CF and normal (N) alleles. Random, phenotypically normal, individuals could not be included as controls in the comparison due to the high frequency of symptomless carriers in the population. As a result, only parents of CF patients, each of whom by definition carries an N and a CF chromosome, were suitable for the analysis. Moreover, because of the strong allelic association observed between CF and some of the closely linked DNA markers, it was necessary to exclude the possibility that sequence differences detected between N and CF were polymorphisms associated with the disease locus.

3.3 IDENTIFICATION OF RFLPs AND FAMILY STUDIES

To determine the relationship of each of the DNA segments isolated from the chromosome walking and jumping experiments to CF, restriction fragment length polymorphisms (RFLPs) were identified and used to study families where crossover events had previously been detected between CF and other flanking DNA markers. As shown in Figure 14, a total of 18 RFLPs were detected in the 500 kb region; 17 of them (from E6 to CE1.0) listed in Table 2; some of them correspond to markers previously reported.

Five of the RFLPs, namely 10-1X.6, T6/20, H1.3 and CE1.0, were identified with cDNA and genomic DNA probes derived from the putative CF gene. The RFLP data are presented in Table 2, with markers in the MET and D7S8 regions included for comparison. The physical distances between these markers as well as their relationship to the MET and D7S8 regions are shown in Figure 14.

TABLE 2. RFLPs ASSOCIATED WITH THE CF GENE

Probe name	Enzyme	Frag-length	N[a]	CF-PI[a]	A[b]	*[c]	Reference
metD	BanI	7.6(kb)	28	48	0.60	0.10	J.E. Spence et al, Am. J. Hum. Genet 39:729 (1986)
		6.8	59	25			
metD	TaqI	6.2	74	75	0.66	0.06	R. White et al, Nature 318:382 (1985
		4.8	19	4			
metH	TaqI	7.5	45	49	0.35	0.05	White et al, supra
		4.0	38	20			
E6	TaqI	4.4	58	62	0.45	0.06	B. Keren et al, Am. J. Hum. Genet. 44:827 (1989)
		3.6	42	17			
E7	TaqI	3.9	40	16	0.47	0.07	
		3+0.9	51	57			

TABLE 2 (continued)

pH131	HinfI	0.4	81	33	0.73	0.15	J.M. Rommens et al, Am. J. Hum. Genet. 43:645 (1988)
		0.3	18	47			
W3D1.4	HindIII	20	82	33	0.68	0.13	B. Kerem et al, supra
		10	22	47			
H2.3A	TaqI	2.1	39	53	0.64	0.09	X. Estivill et al, Nature 326:840 (1987); X. Estivill et al, Genomics 1:257 (1987)
(XV2C)		1.4	37	11			
EG1.4	HincII	3.8	31	69	0.89	0.17	
		2.8	56	7			
EG1.4	BglI	20	27	69	0.89	0.18	
		15	62	9			
JG2E1	PstI	7.8	69	10	0.88	0.18	X. Estivill et al supra and B. Kerem et al supra

TABLE 2 (continued)

(KM19)		6.6	30	70			
E2.6/E.9	MspI	13	34	6	0.85	0.14	
		8.5	26	55			
H2.8A	NcoI	25	22	55	0.87	0.18	
		8	52	9			
E4.1	MspI	12	37	8	0.77	0.11	G. Romeo, personal communication
(Mp6d9)		8.5+3.5	38	64			
J44	XbaI	15.3	40	70	0.86	0.13	
		15+.3	44	6			
10-1X.6	AccI	6.5	67	15	0.90	0.24	
		3.5+3	14	60			
10-1X.6	HaeIII	1.2	14	61	0.91	0.25	
		.6	72	15			
T6/20	MspI	8	56	66	0.51	0.54	
		4.3	21	8			
H1.3	NcoI	2.4	53	7	0.87	0.15	
		1+1.4	35	69			

TABLE 2 (continued)

CE1.0	NdeI	5.5	81	73	0.41	0.03	
		4.7+0.8	8	3			
J32	SacI	15	21	24	0.17	0.02	M.C. Iannuzi et al Am. J. Genet. 44:695 (1989)
		6	47	38			
J3.11	MspI	4.2	36	38	0.29	0.04	B.J. Wainright et al, Nature 318:384 (1985)
		1.8	62	36			
J29	PvuII	9	26	36	0.36	0.06	M.C. Iannuzi et al, supra
		6	55	36			

NOTES FOR TABLE 2

(a) The number of N and CF-PI (CF with pancreatic insufficiency) chromosomes were derived from the parents in the families used in linkage analysis [Tsui et al, Cold Spring Harbor Symp. Quant. Biol. 51:325 (1986)].

(b) Standardized association (A), which is less influenced by the fluctuation of DNA marker allele distribution among the N chromosomes, is used here for the comparison Yule's association coefficient A=(ad-bc)/(ad+bc), where a, b, c, and d are the number of N chromosomes with DNA marker allele 1, CF with 1, N with 2, and CF with 2 respectively. Relative risk can be calculated using the relationship RR = (1+A)/(1-A) or its reverse.

(c) Allelic association (*), calculated according to A. Chakravarti et al, Am. J. Hum. Genet. 36:1239, (1984) assuming the frequency of 0.02 for CF chromosomes in the population is included for comparison.

Because of the small number of recombinant families available for the analysis, as was expected from the close distance between the markers studied and CF, and the possibility of misdiagnosis, alternative approaches were necessary in further fine mapping of the CF gene.

3.4 ALLELIC ASSOCIATION

Allelic association (linkage disequilibrium) has been detected for many closely linked DNA markers. While the utility of using allelic association for measuring genetic distance is uncertain, an overall correlation has been observed between CF and the flanking DNA markers. A strong association with CF was noted for the closer DNA markers, D7S23 and D7S122,

whereas little or no association was detected for the more distant markers MET, D7S8 or D7S424 (see Figure 1).

As shown in Table 2, the degree of association between DNA markers and CF (as measured by the Yule's association coefficient) increased from 0.35 for metH and 0.17 for J32 to 0.91 for 10-1X.6 (only CF-PI patient families were used in the analysis as they appeared to be genetically more homogeneous than CF-PS). The association coefficients appeared to be rather constant over the 300 kb from EG1.4 to H1.3; the fluctuation detected at several locations, most notably at H2.3A, E4.1 and T6/20, were probably due to the variation in the allelic distribution among the N chromosomes (see Table 2). These data are therefore consistent with the result from the study of recombinant families (see Figure 14). A similar conclusion could also be made by inspection of the extended DNA marker haplotypes associated with the CF chromosomes (see below). However, the strong allelic association detected over the large physical distance between EG1.4 and H1.3 did not allow further refined mapping of the CF gene. Since J44 was the last genomic DNA clone isolated by chromosome walking and jumping before a cDNA clone was identified, the strong allelic association detected for the JG2E1-J44 interval prompted us to search for candidate gene sequences over this entire interval. It is of interest to note that the highest degree of allelic association was, in fact, detected between CF and the 2 RFLPs detected by 10-1X.6, a region near the major CF mutation.

Table 3 shows pairwise allelic association between DNA markers closely linked to CF. The average number of chromosomes used in these calculations was 75-80 and only chromosomes from CF-PI families were used in scoring CF chromosomes. Similar results were obtained when Yule's standardized association (A) was used).

TABLE 3

N chromosomes

CF chromosomes	metD BanI	metH TaqI	E6 TaqI	E7 TaqI	pH131 HinfI	W3 D1.4 HdIII	H2.3A TaqI	EG1.4 HcII	JG2E1 BglI	E2.6 PstI	H2.8 MspI	E4.1 NcoI	J44 MspI	10-1X.6 XbaI	T6/20 AccI	H1.3 HaeIII	CE1.0 MspI	J32 NcoI	J3.11 NdeI	J29 SacI	MspI	PvuII
metD BanI	-	0.35	0.49	0.04	0.04	0.05	0.07	0.27	0.06	0.06	0.07	0.14	0.07	0.09	0.03	0.06	0.10	0.03	0.16	0.05	0.07	0.11	0.02
metD TaqI	0.21	-	0.41	0.13	0.15	0.02	0.01	0.02	0.09	0.15	0.11	0.07	0.24	0.03	0.11	0.08	0.02	0.06	0.13	0.15	0.09	0.09	0.05
metH TaqI	0.81	0.14	-	0.01	0.05	0.06	0.06	0.24	0.05	0.08	0.07	0.13	0.15	0.07	0.04	0.02	0.02	0.07	0.02	0.03	0.21	0.04	0.18
E6 TaqI	0.11	0.30	0.00	-	0.89	0.07	0.06	0.04	0.02	0.03	0.00	0.19	0.02	0.09	0.19	0.09	0.11	0.09	0.15	0.07	0.11	0.20	0.00
E7 TaqI	0.16	0.31	0.02	1.00	-	0.11	0.09	0.03	0.03	0.04	0.01	0.11	0.00	0.07	0.22	0.01	0.02	0.09	0.13	0.06	0.06	0.16	0.04
pH131 HinfI	0.45	0.28	0.23	0.38	0.40	-	0.91	0.12	0.04	0.09	0.05	0.08	0.03	0.03	0.08	0.16	0.15	0.20	0.04	0.03	0.06	0.08	0.06
W3D1.4 HindIII	0.45	0.28	0.23	0.45	0.47	0.95	-	0.21	0.02	0.03	0.01	0.06	0.03	0.03	0.10	0.12	0.10	0.23	0.10	0.05	0.05	0.10	0.06
H2.3A TaqI	0.20	0.11	0.15	0.08	0.11	0.36	0.47	-	0.05	0.11	0.07	0.42	0.14	0.29	0.07	0.27	0.22	0.20	0.09	0.23	0.04	0.08	0.12
EG1.4 HincII	0.11	0.06	0.07	0.06	0.07	0.20	0.20	0.24	-	0.95	0.87	0.76	0.86	0.81	0.60	0.07	0.13	0.61	0.56	0.04	0.24	0.14	0.15
EG1.4 BglI	0.03	0.06	0.07	0.06	0.07	0.27	0.27	0.40	1.00	-	0.92	0.77	0.93	0.71	0.55	0.06	0.07	0.56	0.55	0.12	0.28	0.24	0.20
JG2E1 PstI	0.07	0.06	0.03	0.09	0.06	0.30	0.30	0.45	0.93	0.94	-	0.84	1.00	0.76	0.64	0.11	0.11	0.61	0.57	0.13	0.31	0.26	0.22
E2.6/E.9 MspI	0.22	0.06	0.07	0.02	0.03	0.20	0.20	0.34	0.81	0.82	0.92	-	0.83	0.97	0.76	0.56	0.52	0.47	0.70	0.32	0.31	0.25	0.22
H2.8 NcoI	0.05	0.07	0.01	0.08	0.06	0.31	0.31	0.45	0.92	0.93	1.00	0.92	-	0.74	0.65	0.13	0.18	0.60	0.59	0.10	0.28	0.28	0.18
E4.1 MspI	0.12	0.06	0.07	0.05	0.03	0.25	0.25	0.46	0.82	0.86	0.94	1.00	0.93	-	0.71	0.49	0.49	0.49	0.68	0.35	0.27	0.25	0.21
J44 XbaI	0.18	0.05	0.06	0.01	0.01	0.26	0.26	0.43	0.71	0.89	0.80	0.90	0.80	0.85	-	0.33	0.40	0.65	0.64	0.32	0.24	0.22	0.23
10-1X.6 AccI	0.16	0.10	0.24	0.10	0.11	0.42	0.42	0.64	0.54	0.58	0.64	0.70	0.69	0.69	0.59	-	0.91	0.19	0.36	0.56	0.00	0.02	0.03
10-1X.6 HaeIII	0.16	0.10	0.25	0.06	0.11	0.41	0.41	0.65	0.54	0.58	0.64	0.70	0.69	0.69	0.59	1.00	-	0.18	0.43	0.62	0.02	0.02	0.08
T6/20 MspI	0.27	0.07	0.36	0.13	0.13	0.23	0.23	0.29	0.05	0.00	0.01	0.07	0.02	0.01	0.11	0.69	0.69	-	0.56	0.03	0.21	0.18	0.25
H1.3 NcoI	0.06	0.06	0.06	0.09	0.01	0.30	0.30	0.55	0.71	0.76	0.87	0.90	0.87	0.93	0.92	0.64	0.64	0.12	-	0.40	0.19	0.13	0.20
CE1.0 NdeI	0.00	0.04	0.02	0.11	0.11	0.25	0.25	0.08	0.69	0.59	0.55	0.43	0.55	0.37	0.44	0.24	0.24	0.07	0.40	-	0.19	0.20	0.14
J32 SacI	0.03	0.13	0.07	0.17	0.13	0.17	0.24	0.07	0.21	0.21	0.24	0.22	0.24	0.21	0.21	0.27	0.26	0.13	0.21	0.18	-	0.84	0.97
J3.11 MspI	0.14	0.11	0.15	0.07	0.06	0.05	0.05	0.12	0.11	0.10	0.13	0.18	0.19	0.15	0.20	0.28	0.29	0.24	0.14	0.07	0.81	-	0.71
J29 PvuII	0.11	0.12	0.09	0.10	0.10	0.00	0.00	0.09	0.10	0.10	0.14	0.17	0.20	0.16	0.16	0.29	0.29	0.23	0.16	0.06	0.85	0.97	-

Strong allelic association was also detected among subgroups of RFLPs on both the CF and N chromosomes. As shown in Table 3, the DNA markers that are physically close to each other generally appeared to have strong association with each other. For example, strong (in some cases almost complete) allelic association was detected between adjacent markers E6 and E7, between pH131 and W3D1.4 between the AccI and HaeIII polymorphic sites detected by 10-1X.6 and amongst EG1.4, JG2E1, E2.6(E.9), E2.8 and E4.1. The two groups of distal markers in the MET and D7S8 region also showed some degree of linkage disequilibrium among themselves but they showed little association with markers from E6 to CE1.0, consistent with the distant locations for MET and D7S8. On the other hand, the lack of association between DNA markers that are physically close may indicate the presence of recombination hot spots. Examples of these potential hot spots are the region between E7 and pH131, around H2.3A, between J44 and the regions covered by the probes 10-1X.6 and T6/20 (see Figure 14). These regions, containing frequent recombination breakpoints, were useful in the subsequent analysis of extended haplotype data for the CF region.

3.5 HAPLOTYPE ANALYSIS

Extended haplotypes based on 23 DNA markers were generated for the CF and N chromosomes in the collection of families previously used for linkage analysis. Assuming recombination between chromosomes of different haplotypes, it was possible to construct several lineages of the observed CF chromosomes and, also, to predict the location of the disease locus.

To obtain further information useful for understanding the nature of different CF mutations, the F508 deletion data were correlated with the extended DNA marker haplotypes. As shown in Table 4, five major groups of N and CF haplotypes could be defined by the RFLPs within or immediately adjacent to the putative CF gene (regions 6-8).

Table 4 DNA marker haplotypes spanning the CF locus.

	HAPLOTYPES[a]									CF[b]				
										PI	PS	PI	PS	
	1	2	3	4	5	6	7	8	9	(FS08)[c]	(FS08)[c]	others	others	N
I. (a)	A	A	A	A	A	A	A	A	A	10	1	-	-	-
	A	A	A	A	A	A	-	A	A	3	-	-	-	-
	A	A	A	A	-	A	A	-	A	1	-	-	-	1
	A	A	A	A	-	-	A	-	A	-	-	-	-	1
	A	A	A	A	A	A	A	A	B	10	-	-	-	-
	A	A	A	-	A	A	A	A	B	4	-	-	-	-
	A	A	A	A	-	A	A	A	B	1	-	-	-	-
	A	A	A	-	A	A	A	A	C	1	-	-	-	-
	B	A	A	A	A	A	A	A	A	4	-	-	-	-
	B	A	A	-	A	A	A	A	A	1	-	-	-	-
	B	A	A	A	A	A	A	A	-	-	1	-	-	-
	B	A	A	A	-	A	A	A	-	1	-	-	-	-
	B	A	B	A	A	A	A	A	-	1	-	-	-	-
	A	A	D	A	A	A	A	A	A	1	-	-	-	-
	A	A	G	A	A	A	A	A	A	1	-	-	-	-
	B	B	C	A	A	A	A	A	A	1	-	-	-	-
	B	E	C	B	A	-	A	-	A	2	-	-	-	-
	D	D	B	A	A	A	-	A	A	1	-	-	-	-
	B	C	A	A	A	A	A	A	B	1	-	-	-	-
	C	A	D	C	A	A	A	A	B	1	-	-	-	-
	A	D	B	D	A	A	A	A	B	1	-	-	-	-
	D	B	A	B	A	A	A	-	B	1	-	-	-	1
	B	A	A	B	A	A	A	A	B	2	-	-	-	-
	A	A	A	E	A	A	A	A	B	1	1	-	-	-
	A	A	A	C	A	A	A	A	B	1	-	-	-	-
	A	A	A	C	B	A	A	A	B	1	-	-	-	1
	A	B	C	B	A	-	A	A/D	B	1	-	-	-	-
(b)	A	C	-	A	A	A	A	A	A	-	-	-	1	-
	A	C	A	A	A	A	A	A	-	-	-	1	-	-
	D	C	-	A	A	A	A	A	B	-	-	1	-	1
	D	C	-	A	A	A	A	A	B	-	-	1	-	-
	F	C	A	A	A	A	A	A	B	-	-	3	-	-
(c)	B	C	A	B	C	A	A	D	A	-	-	-	-	1
	B	C	A	B	C	A	A	D	B	-	-	1	-	-
	F	C	A	B	C	A	A	D	B	-	-	-	-	1
	A	B	A	B	C	A	A	D	B	-	-	-	-	1
	B	D	A	B	C	A	A	D	C	-	-	-	-	1
	A	B	A	B	A	A	-	D	A	-	-	-	-	1

										0	0	6	4	45
III. (a)	B	C	B	A	A	C	B	A	B	1	.	.	.	.
(b)	B	A	B	A	A	C	B	A	B	.	.	1	.	.
	B	C	B	A	A	C	B	A	A	.	.	1	.	.
	B	C	.	A	A	C	B	A	B	.	.	.	.	1
	A	B	.	A	A	C	B	A	B	.	.	.	.	2
	A	B	.	A	A	C	B	A	C	.	.	.	.	1
	B	B	B	A	A	C	B	A	B	.	.	.	.	1
	D	C	B	A	A	C	B	A	A	.	.	.	1	2
	A	B	B	A	A	C	B	A	B	.	.	.	.	1
	B	B	A	A	A	C	B	A	B	.	.	2	.	1
	B	B	A	A	A	C	B	A	B	.	.	1	.	1
	D	A	A	A	A	C	B	A	B	.	.	.	1	.
	D	C	A	A	A	C	B	A	B	.	.	.	.	1
	A	C	B	A	A	C	B	A	B	.	.	1	.	2
	D	B	A	A	A	C	B	A	A	.	.	.	.	1
(c)	A	A	A	B	B	C	B	A	.	.	.	1	.	.
	F	B	B	B	B	C	B	A	B	.	.	.	.	1
	D	B	B	B	B	C	B	A	A	.	.	.	.	1
										1	0	7	2	17
IV.	F	C	B	A	A	C	B	C	A	.	.	.	1	.
	B	C	A	A	A	C	B	C	.	.	.	.	.	1
	A	B	A	A	A	C	.	C	B	.	.	.	.	1
	A	H	B	A	.	C	.	C	B	.	.	.	.	1
	D	B	B	B	B	C	B	C	B	.	.	.	.	1
										0	0	0	1	4
V. (a)	B	C	B	B	B	C	A	C	A	.	.	1	.	.
	A	C	B	B	.	C	A	C	A	.	.	1	.	.
	B	B	B	B	B	C	A	C	B	.	.	.	.	1
	B	B	C	B	B	C	A	C	B	.	.	.	.	1
	D	C	.	A	B	C	A	C	B	.	.	.	.	1
(b)	B	C	A	B	C	C	A	C	A	.	.	.	1	.
	B	C	.	B	C	C	.	C	D	.	.	.	.	1
										0	0	2	1	5
Others:	B	C	B	A	A	B	B	A	B	.	.	.	.	1
	B	C	B	A	A	D	B	A	B	.	.	.	.	1
	B	C	B	A	B	A	B	D	A	.	.	.	.	1
	B	D	A	B	B	F	A	C	D	.	.	.	.	1
	A	C	B	A	A	C	B	C	A	.	.	.	.	1

TABLE 4 (continued)

(d)	D B	B C	A B	A C	A A	A A	A A	C C	A B	.	.	.	.		1 1	
										57	5	7	1		14	
II. (a)	B	A	-	B	B	B	A	C	B	.	.	1	.		.	
	B	B/C	B	B	B	B	A	C	B	.	.	1	.		.	
	B	A	B	B	B	B	A	A/C	B	.	.	.	1		.	
	A	B	B	B	B	B	A	C	B	.	.	1	.		.	
	B	B	B	B	B	B	A	C	A	.	.	.	.		3	
	A	C	B	B	B	B	A	C	A	.	.	.	.		1	
	A	C	B	B	B	B	A	C	A	.	.	.	.		1	
	F	C	B	B	B	B	A	C	B	.	.	.	.		1	
	A	C	B	B	B	B	A	C	C	.	.	.	.		1	
	B	C	B	B	B	B	A	C	C	.	.	.	1		.	
	B	C	B	B	B	B	A	C	B	.	.	.	.		1	
	B	C	B	B	B	B	A	C	D	.	.	.	.		1	
	B	C	B	B	B	B	A	C	B	.	.	.	.		1	
	D	C	B	B	B	B	A	C	B	.	.	.	.		2	
	D	C	B	B	B	B	A	C	B	.	.	.	.		1	
	F	C	B	B	B	B	A	C	B	.	.	.	.		1	
	C	A	A	B	B	B	A	C	B	.	.	.	.		1	
	B	G	A	B	B	B	A	C	B	.	.	.	.		1	
	F	H	-	B	B	B	A	C	B	.	.	.	.		1	
	B	B	A	B	B	B	A	C	B	.	.	.	.		1	
	A	D	A	B	B	B	A	C	B	.	.	.	.		1	
	F	D	A	B	B	B	A	C	A	.	.	.	.		1	
	C	C	A	B	B	B	A	C	A	.	.	.	.		1	
	B	C	A	B	B	B	A	C	A	.	.	.	.		2	
	A	C	A	B	B	B	A	C	B	.	.	.	.		1	
	A	C	A	B	B	B	A	C	B	.	.	.	.		1	
	B	B/C	A	B	B	B	A	C	C	.	.	.	1		2	
	C	C	A	B	B	B	A	A/C	B	.	.	.	.		1	
	D	C	A	B	B	B	A	C	A	.	.	.	.		1	
	A	C	A	B	B	B	A	A/C	A	.	.	.	.		1	
	A	C	B	D	B	B	A	C	B	.	.	1	.		1	
	B	C	B	B	B	B	A	C	C	.	.	.	.		1	
	A	C	B	B	B	B	A	C	D	.	.	.	1		.	
	A	B/C	A	B	C	B	B/C	C	A	.	.	.	.		1	
(b)	A	C	A	B	B	B	A	B	B	.	.	1	.		.	
	A	C	-	B	B	B	A	B	B	.	.	1	.		.	
(c)	B	D	-	B	-	B	A	A	A	.	.	.	.		1	

G B B A A B/C A A/D B	.	.	.	.	1	
	0	0	0	0	7	
Unclassified	4	10	2	18	6	
Total:	62	15	24	27	98	

(a) The extended haplotype data are derived from the CF families used in previous linkage studies (see footnote (a) of Table 3) with additional CF-PS families collected subsequently (Kerem et al, Am . J. Genet. 44:827 (1989)). The data are shown in groups (regions) to reduce space. The regions are assigned primarily according to pairwise association data shown in Table 3 with regions 6-8 spanning the putative CF locus (the F508 deletion is between regions 6 and 7). A dash (-) is shown at the region where the haplotype has not been determined due to incomplete data or inability to establish phase. Alternative haplotype assignments are also given where data are incomplete. Unclassified includes those chromosomes with more than 3 unknown assignments. The haplotype definitions for each of the 9 regions are:

Region 1-

	metD BanI	metD TaqI	metH TaqI
A =	1	1	1
B =	2	1	2
C =	1	1	2
D =	2	2	1
E =	1	2	-
F =	2	1	1
G =	2	2	2

Region 2-

	E6 TaqI	E7 TaqI	pH131 HincII	W3D1.4 HindIII
A =	1	2	2	2
B =	2	1	1	1
C =	1	2	1	1
D =	2	1	2	2
E =	2	2	2	1
F =	2	2	1	1
G =	1	2	1	2
H =	1	1	2	2

Region 3-

	H2.3A TaqI
A =	1
B =	2

Region 4-

	EG1.4 HincII	EG1.4 BglII	JG2B1 PstI
A =	1	1	2

B				
C	=	2	2	2
D	=	1	1	1
E	=	1	2	1

Region 5-

		E2.6 MspI	E2.8 NcoI	E4.1 MspI
A	=	2	1	2
B	=	1	2	1
C	=	2	2	2

Region 6-

		J44 XbaI	10-1X.6 AccI	10-1X.6 HaeIII
A	=	1	2	1
B	=	2	1	2
C	=	1	1	2
D	=	1	2	2
E	=	2	2	2
F	=	2	2	1

Region 7-

		T6/20 MspI
A	=	1
B	=	2

Region 8-

		H1.3 NcoI	CB1.0 NdeI
A	=	2	1
B	=	1	2
C	=	1	1
D	=	2	2

Region 9-

		J32 SacI	J3.11 MspI	J29 PvuII
A	=	1	1	1
B	=	2	2	2
C	=	2	1	2
D	=	2	2	1
E	=	2	1	1

(b) Number of chromosomes scored in each class:

CF-PI(F)	=	CF chromosomes from CF-PI patients with the F508 deletion;
CF-PS(F)	=	CF chromosomes from CF-PS patients with the F508 deletion;
CF-PI	=	Other CF chromosomes from CF-PI patients;
CF-PS	=	Other CF chromosomes from CF-PS patients;
N	=	Normal chromosomes derived from carrier parents.

It was apparent that most recombinations between haplotypes occurred between regions 1 and 2 and between regions 8 and 9, again in good agreement with the relatively long physical distance between these regions. Other, less frequent, breakpoints were noted between short distance intervals and they generally corresponded to the hot spots identified by pairwise allelic association studies as shown above. It is of interest to note that the F508 deletion associated almost exclusively with Group I, the most frequent CF haplotype, supporting the position that this deletion constitutes the major mutation in CF. More important, while the F508 deletion was detected in 89% (62/70) of the CF chromosomes with the AA haplotype (corresponding to the two regions, 6 and 7) flanking the deletion, none was found in the 14 N chromosomes within the same group ($\chi^2 = 47.3$, $p < 10^{-4}$). The F508 deletion was therefore not a common sequence polymorphism associated with the core of the Group I haplotype (see Table 5).

Together, the results of the oligonucleotide hybridization study and the haplotype analysis support the fact that the gene locus described here is the CF gene and that the 3 bp (F508) deletion is the most common mutation in CF.

3.6 INTRON/EXON BOUNDARIES

Genomic CF gene includes all of the regulatory genetic information as well as intron genetic information which is spliced out in the expression of the CF gene. It has been found that portions of the introns at the intron/exon boundaries for the exons of the CF gene are very helpful in not only locating mutations in the CF gene, but are also very useful in PCR analysis. Such intron information can be employed in PCR analysis for purposes of CF screening which will be discussed in more detail in a later section. As set out in Figure 18 with the headings "Exon 4 and 6 through

Exon 24", there are portions of the bounding introns which are particularly preferred in PCR exon amplification.

In sequencing the intron portions, it has been determined that there are at least 28 exons instead of the previously reported 24 exons in applicants' aforementioned co-pending applications. Exons 6, 14 and 17, as previously reported, are found to be in segments and are now named exons 6a, 6b, 6c and 14a and 14b and exons 17a and 17b.

The intron portions, which have been used in PCR amplification, are identified by arrows in Figure 18. The portions identified by the arrows are preferred, but it is understood that other portions of the intron sequences are also useful in the PCR amplification technique. For example, for exon 9 the relevant genetic information which is preferred in PCR is noted by reference to the 5' and 3' ends of the sequence. The intron section is identified with an "i". Hence in Figure 18 for exon 9, the preferred portions are identified by the arrows 9i-5 and 9i-3. They include the sequence TAA...TGT for 9i-5 and for 9i-3 TGT...CGT. Similarly, the preferred portions of the intron for exon 10 is identified as 10i-5 and 10i-3 and similarly for exons 11 through 24. Similar regions are also identified and/or can be extracted from the bordering intron regions for exons 4 and 6 through 8 of Figure 18.

These oligonucleotides, as derived from the intron sequence, assist in amplifying by PCR the respective exon, thereby providing for analysis for DNA sequence alterations corresponding to mutations of the CF gene. The mutations can be revealed by either direct sequence determination of the PCR products or sequencing the products cloned in plasmid vectors. The amplified exon can also be analyzed by use of gel electrophoresis in the manner to be further described. It has been found that the sections of the intron for each respective exon

are of sufficient length to work particularly well with PCR technique to provide for amplification of the relevant exon.

3.7 CF MUTATIONS - ΔI506 OR ΔI507

The association of the F508 deletion with 1 common and 1 rare CF haplotype provided further insight into the number of mutational events that could contribute to the present patient population. Based on the extensive haplotype data, the original chromosome in which the F508 deletion occurred is likely to carry the haplotype - AAAAAAA- (Group Ia), as defined in Table 4. The other Group I CF chromosomes carrying the deletion are probably recombination products derived from the original chromosome. If the CF chromosomes in each haplotype group are considered to be derived from the same origin, only 3-4 additional mutational events would be predicted (see Table 4). However, since many of the CF chromosomes in the same group are markedly different from each other, further subdivision within each group is possible. As a result, a higher number of independent mutational events could be considered and the data suggest that at least 7 additional, putative mutations also contribute to the CF-PI phenotype (see Table 3). The mutations leading to the CF-PS subgroup are probably more heterogeneous.

The 7 additional CF-PI mutations are represented by the haplotypes: -CAAAAAA- (Group Ib), -CABCAAD- (Group Ic), ---BBBAC- (Group IIa), -CABBBAB- (Group Va). Although the molecular defect in each of these mutations has yet to be defined, it is clear that none of these mutations severely affect the region corresponding to the oligonucleotide binding sites used in the PCR/hybridization experiment.

One CF chromosome hydridizing to the ΔF508-ASO probe, however, has been found to associate with a different haplotype (group IIIa). It appeared that the

ΔI508 should have appeared in both haplotypes, but with the discovery of ΔI507, it is not discovered that it is not. Instead, the ΔI507 is in group I, whereas the ΔI507 is in group IIIa. None of the other CF nor the normal chromosomes of this hoplotype group have shown hybridization to the mutant (ΔF508) ASO [B. Kerem et al, Science 245:1073 (1989)]. In view of the group Ia and IIIa haplotypes being distinctly different from each other, the mutations harbored by these two groups of CF chormosomes must have originated independently. To investigate the molecular nature of the mutation in this group IIIa CF chromosome, we further characterized the region of interest through amplification of the genomic DNA from an individual carrying the chromosome IIIa by the polymerase chain reaction (PCR).

These polymerase chains reactions (PCR) were performed according to the procedure of R.k. Saiki et al Science 230:1350 (1985). A specific DNA segment of 491 bp including exon 10 of the CF gene was amplified with the use of the oligonucleotide primers 10i-5 (5'-GCAGAGTACCTGAAACAGGA-3') and 10i-3 (5'ATGGGTAAGCTACTGTGAATG-3')located in the 5' and 3' flanking regions, respectively, as shown in Figure 18. Both oligonucleotides were purchased from the HSC DNA Biotechnology Service Center (Toronto). Approximately 500 ng of genomic DNA from cultured lymphoblastoid cell lines of the parents and the CF child of Family 5 were used in each reaction. The DNA samples were denatured at 94°C for 30 sec., primers annealed at 55°C for 30 sec., and extended at 72°C for 50 sec. (with 0.5 unit of Taq polymerase, Perkin-Elmer/Cetus, Norwalk, CT) for 30 cycles and a final extension period of 7 min. in a Perkin-Elmer/Cetus DNA Thermal Cycler.

Hydridization analysis of the PCR products from three individuals of Family 5 of group IIIa was performed. The carrier mother and father are

represented by a half-filled circle and square, respectively, and the affected son is a filled square in Figure 19a. The conditions for hybridizaton and washing have been previously described (Kerem et al, supra). There is a relatively weak signal in the father's PCR product with the mutant (oligo ΔF508) probe. In Figure 19c, DNA sequence analysis of the clone 5-3-15 and the PCR products from the affected son and the carrier father. The arrow in the center panel indicates the presence of both A and T nucleotide residue in the same position; the arrow in the right panel indicates the points of divergence between the normal and the ΔI507 sequence. The sequence ladders shown are derived from the reverse-complements as will be described later. Figure 19b shown the DNA sequences and their corresponding amino acid sequences of the normal ΔI507 and ΔI508 alleles spanning the mutation sites are shown. With reference to Figure 19a, the PCR-amplified DNA from the carrier father, who contributed the group IIIa CF chromosome to the affected son, hybridized less efficiently with the ΔF508 ASO than that from the mother who carried the group Ia CF chromosome. The difference became apparent when the hybridization signals were compared to that with the normal ASO probe. This result therefore indicated that the mutation carried by the group IIIa CF chromosome might not be identical to ΔF508.

To define the nucleotide sequence corresponding to the mutant allele on this chromosome, the PCR-amplified product of the father's DNA was excised from a polyacrylamide-electrophoretic gel and cloned into a sequencing vector.

The general procedures for DNA isolation and purification for purposes of cloning into a sequencing vector are described in J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual, 2nd

ed. (Cold Spring Harbor Press, N.Y. 1989). The two homoduplexes generated by PCR amplification of the paternal DNA were purified from a 5% non-denaturing polyarylamide gel (30:1 acrylamide:bis-acrylamide). The appropriate bands were visualized by staining with ethidium bromide, excised and eluted in TE (10 mM Tris-HCl; 1mM EDTA; pH 7.5) for 2 to 12 hours at room temperature. The DNA solution was sequentially treated with Tris-equilibrated phenol, phenol/$CHCl_3$ and $CHCl_3$. The DNA samples were concentrated by precipitation in ethanol and resuspension in TE, incubated with T4 polynucleotide kinase in the presence of ATP, and ligated into diphosphorylated, blunt-ended Bluescript KS™ vector (Stratagene, San Diego, CA). Clones containing amplified product generated from the normal parental chromosome was identified by hybridization with the oligonucleotide N as described in Kerem et al supra.

Clones containing the mutant sequence were identified by their failure to hybridize to the normal ASO (Kerem et al, supra). One clone, 5-3-15 was isolated and its DNA sequence determined. The general protocol for sequencing cloned DNA is essentially as described [J.R. Riordan et al, Science 245:1066 (1989)] with the use of an U.S. Biochemicals Sequenase™ kit. To verify the sequence and to exclude any errors introduced by DNA polymerase during PCR, the DNA sequences for the PCR products from the father and one of the affected children were also determined directly without cloning.

This procedure was accomplished by denaturing 2 pmoles of gel-purified double-stranded PCR product in 0.2 M NaOH/0.2 mM EDTA (5 min. at room temperature), neutralized by adding 0.1 volume of 2 M ammonium acetate (pH 5.4) and precipitated with 2.5 volumes of ethanol at -70°C for 10 min. After washing with 70% ethanol, the DNA pellet was dried and redissolved in a sequencing

reaction buffer containing 4 pmoles of the oligonucleotide primer 10i-3 of Figure 18, dithiothreitol (8.3 mM) and [α-35S]-dATP (0.8 μM, 1000 Ci/mmole). The mixture was incubated at 37°C for 20 min., following which 2 μl of labelling mix and then 2 units of Sequenase were added. Aliquotes of the reaction mixture (3.5 μl) were transferred, without delay, to tubes each containing 2.5 μl of ddGTP, ddATP, ddTTP and ddCTP solutions (U.S. Biochemicals Sequenase kit) and the reactions were stopped by addition of the stop solution.

The DNA sequence for this mutant allele is shown in Figure 19b. The data derived from the cloned DNA and direct sequencing of the PCR products of the affected child and the father are all consistent with a 3 bp deletion when compared to the normal sequence (Figure 19c). The deletion of this 3 bp (ATC) at the I506 or I507 position results in the loss of an isoleucine residue from the putative CFTR, within the same ATP-binding domain where ΔF508 resides, but it is not evident whether this deleted amino acid corresponds to the position 506 or 507. Since the 506 and 507 positions are repeats, it is at present impossible to determine in which position the 3 bp deletion occurs. For convenience in later discussions, however, we refer to this deletion as ΔI507.

The fact that the ΔI507 and ΔI508 mutations occure in the same region of the presumptive ATP-binding domain of CFTR is surprising. Although the entire sequence of ΔI507 allele has not been examined, as has been done for ΔF508, the strategic location of the deletion argues that it is the responsible mutation for this allele. This argument is further supported by the observation that this alteration was not detected in any of the normal chromosomes studied to date (Kerem et al, supra). The identification of a second single amino acid

deletion in the ATP-binding domain of CFTR also provides information about the structure and function of this protein. Since deletion of either the phenylalanine residue at position 508 or isoleucine at position ΔI507 is sufficient to affect the function of CFTR, it is suggested that these residues are involved in the folding of the protein but not directly in the binding of ATP. That is, the length of the peptide is probably more important than the actual amino acid residues in this region. In support of this hypothesis, it has been found that the phenylalanine residue can be replaced by a serine and that isoleucine at position 506 with valine, without apparent loss of function of CFTR.

When the nucleotide sequence of ΔI507 is compared to that of ΔF508 at the ASO-hybridizing region, it was noted that the difference between the two alleles was only an A → T change (Figure 19c). This subtle difference thus explained the cross-hybridization of the ΔF508-ASO to ΔI507. These results therefore exemplified the importance of careful examination of both parental chromosomes in performing ASO-based genetic diagnosis. It has been determined that the ΔF508 and ΔI507 can be distinguished by increasing the stringency of oligonucleotide hybridization condition or by detecting the unique mobility of the heteroduplexes formed between each of these sequences and the normal DNA on a polyacrylaminde gel. The stringency of hybridization can be increased by using a washing temperature at 45°C instead of the prior more mild 39°C in the presence of 2XSSC (150 mM NaCL/15 mM Na citrate).

Identification of the ΔI507 and ΔI508 alleles by polyacrylamide gel electrophoresis is shown in Figure 20. The PCR products were prepared from the three family members and separated on a 5% polyacrylamide gel as described above. A DNA sample from a known

heterozygous ΔI508 carrier is included for comparison. With reference to Figure 20, the banding pattern of the PCR-amplified genomic DNA from the father, who is the carrier of ΔI507, is clearly distinguishable from that of the mother, who is of the type of carriers with the ΔF508 mutation. In this gel electrophoresis test, there were actually three individuals (the carrier father and the two affected sons in Family 5) who carried the ΔI507 deletion. Since they all belong to the same family, they only represent one single CF chromosome in our population analysis [Kerem et al, supra] The two patients who also inherited the ΔF508 mutation from their mother showed typical symptoms of CF with pancreatic insufficiency. The father of this family was the only parent who carries this ΔI507 mutation; no other CF parents showed reduced hybridization intensity signal with the ΔF508 mutant oligonucleotide probe or a peculiar heteroduplex pattern for the PCR product (as defined above) in the retrospective study. In addition, two representatives of the group IIIb and one of the group IIIc CF chromosomes from our collection [Kerem et al, supra] were sequenced, but none were found to contain ΔI507. Since the electrophoresis technique eliminates the need for probe-labelling and hybridization, it may prove to be the method of choice for detecting carriers in a large population scale.

The present data also indicate that there is a strict correlation between DNA market haplotype and mutation in CF. The ΔF508 deletion is the most common CF mutation that occurred on a group Ia chromosome background [Kerem et al, supra]. The ΔI507 mutation is, however, rare in the CF population; the one group IIIa CF chromosome carrying this deletion is the only example in our studied population (1/219) [J. M. Rommens et al, Am. J. Hum. Genet. in press (1990)]. Since the group III haplotype is relatively common among the normal

chromosomes (17/198), the ΔI507 deletion probably occurred recently. Additional studies with larger populations of different geographic and ethnic backgrounds should provide further insight in understanding the origins of these mutations.

4.0 CFTR PROTEIN

As discussed with respect to the DNA sequence of Figure 1, analysis of the sequence of the overlapping cDNA clones predicted an unprocessed polypeptide of 1480 amino acids with a molecular mass of 168,138 daltons. As later described, due to polymorphisms in the protein, the molecular weight of the protein can vary due to possible substitutions or deletion of certain amino acids. The molecular weight will also change due to the addition of carbohydrate units to form a glycoprotein. It is also understood that the functional protein in the cell will be similar to the unprocessed polypeptide, but may be modified due to cell metabolism.

Accordingly, purified normal CFTR polypeptide is characterized by a molecular weight of about 170,000 daltons and having epithelial cell transmembrane ion conductance activity. The normal CFTR polypeptide, which is substantially free of other human proteins, is encoded by the aforementioned DNA sequences and according to one embodiment, that of Figure 1. Such polypeptide displays the immunological or biological activity of normal CFTR polypeptide. As will be later discussed, the CFTR polypeptide and fragments thereof may be made by chemical or enzymatic peptide synthesis or expressed in an appropriate cultured cell system. The invention provides purified 507 mutant CFTR polypeptide which is characterized by cystic fibrosis-associated activity in human epithelial cells. Such 507 mutant CFTR polypeptide, as substantially free of other human proteins, can be encoded by the 507 mutant DNA sequence.

4.1 STRUCTURE OF CFTR

The most characteristic feature of the predicted protein is the presence of two repeated motifs, each of which consists of a set of amino acid residues capable of spanning the membrane several times followed by sequence resembling consensus nucleotide (ATP)-binding folds (NBFs) (Figures 11, 12 and 15). These characteristics are remarkably similar to those of the mammalian multidrug resistant P-glycoprotein and a number of other membrane-associated proteins, thus implying that the predicted CF gene product is likely to be involved in the transport of substances (ions) across the membrane and is probably a member of a membrane protein super family.

Figure 13 is a schematic model of the predicted CFTR protein. In Figure 13, cylinders indicate membrane spanning helices, hatched spheres indicate NBFs. The stippled sphere is the polar R-domain. The 6 membrane spanning helices in each half of the molecule are depicted as cylinders. The inner cytoplasmically oriented NBFs are shown as hatched spheres with slots to indicate the means of entry by the nucleotide. The large polar R-domain which links the two halves is represented by an stippled sphere. Charged individual amino acids within the transmembrane segments and on the R-domain surface are depicted as small circles containing the charge sign. Net charges on the internal and external loops joining the membrane cylinders and on regions of the NBFs are contained in open squares. Sites for phosphorylation by protein kinases A or C are shown by closed and open triangles respectively. K,R,H,D, and E are standard nomenclature for the amino acids, lysine, arginine, histidine, aspartic acid and glutamic acid respectively.

Each of the predicted membrane-associated regions of the CFTR protein consists of 6 highly hydrophobic

segments capable of spanning a lipid bilayer according to the algorithms of Kyte and Doolittle and of Garnier et al (J. Mol. Biol. 120, 97 (1978) (Figure 13). The membrane-associated regions are each followed by a large hydrophilic region containing the NBFs. Based on sequence alignment with other known nucleotide binding proteins, each of the putative NBFs in CFTR comprises at least 150 residues (Figure 13). The 3 bp deletion at position 507 as detected in CF patients is located between the 2 most highly conserved segments of the first NBF in CFTR. The amino acid sequence identity between the region surrounding the isoleucine deletion and the corresponding regions of a number of other proteins suggests that this region is of functional importance (Figure 15). A hydrophobic amino acid, usually one with an aromatic side chain, is present in most of these proteins at the position corresponding to I507 of the CFTR protein. It is understood that amino acid polymorphisms may exist as a result of DNA polymorphisms.

Figure 15 shows alignment of the 3 most conserved segments of the extended NBF's of CFTR with comparable regions of other proteins. These 3 segments consist of residues 433-473, 488-513, and 542-584 of the N-terminal half and 1219-1259, 1277-1302, and 1340-1382 of the C-terminal half of CFTR. The heavy overlining points out the regions of greatest similarity. Additional general homology can be seen even without the introduction of gaps.

Despite the overall symmetry in the structure of the protein and the sequence conservation of the NBFs, sequence homology between the two halves of the predicted CFTR protein is modest. This is demonstrated in Figure 12, where amino acids 1-1480 are represented on each axis. Lines on either side of the identity diagonal indicate the positions of internal

similarities. Therefore, while four sets of internal sequence identity can be detected as shown in Figure 12, using the Dayhoff scoring matrix as applied by Lawrence et al. [C. B. Lawrence, D. A. Goldman, and R. T. Hood, Bull Math Biol. 48, 569 (1986)], three of these are only apparent at low threshold settings for standard deviation. The strongest identity is between sequences at the carboxyl ends of the NBFs. Of the 66 residues aligned 27% are identical and another 11% are functionally similar. The overall weak internal homology is in contrast to the much higher degree (>70%) in P-glycoprotein for which a gene duplication hypothesis has been proposed (Gros et al, Cell 47, 371, 1986, C. Chen et al, Cell 47, 381, 1986, Gerlach et al, Nature, 324, 485, 1986, Gros et al, Mol. Cell. Biol. 8, 2770, 1988). The lack of conservation in the relative positions of the exon-intron boundaries may argue against such a model for CFTR (Figure 2).

Since there is apparently no signal-peptide sequence at the amino-terminus of CFTR, the highly charged hydrophilic segment preceding the first transmembrane sequence is probably oriented in the cytoplasm. Each of the 2 sets of hydrophobic helices are expected to form 3 transversing loops across the membrane and little sequence of the entire protein is expected to be exposed to the exterior surface, except the region between transmembrane segment 7 and 8. It is of interest to note that the latter region contains two potential sites for N-linked glycosylation.

Each of the membrane-associated regions is followed by a NBF as indicated above. In addition, a highly charged cytoplasmic domain can be identified in the middle of the predicted CFTR polypeptide, linking the 2 halves of the protein. This domain, named the R-domain, is operationally defined by a single large exon in which 69 of the 241 amino acids are polar residues arranged in

alternating clusters of positive and negative charges. Moreover, 9 of the 10 consensus sequences required for phosphosphorylation by protein kinase A (PKA), and, 7 of the potential substrate sites for protein kinase C (PKC) found in CFTR are located in this exon.

4.2 FUNCTION OF CFTR

Properties of CFTR can be derived from comparison to other membrane-associated proteins (Figure 15). In addition to the overall structural similarity with the mammalian P-glycoprotein, each of the two predicted domains in CFTR also shows remarkable resemblance to the single domain structure of hemolysin B of *E. coli* and the product of the White gene of Drosophila. These latter proteins are involved in the transport of the lytic peptide of the hemolysin system and of eye pigment molecules, respectively. The vitamin B12 transport system of *E. coli*, BtuD and MbpX which is a liverwort chloroplast gene whose function is unknown also have a similar structural motif. Furthermore, the CFTR protein shares structural similarity with several of the periplasmic solute transport systems of gram negative bacteria where the transmembrane region and the ATP-binding folds are contained in separate proteins which function in concert with a third substrate-binding polypeptide.

The overall structural arrangement of the transmembrane domains in CFTR is similar to several cation channel proteins and some cation-translocating ATPases as well as the recently described adenylate cyclase of bovine brain. The functional significance of this topological classification, consisting of 6 transmembrane domains, remains speculative.

Short regions of sequence identity have also been detected between the putative transmembrane regions of CFTR and other membrane-spanning proteins. Interestingly, there are also sequences, 18 amino acids

in length situated approximately 50 residues from the carboxyl terminus of CFTR and the raf serine/threonine kinase protooncogene of Xenopus laevis which are identical at 12 of these positions.

Finally, an amino acid sequence identity (10/13 conserved residues) has been noted between a hydrophilic segment (position 701-713) within the highly charged R-domain of CFTR and a region immediately preceding the first transmembrane loop of the sodium channels in both rat brain and eel. The charged R-domain of CFTR is not shared with the topologically closely related P-glycoprotein; the 241 amino acid linking-peptide is apparently the major difference between the two proteins.

In summary, features of the primary structure of the CFTR protein indicate its possession of properties suitable to participation in the regulation and control of ion transport in the epithelial cells of tissues affected in CF. Secure attachment to the membrane in two regions serve to position its three major intracellular domains (nucleotide-binding folds 1 and 2 and the R-domain) near the cytoplasmic surface of the cell membrane where they can modulate ion movement through channels formed either by CFTR transmembrane segments themselves or by other membrane proteins.

In view of the genetic data, the tissue-specificity, and the predicted properties of the CFTR protein, it is reasonable to conclude that CFTR is directly responsible for CF. It, however, remains unclear how CFTR is involved in the regulation of ion conductance across the apical membrane of epithelial cells.

It is possible that CFTR serves as an ion channel itself. As depicted in Figure 13, 10 of the 12 transmembrane regions contain one or more amino acids with charged side chains, a property similar to the

brain sodium channel and the GABA receptor chloride channel subunits, where charged residues are present in 4 of the 6, and 3 of the 4, respective membrane-associated domains per subunit or repeat unit. The amphipathic nature of these transmembrane segments is believed to contribute to the channel-forming capacity of these molecules. Alternatively, CFTR may not be an ion channel but instead serve to regulate ion channel activities. In support of the latter assumption, none of the purified polypeptides from trachea and kidney that are capable of reconstituting chloride channels in lipid membranes [Landry et al, Science 224:1469 (1989)] appear to be CFTR if judged on the basis of the molecular mass.

In either case, the presence of ATP-binding domains in CFTR suggests that ATP hydrolysis is directly involved and required for the transport function. The high density of phosphorylation sites for PKA and PKC and the clusters of charged residues in the R-domain may both serve to regulate this activity. The deletion of a phenylalanine residue in the NBF may prevent proper binding of ATP or the conformational change which this normally elicits and consequently result in the observed insensitivity to activation by PKA- or PKC-mediated phosphorylation of the CF apical chloride conductance pathway. Since the predicted protein contains several domains and belongs to a family of proteins which frequently function as parts of multi-component molecular systems, CFTR may also participate in epithelial tissue functions of activity or regulation not related to ion transport.

With the isolated CF gene (cDNA) now in hand it is possible to define the basic biochemical defect in CF and to further elucidate the control of ion transport pathways in epithelial cells in general. Most important, knowledge gained thus far from the predicted

structure of CFTR together with the additional information from studies of the protein itself provide a basis for the development of improved means of treatment of the disease. In such studies, antibodies have been raised to the CFTR protein as later described.

5.0 CF SCREENING

5.1 DNA BASED DIAGNOSIS

Given the knowledge of the 507 mutation as disclosed herein, carrier screening and prenatal diagnosis can be carried out as follows.

The high risk population for cystic fibrosis is Caucasians. For example, each Caucasian woman and/or man of child-bearing age would be screened to determine if she or he was a carrier (approximately a 5% probability for each individual). If both are carriers, they are a couple at risk for a cystic fibrosis child. Each child of the at risk couple has a 25% chance of being affected with cystic fibrosis. The procedure for determining carrier status using the probes disclosed herein is as follows.

One major application of the DNA sequence information of the normal and 507 mutant CF gene is in the area of genetic testing, carrier detection and prenatal diagnosis. Individuals carrying mutations in the CF gene (disease carrier or patients) may be detected at the DNA level with the use of a variety of techniques. The genomic DNA used for the diagnosis may be obtained from body cells, such as those present in peripheral blood, urine, saliva, tissue biopsy, surgical specimen and autopsy material. The DNA may be used directly for detection of specific sequence or may be amplified enzymatically in vitro by using PCR [Saiki et al. Science 230: 1350-1353, (1985), Saiki et al. Nature 324: 163-166 (1986)] prior to analysis. RNA or its cDNA form may also be used for the same purpose. Recent reviews of this subject have been presented by Caskey,

[Science 236: 1223-8 (1989) and by Landegren et al (Science 242: 229-237 (1989)].

The detection of specific DNA sequence may be achieved by methods such as hybridization using specific oligonucleotides [Wallace et al. Cold Spring Harbour Symp. Quant. Biol. 51: 257-261 (1986)], direct DNA sequencing [Church and Gilbert, Proc. Nat. Acad. Sci. U. S. A. 81: 1991-1995 (1988)], the use of restriction enzymes [Flavell et al. Cell 15: 25 (1978), Geever et al Proc. Nat. Acad. Sci. U. S. A. 78: 5081 (1981)], discrimination on the basis of electrophoretic mobility in gels with denaturing reagent (Myers and Maniatis, Cold Spring Harbour Sym. Quant. Biol. 51: 275-284 (1986)), RNase protection (Myers, R. M., Larin, J., and T. Maniatis Science 230: 1242 (1985)), chemical cleavage (Cotton et al Proc. Nat. Acad. Sci. U. S. A. 85: 4397-4401, (1985)) and the ligase-mediated detection procedure [Landegren et al Science 241:1077 (1988)].

Oligonucleotides specific to normal or mutant sequences are chemically synthesized using commercially available machines, labelled radioactively with isotopes (such as ^{32}P) or non-radioactively (with tags such as biotin (Ward and Langer et al. Proc. Nat. Acad. Sci. U. S. A. 78: 6633-6657 (1981)), and hybridized to individual DNA samples immobilized on membranes or other solid supports by dot-blot or transfer from gels after electrophoresis. The presence or absence of these specific sequences are visualized by methods such as autoradiography or fluorometric (Landegren et al, 1989, supra) or colorimetric reactions (Gebeyehu et a. Nucleic Acids Research 15: 4513-4534 (1987)). An embodiment of this oligonucleotide screening method has been applied in the detection of the I507 deletion as described herein.

Sequence differences between normal and mutants may be revealed by the direct DNA sequencing method of

Church and Gilbert (supra). Cloned DNA segments may be used as probes to detect specific DNA segments. The sensitivity of this method is greatly enhanced when combined with PCR [Wrichnik et al, Nucleic Acids Res. 15:529-542 (1987); Wong et al, Nature 330:384-386 (1987); Stoflet et al, Science 239:491-494 (1988)]. In the latter procedure, a sequencing primer which lies within the amplified sequence is used with double-stranded PCR product or single-stranded template generated by a modified PCR. The sequence determination is performed by conventional procedures with radiolabeled nucleotides or by automatic sequencing procedures with fluorescent-tags.

Sequence alterations may occasionally generate fortuitous restriction enzyme recognition sites which are revealed by the use of appropriate enzyme digestion followed by conventional gel-blot hybridization (Southern, J. Mol. Biol 98: 503 (1975)). DNA fragments carrying the site (either normal or mutant) are detected by their reduction in size or increase of corresponding restriction fragment numbers. Genomic DNA samples may also be amplified by PCR prior to treatment with the appropriate restriction enzyme; fragments of different sizes are then visualized under UV light in the presence of ethidium bromide after gel electrophoresis.

Genetic testing based on DNA sequence differences may be achieved by detection of alteration in electrophoretic mobility of DNA fragments in gels with or without denaturing reagent. Small sequence deletions and insertions can be visualized by high resolution gel electrophoresis. For example, the PCR product with the 3 bp deletion is clearly distinguishable from the normal sequence on an 8% non-denaturing polyacrylamide gel. DNA fragments of different sequence compositions may be distinguished on denaturing formamide gradient gel in which the mobilities of different DNA fragments are

retarded in the gel at different positions according to their specific "partial-melting" temperatures (Myers, supra). In addition, sequence alterations, in particular small deletions, may be detected as changes in the migration pattern of DNA heteroduplexes in non-denaturing gel electrophoresis, as have been detected for the 3 bp (I507) mutation and in other experimental systems [Nagamine et al, Am. J. Hum. Genet, 45:337-339 (1989)]. Alternatively, a method of detecting a mutation comprising a single base substitution or other small change could be based on differential primer length in a PCR. For example, one invariant primer could be used in addition to a primer specific for a mutation. The PCR products of the normal and mutant genes can then be differentially detected in acrylamide gels.

Sequence changes at specific locations may also be revealed by nuclease protection assays, such as RNase (Myers, supra) and S1 protection (Berk, A. J., and P. A. Sharpe Proc. Nat. Acad. Sci. U. S. A. 75: 1274 (1978)), the chemical cleavage method (Cotton, supra) or the ligase-mediated detection procedure (Landegren supra).

In addition to conventional gel-electrophoresis and blot-hybridization methods, DNA fragments may also be visualized by methods where the individual DNA samples are not immobilized on membranes. The probe and target sequences may be both in solution or the probe sequence may be immobilized [Saiki et al, Proc. Natl. Acad. Sci USA, 86:6230-6234 (1989)]. A variety of detection methods, such as autoradiography involving radioisotopes, direct detection of radioactive decay (in the presence or absence of scintillant), spectrophotometry involving colorigenic reactions and fluorometry involving fluorogenic reactions, may be used to identify specific individual genotypes.

Since more than one mutation is anticipated in the CF gene such as I507 and F508, a multiples system is an ideal protocol for screening CF carriers and detection of specific mutations. For example, a PCR with multiple, specific oligonucleotide primers and hybridization probes, may be used to identify all possible mutations at the same time (Chamberlain et al. Nucleic Acids Research 16: 1141-1155 (1988)). The procedure may involve immobilized sequence-specific oligonucleotides probes (Saiki et al, supra).

5.2 DETECTING THE CF 507 MUTATION

These detection methods may be applied to prenatal diagnosis using amniotic fluid cells, chorionic villi biopsy or sorting fetal cells from maternal circulation. The test for CF carriers in the population may be incorporated as an essential component in a broad-scale genetic testing program for common diseases.

According to an embodiment of the invention, the portion of the DNA segment that is informative for a mutation, such as the mutation according to this embodiment, that is, the portion that immediately surrounds the I507 deletion, can then be amplified by using standard PCR techniques [as reviewed in Landegren, Ulf, Robert Kaiser, C. Thomas Caskey, and Leroy Hood, DNA Diagnostics - Molecular Techniques and Automation, in Science 242: 229-237 (1988)]. It is contemplated that the portion of the DNA segment which is used may be a single DNA segment or a mixture of different DNA segments. A detailed description of this technique now follows.

A specific region of genomic DNA from the person or fetus is to be screened. Such specific region is defined by the oligonucleotide primers C16B (5'GTTTTCCTGGATTATGCCTGGCAC3') and C16D (5'GTTGGCATGCTTTGATGACGCTTC3') or as shown in Figure 18 by primers 10i-5 and 10i-3. The specific regions using

10i-5 and 10i-3 were amplified by the polymerase chain reaction (PCR). 200-400 ng of genomic DNA, from either cultured lymphoblasts or peripheral blood samples of CF individuals and their parents, were used in each PCR with the oligonucleotides primers indicated above. The oligonucleotides were purified with Oligonucleotide Purification Cartridges™ (Applied Biosystems) or NENSORB™ PREP columns (Dupont) with procedures recommended by the suppliers. The primers were annealed at 55°C for 30 sec, extended at 72°C for 60 sec (with 2 units of Taq DNA polymerase) and denatured at 94°C for 60 sec, for 30 cycles with a final cycle of 7 min for extension in a Perkin-Elmer/Cetus automatic thermocycler with a Step-Cycle program (transition setting at 1.5 min). Portions of the PCR products were separated by electrophoresis on 1.4% agarose gels, transferred to Zetabind™; (Biorad) membrane according to standard procedures.

The normal and ΔI507 oligonucleotide probes of Figure 19 (10 ng each) are labeled separately with 10 units of T4 polynucleotide kinase (Pharmacia) in a 10 μl reaction containing 50 mM Tris-HCl (pH7.6), 10 mM $MgCl_2$, 0.5 mM dithiothreitol, 10 mM spermidine, 1 mM EDTA and 30-40 μCi of $\gamma[^{32}P]$ - ATP for 20-30 min at 37°C. The unincorporated radionucleotides were removed with a Sephadex G-25 column before use. The hybridization conditions were as described previously (J.M. Rommens et al Am. J. Hum. Genet. 43,645 (1988)) except that the temperature can be 37°C. The membranes are washed twice at room temperature with 5xSSC and twice at 39°C with 2 x SSC (1 x SSC = 150 mM NaCl and 15 mM Na citrate). Autoradiography is performed at room temperature overnight. Autoradiographs are developed to show the hybridization results of genomic DNA with the 2 specific oligonucleotide probes. Probe C normal detects

the normal DNA sequence and Probe C ΔI507 detects the mutant sequence.

Genomic DNA sample from each family member can, as explained, be amplified by the polymerase chain reaction using the intron sequences of Figure 18 and the products separated by electrophoresis on a 1.4% agarose gel and then transferred to Zetabind (Biorad) membrane according to standard procedures. The 3bp deletion of ΔI507 can be revealed by a very convenient polyacrylamide gel electrophoresis procedure. When the PCR products generated by the above-mentioned 10i-5 and 10i-3 primers are applied to an 5% polyacrylamide gel, electrophoresed for 3 hrs at 20V/cm in a 90mM Tris-borate buffer (pH 8.3), DNA fragments of a different mobility are clearly detectable for individuals without the 3 bp deletion, heterozygous or homozygous for the deletion.

As already explained with respect to Figure 20, the PCR amplified genomic DNA can be subjected to gel electrophoresis to identify the 3 bp deletion. As shown in Figure 20, in the four lanes the first lane is a control with a normal/ΔF508 deletion. The next lane is the father with a normal/ΔI507 deletion. The third lane is the mother with a normal/ΔF508 deletion and the fourth lane is the child with a ΔF508/ΔI507 deletion. The homoduplexes show up as solid bands across the base of each lane. In lanes 1 and 3, the two heteroduplexes show up very clealy as two spaced apart bands. In lane 2, the father's ΔI507 mutation shows up very clearly, whereas in the fourth lane, the child with the adjacent 507, 508 mutations, there is no distinguishable heteroduplexes. Hence the showing is at the homoduplex line. Since the father in lane 2 and the mother in lane 3 show heteroduplex banding and the child does not, indicates either the child is normal or is a patient. This can be futher checked if needed, such as in

embryoic analysis by mixing the 507 and 508 probes to determine the presence of the ΔI507 and ΔF508 mutations.

Similar alteration in gel mobility for heteroduplexes formed during PCR has also been reported for experimental systems where small deletions are involved (Nagamine et al supra). These mobility shifts may be used in general as the basis for the non-radioactive genetic screening tests.

5.3 CF SCREENING PROGRAMS

It is appreciated that approximately 1% of the carriers can be detected using the specific ΔI507 probes of this particular embodiment of the invention. Thus, if an individual tested is not a carrier using the ΔI507 probes, their carrier status can not be excluded, they may carry some other mutation, such as the ΔF508 as previously noted. However, if both the individual and the spouse of the individual tested are a carrier for the ΔI507 mutation, it can be stated with certainty that they are an at risk couple. The sequence of the gene as disclosed herein is an essential prerequisite for the determination of the other mutations.

Prenatal diagnosis is a logical extension of carrier screening. A couple can be identified as at risk for having a cystic fibrosis child in one of two ways: if they already have a cystic fibrosis child, they are both, by definition, obligate carriers of the defective CFTR gene, and each subsequent child has a 25% chance of being affected with cystic fibrosis. A major advantage of the present invention eliminates the need for family pedigree analysis, whereas, according to this invention, a gene mutation screening program as outlined above or other similar method can be used to identify a genetic mutation that leads to a protein with altered function. This is not dependent on prior ascertainment of the family through an affected child. Fetal DNA

samples, for example, can be obtained, as previously mentioned, from amniotic fluid cells and chorionic villi specimens. Amplification by standard PCR techniques can then be performed on this template DNA.

If both parents are shown to be carriers with the ΔI507 deletion, the interpretation of the results would be the following. If there is hybridization of the fetal DNA to the normal probe, the fetus will not be affected with cystic fibrosis, although it may be a CF carrier (50% probability for each fetus of an at risk couple). If the fetal DNA hybridizes only to the ΔI507 deletion probe and not to the normal probe, the fetus will be affected with cystic fibrosis.

It is appreciated that for this and other mutations in the CF gene, a range of different specific procedures can be used to provide a complete diagnosis for all potential CF carriers or patients. A complete description of these procedures is later described.

The invention therefore provides a method and kit for determining if a subject is a CF carrier or CF patient. In summary, the screening method comprises the steps of:

providing a biological sample of the subject to be screened; and providing an assay for detecting in the biological sample, the presence of at least a member from the group consisting of a 507 mutant CF gene, 507 mutant CF gene products and mixtures thereof.

The method may be further characterized by including at least one more nucleotide probe which is a different DNA sequence fragment of, for example, the DNA of Figure 1, or a different DNA sequence fragment of human chromosome 7 and located to either side of the DNA sequence of Figure 1. In this respect, the DNA fragments of the intron portions of Figure 2 are useful in further confirming the presence of the mutation. Unique aspects of the introns at the exon boundaries may

be relied upon in screening procedures to further confirm the presence of the mutation at the I507 position.

A kit, according to an embodiment of the invention, suitable for use in the screening technique and for assaying for the presence of the 507 mutant CF gene by an immunoassay comprises:

(a) an antibody which specifically binds to a gene product of the 507 mutant CF gene;

(b) reagent means for detecting the binding of the antibody to the gene product; and

(c) the antibody and reagent means each being present in amounts effective to perform the immunoassay.

The kit for assaying for the presence for the 507 mutant CF gene may also be provided by hybridization techniques. The kit comprises:

(a) an oligonucleotide probe which specifically binds to the 507 mutant CF gene;

(b) reagent means for detecting the hybridization of the oligonucleotide probe to the 507 mutant CF gene; and

(c) the probe and reagent means each being present in amounts effective to perform the hybridization assay.

5.4 ANTIBODIES TO DETECT MUTANT CFTR

As mentioned, antibodies to epitopes within the 507 mutant CFTR protein are raised to provide extensive information on the characteristics of the mutant protein and other valuable information which includes:

1. The antibodies can be used to provide another technique in detecting any of the other CF mutations which result in the synthesis of a protein with an altered size.

2. Antibodies to distinct domains of the mutant protein can be used to determine the topological arrangement of the protein in the cell membrane. This provides information on segments of the

protein which are accessible to externally added modulating agents for purposes of drug therapy.

3. The structure-function relationships of portions of the protein can be examined using specific antibodies. For example, it is possible to introduce into cells antibodies recognizing each of the charged cytoplasmic loops which join the transmembrane sequences as well as portions of the nucleotide binding folds and the R-domain. The influence of these antibodies on functional parameters of the protein provide insight into cell regulatory mechanisms and potentially suggest means of modulating the activity of the defective protein in a CF patient.

4. Antibodies with the appropriate avidity also enable immunoprecipitation and immuno-affinity purification of the protein. Immunoprecipitation will facilitate characterization of synthesis and post translational modification including ATP binding and phosphorylation. Purification will be required for studies of protein structure and for reconstitution of its function, as well as protein based therapy.

In order to prepare the antibodies, fusion proteins containing defined portions of 507 mutant CFTR polypeptides can be synthesized in bacteria by expression of corresponding 507 mutant DNA sequence in a suitable cloning vehicle. Smaller peptide may be synthesized chemically. The fusion proteins can be purified, for example, by affinity chromatography on glutathione-agarose and the peptides coupled to a carrier protein (hemocyanin), mixed with Freund's adjuvant and injected into rabbits. Following booster injections at bi-weekly intervals, the rabbits are bled and sera isolated. The developed polyclonal antibodies in the sera may then be combined with the fusion

proteins. Immunoblots are then formed by staining with, for example, alkaline-phosphatase conjugated second antibody in accordance with the procedure of Blake et al, Anal. Biochem. 136:175 (1984).

Thus, it is possible to raise polyclonal antibodies specific for both fusion proteins containing portions of the 507 mutant CFTR protein and peptides corresponding to short segments of its sequence. Similarly, mice can be injected with KLH conjugates of peptides to initiate the production of monoclonal antibodies to corresponding segments of 507 mutant CFTR protein.

As for the generation of monoclonal antibodies, immunogens for the raising of monoclonal antibodies (mAbs) to the 507 mutant CFTR protein are bacterial fusion proteins [Smith et al, Gene 67:31 (1988)] containing portions of the CFTR polypeptide or synthetic peptides corresponding to short (12 to 25 amino acids in length) segments of the mutant sequence. The essential methodology is that of Kohler and Milstein [Nature 256: 495 (1975)].

Balb/c mice are immunized by intraperitoneal injection with 500 μg of pure fusion protein or synthetic peptide in incomplete Freund's adjuvant. A second injection is given after 14 days, a third after 21 days and a fourth after 28 days. Individual animals so immunized are sacrificed one, two and four weeks following the final injection. Spleens are removed, their cells dissociated, collected and fused with Sp2/O-Ag14 myeloma cells according to Gefter et al, Somatic Cell Genetics 3:231 (1977). The fusion mixture is distributed in culture medium selective for the propagation of fused cells which are grown until they are about 25% confluent. At this time, culture supernatants are tested for the presence of antibodies reacting with a particular CFTR antigen. An alkaline phosphatase labelled anti-mouse second antibody is then

used for detection of positives. Cells from positive culture wells are then expanded in culture, their supernatants collected for further testing and the cells stored deep frozen in cryoprotectant-containing medium. To obtain large quantities of a mAb, producer cells are injected into the peritoneum at 5 x 10^6 cells per animal, and ascites fluid is obtained. Purification is by chromotography on Protein G- or Protein A-agarose according to Ey et al, Immunochemistry 15:429 (1977).

Reactivity of these mAbs with the 507 mutant CFTR protein can be confirmed by polyacrylamide gel electrophoresis of membranes isolated from epithelial cells in which it is expressed and immunoblotted [Towbin et al, Proc. Natl. Acad. Sci. USA 76:4350 (1979)].

In addition to the use of monoclonal antibodies specific for 507 mutant domain of the CFTR protein to probe their individual functions, other mAbs, which can distinguish between the normal and 507 mutant forms of CFTR protein, are used to detect the mutant protein in epithelial cell samples obtained from patients, such as nasal mucosa biopsy "brushings" [R. De-Lough and J. Rutland, J. Clin. Pathol. 42, 613 (1989)] or skin biopsy specimens containing sweat glands.

Antibodies capable of this distinction are obtained by differentially screening hybridomas from paired sets of mice immunized with a peptide containing the isoleucine at amino acid position 507 (e.g. GTIKENIIFGVSY) or a peptide which is identical except for the absence of I507 (GTIKENIFGVSY). mAbs capable of recognizing the other mutant forms of CFTR protein present in patients in addition or instead of I507 deletion are obtained using similar monoclonal antibody production strategies.

Antibodies to normal and CF versions of CFTR protein and of segments thereof are used in

diagnostically immunocytochemical and immunofluorescence light microscopy and immunoelectron microscopy to demonstrate the tissue, cellular and subcellular distribution of CFTR within the organs of CF patients, carriers and non-CF individuals.

Antibodies are used to therapeutically modulate by promoting the activity of the CFTR protein in CF patients and in cells of CF patients. Possible modes of such modulation might involve stimulation due to cross-linking of CFTR protein molecules with multivalent antibodies in analogy with stimulation of some cell surface membrane receptors, such as the insulin receptor [O'Brien et al, Euro. Mol. Biol. Organ. J. 6:4003 (1987)], epidermal growth factor receptor [Schreiber et al, J. Biol. Chem. 258:846 (1983)] and T-cell receptor-associated molecules such as CD4 [Veillette et al Nature, 338:257 (1989)].

Antibodies are used to direct the delivery of therapeutic agents to the cells which express defective CFTR protein in CF. For this purpose, the antibodies are incorporated into a vehicle such as a liposome [Matthay et al, Cancer Res. 46:4904 (1986)] which carries the therapeutic agent such as a drug or the normal gene.

5.5 RFLP ANALYSIS

DNA diagnosis is currently being used to assess whether a fetus will be born with cystic fibrosis, but historically this has only been done after a particular set of parents has already had one cystic fibrosis child which identifies them as obligate carriers. However, in combination with carrier detection as outlined above, DNA diagnosis for all pregnancies of carrier couples will be possible. If the parents have already had a cystic fibrosis child, an extended haplotype analysis can be done on the fetus and thus the percentage of false positive or false negative will be greatly

reduced. If the parents have not already had an affected child and the DNA diagnosis on the fetus is being performed on the basis of carrier detection, haplotype analysis can still be performed.

Although it has been thought for many years that there is a great deal of clinical heterogeneity in the cystic fibrosis disease, it is now emerging that there are two general categories, called pancreatic sufficiency (CF-PS) and pancreatic insufficiency (CF-PI). If the mutations related to these disease categories are well characterized, one can associate a particular mutation with a clinical phenotype of the disease. This allows changes in the treatment of each patient. Thus the nature of the mutation will to a certain extent predict the prognosis of the patient and indicate a specific treatment.

6.0 MOLECULAR BIOLOGY OF CYSTIC FIBROSIS

The postulate that CFTR may regulate the activity of ion channels, particularly the outwardly rectifying Cl channel implicated as the functional defect in CF, can be tested by the injection and translation of full length *in vitro* transcribed CFTR mRNA in Xenopus oocytes. The ensuing changes in ion currents across the oocyte membrane can be measured as the potential is clamped at a fixed value. CFTR may regulate endogenous oocyte channels or it may be necessary to also introduce epithelial cell RNA to direct the translation of channel proteins. Use of mRNA coding for normal and for mutant CFTR, as provided by this invention, makes these experiments possible.

Other modes of expression in heterologous cell system also facilitate dissection of structure-function relationships. The complete CFTR DNA sequence ligated into a plasmid expression vector is used to transfect cells so that its influence on ion transport can be assessed. Plasmid expression vectors containing part of

the normal CFTR sequence along with portions of modified sequence at selected sites can be used in vitro mutagenesis experiments performed in order to identify those portions of the CFTR protein which are crucial for regulatory function.

6.1 EXPRESSION OF 507 MUTANT DNA SEQUENCE

The 507 mutant DNA sequence can be manipulated in studies to understand the expression of the gene and its product, and, to achieve production of large quantities of the protein for functional analysis, antibody production, and patient therapy. The changes in the sequence may or may not alter the expression pattern in terms of relative quantities, tissue-specificity and functional properties. The partial or full-length cDNA sequences, which encode for the subject protein, unmodified or modified, may be ligated to bacterial expression vectors such as the pRIT (Nilsson et al. EMBO J. 4: 1075-1080 (1985)), pGEX (Smith and Johnson, Gene 67: 31-40 (1988)) or pATH (Spindler et al. J. Virol. 49: 132-141 (1984)) plasmids which can be introduced into E. coli cells for production of the corresponding proteins which may be isolated in accordance with the previously discussed protein purification procedures. The DNA sequence can also be transferred from its existing context to other cloning vehicles, such as other plasmids, bacteriophages, cosmids, animal virus, yeast artificial chromosomes (YAC)(Burke et al. Science 236: 806-812, (1987)), somatic cells, and other simple or complex organisms, such as bacteria, fungi (Timberlake and Marshall, Science 244: 1313-1317 (1989), invertebrates, plants (Gasser and Fraley, Science 244: 1293 (1989), and pigs (Pursel et al. Science 244: 1281-1288 (1989)).

For expression in mammalian cells, the cDNA sequence may be ligated to heterologous promoters, such as the simian virus (SV) 40, promoter in the pSV2 vector

[Mulligan and Berg, Proc. Natl. Acad. Sci USA, 78:2072-2076 (1981)] and introduced into cells, such as monkey COS-1 cells [Gluzman, Cell, 23:175-182 (1981)], to achieve transient or long-term expression. The stable integration of the chimeric gene construct may be maintained in mammalian cells by biochemical selection, such as neomycin [Southern and Berg, J. Mol. Appln. Genet. 1:327-341 (1982)] and mycophoenolic acid [Mulligan and Berg, supra].

DNA sequences can be manipulated with standard procedures such as restriction enzyme digestion, fill-in with DNA polymerase, deletion by exonuclease, extension by terminal deoxynucleotide transferase, ligation of synthetic or cloned DNA sequences, site-directed sequence-alteration via single-stranded bacteriophage intermediate or with the use of specific oligonucleotides in combination with PCR.

The cDNA sequence (or portions derived from it), or a mini gene (a cDNA with an intron and its own promoter) is introduced into eukaryotic expression vectors by conventional techniques. These vectors are designed to permit the transcription of the cDNA in eukaryotic cells by providing regulatory sequences that initiate and enhance the transcription of the cDNA and ensure its proper splicing and polyadenylation. Vectors containing the promoter and enhancer regions of the simian virus (SV)40 or long terminal repeat (LTR) of the Rous Sarcoma virus and polyadenylation and splicing signal from SV 40 are readily available [Mulligan et al Proc. Natl. Acad. Sci. USA 78:1078-2076, (1981); Gorman et al Proc Natl. Acad. Sci USA 79: 6777-6781 (1982)]. Alternatively, the CFTR endogenous promoter may be used. The level of expression of the cDNA can be manipulated with this type of vector, either by using promoters that have different activities (for example, the baculovirus pAC373 can express cDNAs at high levels in S.

frungiperda cells [M. D. Summers and G. E. Smith in, Genetically Altered Viruses and the Environment (B. Fields, et al, eds.) vol. 22 no 319-328, Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York, 1985] or by using vectors that contain promoters amenable to modulation, for example the glucocorticoid-responsive promoter from the mouse mammary tumor virus [Lee et al, Nature 294:228 (1982)]. The expression of the cDNA can be monitored in the recipient cells 24 to 72 hours after introduction (transient expression).

In addition, some vectors contain selectable markers [such as the gpt [Mulligan et Berg supra] or neo [Southern and Berg J. Mol. Appln. Genet 1:327-341 (1982)] bacterial genes that permit isolation of cells, by chemical selection, that have stable, long term expression of the vectors (and therefore the cDNA) in the recipient cell. The vectors can be maintained in the cells as episomal, freely replicating entities by using regulatory elements of viruses such as papilloma [Sarver et al Mol. Cell Biol. 1:486 (1981)] or Epstein-Barr (Sugden et al Mol. Cell Biol. 5:410 (1985)]. Alternatively, one can also produce cell lines that have integrated the vector into genomic DNA. Both of these types of cell lines produce the gene product on a continuous basis. One can also produce cell lines that have amplified the number of copies of the vector (and therefore of the cDNA as well) to create cell lines that can produce high levels of the gene product [Alt et al. J. Biol. Chem. 253: 1357 (1978)].

The transfer of DNA into eukaryotic, in particular human or other mammalian cells is now a conventional technique. The vectors are introduced into the recipient cells as pure DNA (transfection) by, for example, precipitation with calcium phosphate [Graham and vander Eb, Virology 52:466 (1973) or strontium phosphate [Brash et al Mol. Cell Biol. 7:2013 (1987)],

electroporation [Neumann et al EMBO J 1:841 (1982)], lipofection [Felgner et al Proc Natl. Acad. Sci USA 84:7413 (1987)], DEAE dextran [McCuthan et al J. Natl Cancer Inst. 41:351 1968)], microinjection [Mueller et al Cell 15:579 1978)], protoplast fusion [Schafner, Proc Natl. Aca. Sci USA 72:2163] or pellet guns [Klein et al, Nature 327: 70 (1987)]. Alternatively, the cDNA can be introduced by infection with virus vectors. Systems are developed that use, for example, retroviruses [Bernstein et al. Genetic Engineering 7: 235, (1985)], adenoviruses [Ahmad et al J. Virol 57:267 (1986)] or Herpes virus [Spaete et al Cell 30:295 (1982)].

These eukaryotic expression systems can be used for many studies of the 507 mutant CF gene and the 507 mutant CFTR product. These include, for example: (1) determination that the gene is properly expressed and that all post-translational modifications necessary for full biological activity have been properly completed (2) identify regulatory elements located in the 5' region of the CF gene and their role in the tissue- or temporal-regulation of the expression of the CF gene (3) production of large amounts of the normal protein for isolation and purification (4) to use cells expressing the CFTR protein as an assay system for antibodies generated against the CFTR protein or an assay system to test the effectiveness of drugs, (5) study the function of the normal complete protein, specific portions of the protein, or of naturally occurring or artificially produced mutant proteins. Naturally occurring mutant proteins exist in patients with CF while artificially produced mutant protein can be designed by site directed sequence alterations. These latter studies can probe the function of any desired amino acid residue in the protein by mutating the nucleotides coding for that amino acid.

Using the above techniques, the expression vectors containing the 507 mutant CF gene sequence or fragments thereof can be introduced into human cells, mammalian cells from other species or non-mammalian cells as desired. The choice of cell is determined by the purpose of the treatment. For example, one can use monkey COS cells [Gluzman, Cell 23:175 (1981)], that produce high levels of the SV40 T antigen and permit the replication of vectors containing the SV40 origin of replication, can be used to show that the vector can express the protein product, since function is not required. Similar treatment could be performed with Chinese hamster ovary (CHO) or mouse NIH 3T3 fibroblasts or with human fibroblasts or lymphoblasts.

The recombinant cloning vector, according to this invention, then comprises the selected DNA of the DNA sequences of this invention for expression in a suitable host. The DNA is operatively linked in the vector to an expression control sequence in the recombinant DNA molecule so that normal CFTR polypeptide can be expressed. The expression control sequence may be selected from the group consisting of sequences that control the expression of genes of prokaryotic or eukaryotic cells and their viruses and combinations thereof. The expression control sequence may be specifically selected from the group consisting of the lac system, the trp system, the tac system, the trc system, major operator and promoter regions of phage lambda, the control region of fd coat protein, the early and late promoters of SV40, promoters derived from polyoma, adenovirus, retrovirus, baculovirus and simian virus, the promoter for 3-phosphoglycerate kinase, the promoters of yeast acid phosphatase, the promoter of the yeast alpha-mating factors and combinations thereof.

The host cell, which may be transfected with the vector of this invention, may be selected from the group

consisting of E. coli, Pseudomonas, Bacillus subtilis, Bacillus stearothermophilus or other bacili; other bacteria; yeast; fungi; insect; mouse or other animal; or plant hosts; or human tissue cells.

It is appreciated that for the mutant DNA sequence similar systems are employed to express and produce the mutant product.

6.2 PROTEIN FUNCTION CONSIDERATIONS

To study the function of the mutant CFTR protein, it is preferable to use epithelial cells as recipients, since proper functional expression may require the presence of other pathways or gene products that are only expressed in such cells. Cells that can be used include, for example, human epithelial cell lines such as T84 (ATCC #CRL 248) or PANC-1 (ATCC # CLL 1469), or the T43 immortalized CF nasal epithelium cell line [Jettan et al, Science (1989)] and primary [Yanhoskes et al. Ann. Rev. Resp. Dis. 132: 1281 (1985)] or transformed [Scholte et al. Exp. Cell. Res. 182: 559(1989)] human nasal polyp or airways cells, pancreatic cells [Harris and Coleman J. Cell. Sci. 87: 695 (1987)], or sweat gland cells [Collie et al. In Vitro 21: 597 (1985)] derived from normal or CF subjects. The CF cells can be used to test for the functional activity of mutant CF genes. Current functional assays available include the study of the movement of anions (Cl or I) across cell membranes as a function of stimulation of cells by agents that raise intracellular AMP levels and activate chloride channels [Stutto et al. Proc. Nat. Acad. Sci. U. S. A. 82: 6677 (1985)]. Other assays include the measurement of changes in cellular potentials by patch clamping of whole cells or of isolated membranes [Frizzell et al. Science 233: 558 (1986), Welsch and Liedtke Nature 322: 467 (1986)]or the study of ion fluxes in epithelial sheets of confluent cells [Widdicombe et al. Proc. Nat. Acad. Sci.

82: 6167 (1985)]. Alternatively, RNA made from the CF gene could be injected into Xenopus oocytes. The oocyte will translate RNA into protein and allow its study. As other more specific assays are developed these can also be used in the study of transfected mutant CFTR protein function.

"Domain-switching" experiments between mutant CFTR and the human multidrug resistance P-glycoprotein can also be performed to further the study of the mutant CFTR protein. In these experiments, plasmid expression vectors are constructed by routine techniques from fragments of the mutant CFTR sequence and fragments of the sequence of P-glycoprotein ligated together by DNA ligase so that a protein containing the respective portions of these two proteins will be synthesized by a host cell transfected with the plasmid. The latter approach has the advantage that many experimental parameters associated with multidrug resistance can be measured. Hence, it is now possible to assess the ability of segments of mutant CFTR to influence these parameters.

These studies of the influence of mutant CFTR on ion transport will serve to bring the field of epithelial transport into the molecular arena.

6.3 THERAPIES

It is understood that the major aim of the various biochemical studies using the compositions of this invention is the development of therapies to circumvent or overcome the CF defect, using both the pharmacological and the "gene-therapy" approaches.

In the pharmacological approach, drugs which circumvent or overcome the CF defect are sought. Initially, compounds may be tested essentially at random, and screening systems are required to discriminate among many candidate compounds. This invention provides host cell systems, expressing various

of the mutant CF genes, which are particularly well suited for use as first level screening systems. Preferably, a cell culture system using mammalian cells (most preferably human cells) transfected with an expression vector comprising a DNA sequence coding for CFTR protein containing a CF-generating mutation, for example the I507 deletion, is used in the screening process. Candidate drugs are tested by incubating the cells in the presence of the candidate drug and measuring those cellular functions dependent on CFTR, especially by measuring ion currents where the transmembrane potential is clamped at a fixed value. To accommodate the large number of assays, however, more convenient assays are based, for example, on the use of ion-sensitive fluorescent dyes. To detect changes in Cl^-ion concentration SPQ or its analogues are useful.

Alternatively, a cell-free system could be used. Purified CFTR could be reconstituted into articifial membranes and drugs could be screened in a cell-free assay [Al-Aqwatt, Science, (1989)].

At the second level, animal testing is required. It is possible to develop a model of CF by interfering with the normal expression of the counterpart of the CF gene in an animal such as the mouse. The "knock-out" of this gene by introducing a mutant form of it into the germ line of animals will provide a strain of animals with CF-like syndromes. This enables testing of drugs which showed a promise in the first level cell-based screen.

As further knowledge is gained about the nature of the protein and its function, it will be possible to predict structures of proteins or other compounds that interact with the CFTR protein. That in turn will allow for certain predictions to be made about potential drugs that will interact with this protein and have some effect on the treatment of the patients. Ultimately

such drugs may be designed and synthesized chemically on the basis of structures predicted to be required to interact with domains of CFTR. This approach is reviewed in Capsey and Delvatte, Genetically Engineered Human Therapeutic Drugs Stockton Press, New York, 1988. These potential drugs must also be tested in the screening system.

6.3.1 PROTEIN REPLACEMENT THERAPY

Treatment of CF can be performed by replacing the defective protein with normal protein, by modulating the function of the defective protein or by modifying another step in the pathway in which CFTR participates in order to correct the physiological abnormality.

To be able to replace the defective protein with the normal version, one must have reasonably large amounts of pure CFTR protein. Pure protein can be obtained as described earlier from cultured cell systems. Delivery of the protein to the affected airways tissue will require its packaging in lipid-containing vesicles that facilitate the incorporation of the protein into the cell membrane. It may also be feasible to use vehicles that incorporate proteins such as surfactant protein, such as SAP(Val) or SAP(Phe) that performs this function naturally, at least for lung alveolar cells. (PCT Patent Application WO/8803170, Whitsett et al, May 7, 1988 and PCT Patent Application WO89/04327, Benson et al, May 18, 1989). The CFTR-containing vesicles are introduced into the airways by inhalation or irrigation, techniques that are currently used in CF treatment (Boat et al, supra).

6.3.2 DRUG THERAPY

Modulation of CFTR function can be accomplished by the use of therapeutic agents (drugs). These can be identified by random approaches using a screening program in which their effectiveness in modulating the defective CFTR protein is monitored in vitro. Screening

programs can use cultured cell systems in which the defective CFTR protein is expressed. Alternatively, drugs can be designed to modulate CFTR activity from knowledge of the structure and function correlations of CFTR protein and from knowledge of the specific defect in the 507 CFTR mutant protein (Capsey and Delvatte, supra). It is possible that the 507 mutant CFTR protein will require a different drug for specific modulation. It will then be necessary to identify the specific mutation(s) in each CF patient before initiating drug therapy.

Drugs can be designed to interact with different aspects of CFTR protein structure or function. For example, a drug (or antibody) can bind to a structural fold of the protein to correct a defective structure. Alternatively, a drug might bind to a specific functional residue and increase its affinity for a substrate or cofactor. Since it is known that members of the class of proteins to which CFTR has structural homology can interact, bind and transport a variety of drugs, it is reasonable to expect that drug-related therapies may be effective in treatment of CF.

A third mechanism for enhancing the activity of an effective drug would be to modulate the production or the stability of CFTR inside the cell. This increase in the amount of CFTR could compensate for its defective function.

Drug therapy can also be used to compensate for the defective CFTR function by interactions with other components of the physiological or biochemical pathway necessary for the expression of the CFTR function. These interactions can lead to increases or decreases in the activity of these ancillary proteins. The methods for the identification of these drugs would be similar to those described above for CFTR-related drugs.

In other genetic disorders, it has been possible to correct for the consequences of altered or missing normal functions by use of dietary modifications. This has taken the form of removal of metabolites, as in the case of phenylketonuria, where phenylalanine is removed from the diet in the first five years of life to prevent mental retardation, or by the addition of large amounts of metabolites to the diet, as in the case of adenosime deaminase deficiency where the functional correction of the activity of the enzyme can be produced by the addition of the enzyme to the diet. Thus, once the details of the CFTR function have been elucidated and the basic defect in CF has been defined, therapy may be achieved by dietary manipulations.

The second potential therapeutic approach is so-called "gene-therapy" in which normal copies of the CF gene are introduced in to patients so as to successfully code for normal protein in the key epithelial cells of affected tissues. It is most crucial to attempt to achieve this with the airway epithelial cells of the respiratory tract. The CF gene is delivered to these cells in form in which it can be taken up and code for sufficient protein to provide regulatory function. As a result, the patient's quality and length of life will be greatly extended. Ultimately, of course, the aim is to deliver the gene to all affected tissues.

6.3.3 GENE THERAPY

One approach to therapy of CF is to insert a normal version of the CF gene into the airway epithelium of affected patients. It is important to note that the respiratory system is the primary cause of mordibity and mortality in CF; while pancreatic disease is a major feature, it is relatively well treated today with enzyme supplementation. Thus, somatic cell gene therapy [for a review, see T. Friedmann, Science 244:1275 (1989)]

targeting the airway would alleviate the most severe problems associated with CF.

A. Retroviral Vectors. Retroviruses have been considered the preferred vector for experiments in somatic gene therapy, with a high efficiency of infection and stable integration and expression [Orkin et al Prog. Med. Genet 7:130, (1988)]. A possible drawback is that cell division is necessary for retroviral integration, so that the targeted cells in the airway may have to be nudged into the cell cycle prior to retroviral infection, perhaps by chemical means. The full length CF gene cDNA can be cloned into a retroviral vector and driven from either its endogenous promoter or from the retroviral LRT (long terminal repeat). Expression of levels of the normal protein as low as 10% of the endogenous mutant protein in CF patients would be expected to be beneficial, since this is a recessive disease. Delivery of the virus could be accomplished by aerosol or instillation into the trachea.

B. Other Viral Vectors. Other delivery systems which can be utilized include adeno-associated virus [AAV, McLaughlin et al, J. Virol 62:1963 (1988)], vaccinia virus [Moss et al Annu. Rev.Immunol, 5:305, 1987)], bovine papilloma virus [Rasmussen et al, Methods Enzymol 139:642 (1987)] or member of the herpesvirus group such as Epstein-Barr virus (Margolskee et al Mol. Cell. Biol 8:2937 (1988)]. Though much would need to be learned about their basic biology, the idea of using a viral vector with natural tropism for the respiratory track (e.g. respiratory syncytial virus, echovirus, Coxsackie virus, etc.) is possible.

C. Non-viral Gene Transfer. Other methods of inserting the CF gene into respiratory epithelium may also be productive; many of these are lower efficiency and would potentially require infection in vitro,

selection of transfectants, and reimplantation. This would include calcium phosphate, DEAE dextran, electroporation, and protoplast fusion. A particularly attractive idea is the use of liposome, which might be possible to carry out in vivo [Ostro, Liposomes, Marcel-Dekker, 1987]. Synthetic cationic lipids such as DOTMA [Felger et al Proc. Natl. Acad.Sci USA 84:7413 (1987)] may increase the efficiency and ease of carrying out this approach.

6.4 CF ANIMAL MODELS

The creation of a mouse or other animal model for CF will be crucial to understanding the disease and for testing of possible therapies (for general review of creating animal models, see Erickson, Am. J. Hum.Genet 43:582 (1988)]. Currently no animal model of the CF exists. The evolutionary conservation of the CF gene (as demonstrated by the cross-species hybridization blots for E4.3 and H1.6), as is shown in Figure 4, indicate that an orthologous gene exists in the mouse (hereafter to be denoted mCF, and its corresponding protein as mCFTR), and this will be possible to clone in mouse genomic and cDNA libraries using the human CF gene probes. It is expected that the generation of a specific mutation in the mouse gene analogous to the I507 mutation will be most optimum to reproduce the phenotype, though complete inactivation of the mCFTR gene will also be a useful mutant to generate.

A. Mutagenesis. Inactivation of the mCF gene can be achieved by chemical [e.g. Johnson et al Proc. Natl. Acad. Sci. USA 78:3138 (1981)] or X-ray mutagenesis [Popp et al J. Mol. Biol. 127:141 (1979)] of mouse gametes, followed by fertilization. Offspring heterozygous for inactivation of mCFTR can then be identified by Southern blotting to demonstrate loss of one allele by dosage, or failure to inherit one parental allele if an RFLP marker is being assessed. This

approach has previously been successfully used to identify mouse mutants for α-globin [Whitney et al Proc. Natl. Acad. Sci. USA 77:1087 (1980)], phenylalanine hydroxylase [McDonald et al Pediatr. Res 23:63 (1988)], and carbonic anhydrase II [Lewis et al Proc. Natl. Acad. Sci. USA 85:1962, (1988)].

B. Transgenics A mutant version of CFTR or mouse CFTR can be inserted into the mouse germ line using now standard techniques of oocyte injection [Camper, Trends in Genetics (1988)]; alternatively, if it is desirable to inactivate or replace the endogenous mCF gene, the homologous recombination system using embryonic stem (ES) cells [Capecchi, Science 244:1288 (1989)] may be applied.

1. Oocyte Injection Placing one or more copies of the normal or mutant mCF gene at a random location in the mouse germline can be accomplished by microinjection of the pronucleus of a just-fertilized mouse oocyte, followed by reimplantation into a pseudo-pregnant foster mother. The liveborn mice can then be screened for integrants using analysis of tail DNA for the presence of human CF gene sequences. The same protocol can be used to insert a mutant mCF gene. To generate a mouse model, one would want to place this transgene in a mouse background where the endogenous mCF gene has been inactivated, either by mutagenesis (see above) or by homologous recombination (see below). The transgene can be either: a) a complete genomic sequence, though the size of this (about 250 kb) would require that it be injected as a yeast artificial chromosome or a chromosome fragment; b) a cDNA with either the natural promoter or a heterologous promoter; c) a "minigene" containing all of the coding region and various other elements such as introns, promoter, and 3' flanking elements found to be necessary for optimum expression.

2. Retroviral Infection of Early Embryos. This alternative involves inserting the CFTR or mCF gene into a retroviral vector and directly infecting mouse embroyos at early stages of development generating a chimera [Soriano et al Cell 46:19 (1986)]. At least some of these will lead to germline transmission.

3. ES Cells and Homologous Recombination. The embryonic stem cell approach (Capecchi, supra and Capecchi, Trends Genet 5:70 (1989)] allows the possibility of performing gene transfer and then screening the resulting totipotent cells to identify the rare homologous recombination events. Once identified, these can be used to generate chimeras by injection of mouse blastocysts, and a proportion of the resulting mice will show germline transmission from the recombinant line. There are several ways this could be useful in the generation of a mouse model for CF:

a) Inactivation of the mCF gene can be conveniently accomplished by designing a DNA fragment which contains sequences from a mCFTR exon flanking a selectable marker such as neo. Homologous recombination will lead to insertion of the neo sequences in the middle of an exon, inactivating mCFTR. The homologous recombination events (usually about 1 in 1000) can be recognized from the heterologous ones by DNA analysis of individual clones [usually using PCR, Kim et al Nucleic Acids Res. 16:8887 (1988), Joyner et al Nature 338:153 (1989); Zimmer et al supra, p. 150] or by using a negative selection against the heterologous events [such as the use of an HSV TK gene at the end of the construct, followed by the gancyclovir selection, Mansour et al, Nature 336:348 (1988)]. This inactivated mCFTR mouse can then be used to introduce a mutant CF gene or mCF gene containing the I507 abnormality or any other desired mutation.

b) It is possible that specific mutants of mCFTR cDNA be created in one step. For example, one can make a construct containing mCF intron 9 sequences at the 5' end, a selectable neo gene in the middle, and intro 9 + exon 10 (containing the mouse version of the I507 mutation) at the 3' end. A homologous recombination event would lead to the insertion of the neo gene in intron 9 and the replacement of exon 10 with the mutant version.

c) If the presence of the selectable neo marker in the intron altered expresson of the mCF gene, it would be possible to excise it in a second homologous recombination step.

d) It is also possible to create mutations in the mouse germline by injecting oligonucleotides containing the mutation of interest and screening the resulting cells by PCR.

This embodiment of the invention has considered primarily a mouse model for cystic fibrosis. Figure 4 shows cross-species hybridization not only to mouse DNA, but also to bovine, hamster and chichen DNA. Thus, it is contemplated that an orthologous gene will exist in many other species also. It is thus contemplated that it will be possible to generate other animal models using similar technology.

Although preferred embodiments of the invention have been described herein in detail, it will be understood by those skilled in the art that variations may be made thereto without departing from the spirit of the invention or the scope of the appended claims.

THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS

1. A DNA molecule comprising an intronless DNA sequence encoding a mutant CFTR polypeptide having the sequence according to Figure 1 for amino acid residue positions 1 to 1480, further characterized by a three base pair deletion which results in the deletion of isoleucine from amino acid residue position 506 or 507.

2. A DNA molecule comprising an intronless DNA sequence selected from the group consisting of:

(a) DNA sequences which correspond to the sequence of claim 1 and which encode, on expression, for mutant CFTR polypeptide;

(b) DNA sequences which correspond to a fragment of the sequences in claim 1 including at least 16 nucleotides;

(c) DNA sequences which comprise at least 16 nucleotides and encode a fragment of the amino acid sequence of claim 1; and

(d) DNA sequences encoding an epitope characteristic of the mutant CFTR protein encoded by at least 18 sequential nucleotides in the sequence of claim 1.

3. The DNA molecule of claim 1 wherein the DNA molecule is a cDNA.

4. The DNA molecule of claim 2 wherein the DNA molecule is a cDNA.

5. A purified RNA molecule comprising an RNA sequence corresponding to the DNA sequence recited in claim 2.

6. A purified nucleic acid probe comprising a DNA or

RNA nucleotide sequence corresponding to the sequence recited in parts (b), (c), or (d) of claim 2.

7. A nucleic acid probe according to claim 24 wherein said sequence comprises AAA GAA AAT ATC TTT GGT GTT, and its complement.

8. A recombinant cloning vector comprising the DNA molecule of claim 2.

9. The vector of claim 8 wherein said DNA molecule is operatively linked to an expression control sequence in said recombinant DNA molecule so that 506 or 507 mutant CFTR polypeptide can be expressed, said expression control sequence being selected from the group consisting of sequences that control the expression of genes of prokaryotic or eukaryotic cells and their viruses and combinations thereof.

10. The vector of claim 9 wherein the expression control sequence is selected from the group consisting of the lac system, the trp system, the tac system, the trc system, major operator and promoter regions of phage lambda, the control region of fd coat protein, the early and late promoters of SV40, promoters derived from polyoma, adenovirus, retrovirus, baculovirus and simian virus, the promoter for 3-phosphoglycerate kinase, the promoters of yeast acid phosphatase, the promoter of the yeast alpha-mating factors and combinations thereof.

11. A host transformed with the vector according to claim 8.

12. The host of claim 11 selected from the group consisting of strains of E. coli, Pseudomonas, Bacillus subtilis, Bacillus stearothermophilus, or other bacili;

other bacteria; yeast; fungi; insect; mouse or other animal; plant hosts; or human tissue cells.

13. The host of claim 12 wherein said human tissue cells are human epithelial cells.

14. A method for producing a 506 or 507 mutant CFTR polypeptide comprising the steps of:

(a) culturing a host cell transfected by the vector of claim 8 in a medium and under conditions favorable for expression of the 506 or 507 mutant CFTR polypeptide; and

(b) isolating the expressed 506 or 507 mutant CFTR polypeptide.

15. A mutant CFTR polypeptide substantially free of other human proteins and encoded by the DNA sequence recited in claim 2.

16. A substantially pure mutant CFTR polypeptide according to claim 15 made by chemical or enzymatic peptide synthesis.

17. A polypeptide coded for by expression of a DNA sequence recited in claim 2.

18. A method for screening a subject to determine if said subject is a CF carrier or a CF patient comprising the steps of:

providing a biological sample of the subject to be screened; and providing an assay for detecting in the biological sample, the presence of at least a member from the group consisting of a 506 or 507 mutant CF gene, 506 or 507 mutant CFTR polypeptide products and mixtures thereof.

19. The method of claim 18 wherein the biological sample includes at least part of the genome of the subject and the assay comprises an hybridization assay.

20. The method of claim 19 wherein the assay further comprises a labelled nucleotide probe according to claim 6.

21. The method of claim 20 wherein said probe comprises the nucleotide sequence of claim 7.

22. The method of claim 18 wherein the biological sample includes a CFTR polypeptide of the subject and the assay comprises an immunological assay.

23. The method of claim 22 wherein the assay further includes an antibody specific for said 506 or 507 mutant CFTR polypeptide.

24. The method of claim 22 wherein the assay is a radioimmunoassay.

25. The method of claim 23 wherein the antibody is at least one monoclonal antibody.

26. The method of claim 18 wherein the subject is a human fetus <u>in utero</u>.

27. The method of claim 20 wherein the assay further includes at least one additional nucleotide probe according to claim 6.

28. The method of claim 27, wherein the assay further includes a second nucleotide probe comprising a different DNA sequence fragment of the DNA of Figure 1 or its RNA homologue or a different DNA sequence

fragment of human chromosome 7 and located to either side of the DNA sequence of Figure 1.

29. In a process for screening a potential CF carrier or patient to indicate the presence of an identified cystic fibrosis 506 or 507 mutation in the CF gene, said process including the steps of:

(a) isolating genomic DNA from said potential CF carrier or said potential patient;

(b) hybridizing a DNA probe onto said isolated genomic DNA, said DNA probe spanning a 506 or 507 mutation in said CF gene wherein said DNA probe is capable of detecting said mutation;

(c) treating said genomic DNA to determine presence or absence of said DNA probe and thereby indicating in accordance with a predetermined manner of hybridization, the presence or absence of said cystic fibrosis mutation.

30. A process for detecting cystic fibrosis carriers of the 506 or 507 mutant CF gene wherein said process consists of determining differential mobility of heteroduplex PCR products in polyacrylamide gels as a result of deletions in the 506 or 507 mutant CF gene.

31. A kit for assaying for the presence of a 506 or 507 mutant CF gene by immunoassay comprising:

(a) an antibody which specifically binds to a gene product of the 506 or 507 mutant CF gene;

(b) reagent means for detecting the binding of the antibody to the gene product; and

(c) the antibody and reagent means each being present in amounts effective to perform the immunoassay.

32. The kit of claim 31 wherein said reagent means for detecting binding is selected from the group consisting

of fluorescence detection, radioactive decay detection, enzyme activity detection or colorimetric detection.

33. A kit for assaying for the presence of a CF gene by hybridization comprising:

(a) an oligonucleotide probe which specifically binds to the 507 mutant CF gene;

(b) reagent means for detecting the hybridization of the oligonucleotide probe to the 506 or 507 mutant CF gene; and

(c) the probe and reagent means each being present in amounts effective to perform the hybridization assay.

34. An animal comprising a heterologous cell system comprising a recombinant cloning vector of claim 8 which induces cystic fibrosis symptoms in said animal.

35. The animal of claim 34 wherein said animal is a mammal.

36. The animal of claim 35 wherein said mammal is a rodent.

37. The animal of claim 36 wherein said rodent is a mouse.

38. A transgenic mouse exhibiting cystic fibrosis symptoms.

39. In a polymerase chain reaction to amplify a selected exon of a cDNA sequence of Figure 1, the use of oligonucleotide primers from intron portions near the 5′ and 3′ boundaries of the selected exon of Figure 18.

40. In a polymerase chain reaction of claim 39, the use of oligonucleotide primers xi-5 and xi-3 of Figure 18

where X is the exon number 4, 6a, 6b, 6c, 7 through 13, 14a, 14b, 15 and 16, 17a, 17b and 18 through 24.

41. In a polymerase chain reaction of claim 40, said oligonucleotide primers being:

i) 5'-GCA GAG TAC CTG AAA CAG GAA-3'

ii) 5'-AGT GGT AAG CTA CTG TGA ATG-3'

from exon 10.

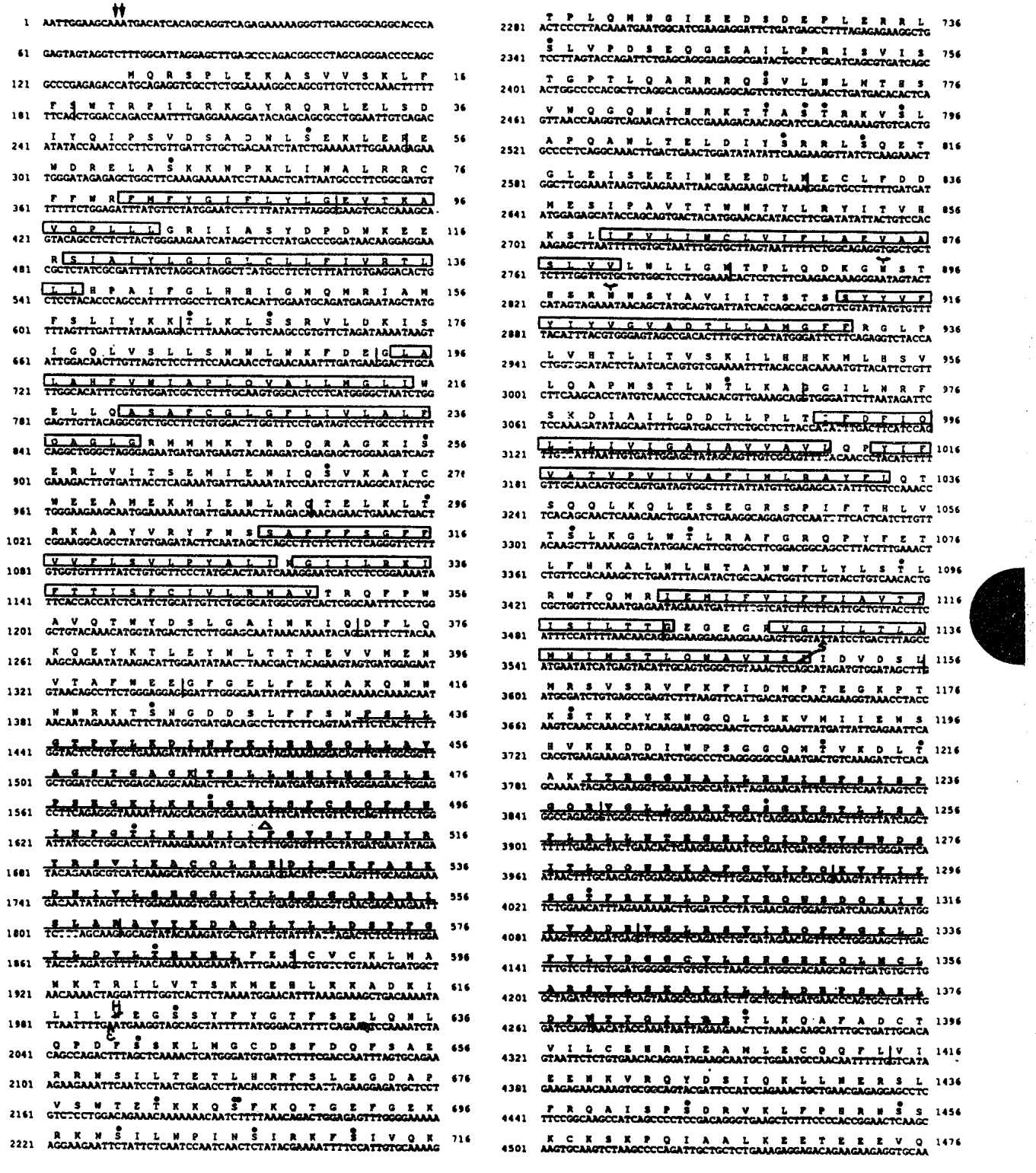


FIGURE 1

```
      D  T  R  L  =
4561  GATACAAGGCTTTAGAGAGCAGCATAAATGTTGACATGGGACATTTGCTCATGGAATTGG
4621  AGCTCGTGGGACAGTCACCTCATGGAATTGGAGCTCGTGGAACAGTTACCTCTGCCTCAG
4681  AAAACAAGGATGAATTAAGTTTTTTTTTAAAAAAGAAACATTTGGTAAGGGGAATTGAGG
4741  ACACTGATATGGGTCTTGATAAATGGCTTCCTGGCAATAGTCAAATTGTGTGAAAGGTAC
4801  TTCAAATCCTTGAAGATTTACCACTTGTGTTTTGCAAGCCAGATTTTCCTGAAAACCCTT
4861  GCCATGTGCTAGTAATTGGAAAGGCAGCTCTAAATGTCAATCAGCCTAGTTGATCAGCTT
4921  ATTGTCTAGTGAAACTCGTTAATTTGTAGTGTTGGAGAAGAACTGAAATCATACTTCTTA
4981  GGGTTATGATTAAGTAATGATAACTGGAAACTTCAGCGGTTTATATAAGCTTGTATTCCT
5041  TTTTCTCTCCTCTCCCCATGATGTTTAGAAACACAACTATATTGTTTGCTAAGCATTCCA
5101  ACTATCTCATTTCCAAGCAAGTATTAGAATACCACAGGAACCACAAGACTGCACATCAAA
5161  ATATGCCCCATTCAACATCTAGTGAGCAGTCAGGAAAGAGAACTTCCAGATCCTGGAAAT
5221  CAGGGTTAGTATTGTCCAGGTCTACCAAAAACTCTCAATATTTCAGATAATCACAATACAT
5281  CCCTTACCTGGGAAAGGGCTGTTATAATCTTTCACAGGGGACAGGATGGTTCCCTTGATG
5341  AAGAAGTTGATATGCCTTTTCCCAACTCCAGAAAGTGACAAGCTCACAGACCTTTGAACT
5401  AGAGTTTAGCTGGAAAAGTATGTTAGTGCAAATTGTCACAGGACAGCCCTTCTTTCCACA
5461  GAAGCTCCAGGTAGAGGGTGTGTAAGTAGATAGGCCATGGGCACTGTGGGTAGACACACA
5521  TGAAGTCCAAGCATTTAGATGTATAGGTTGATGGTGGTATGTTTTCAGGCTAGATGTATG
5581  TACTTCATGCTGTCTACACTAAGAGAGAATGAGAGACACACTGAAGAAGCACCAATCATG
5641  AATTAGTTTTATATGCTTCTGTTTTATAATTTTGTGAAGCAAAATTTTTTCTCTAGGAAA
5701  TATTTATTTTAATAATGTTTCAAACATATATTACAATGCTGTATTTTAAAAGAATGATTA
5761  TGAATTACATTTGTATAAAATAATTTTTATATTTGAAATATTGACTTTTTATGGCACTAG
5821  TATTTTTATGAAATATTATGTTAAAACTGGGACAGGGGAGAACCTAGGGTGATATTAACC
5881  AGGGGCCATGAATCACCTTTTGGTCTGGAGGGAAGCCTTGGGGCTGATCGAGTTGTTGCC
5941  CACAGCTGTATGATTCCCAGCCAGACACAGCCTCTTAGATGCAGTTCTGAAGAAGATGGT
6001  ACCACCAGTCTGACTGTTTCCATCAAGGGTACACTGCCTTCTCAACTCCAAACTGACTCT
6061  TAAGAAGACTGCATTATATTTATTACTGTAAGAAAATATCACTTGTCAATAAAATCCATA
6121  CATTTGTGT(A)n
```

FIGURE 1 (continued)

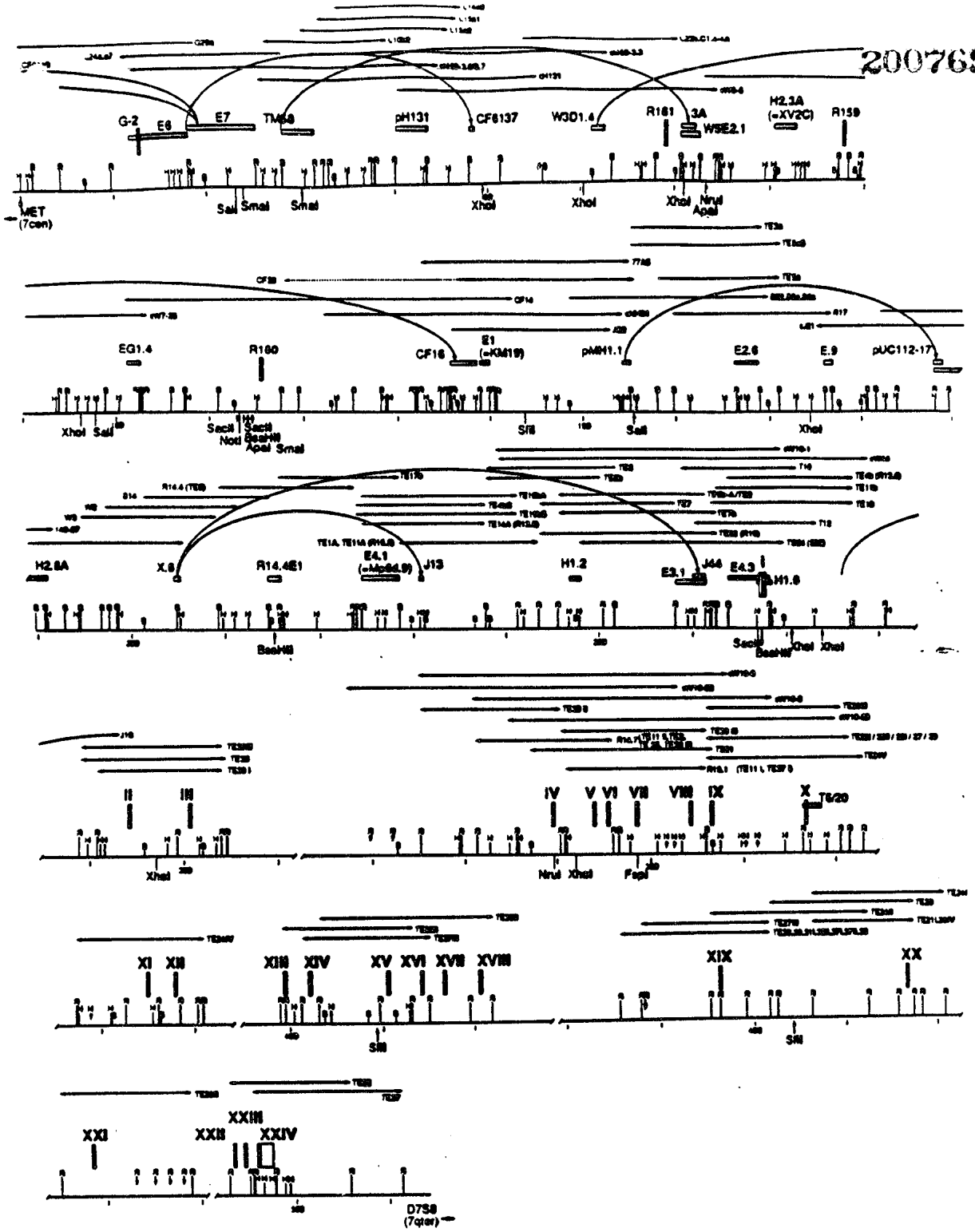

G-2
E6
E7
TM58
pH131
CF6137
W3D1.4
R161
3A
WSE2.1
H2.3A
(=XV2C)
R159
MET
(7cen)
SalI
SmaI
SmaI
XhoI
XhoI
XhoI
NruI
ApaI
EG1.4
R160
CF16
E1
(=KM19)
pMH1.1
E2.6
E.9
pUC112-17
XhoI
SalI
SacII
NotI
SacII
BssHII
ApaI
SmaI
SfiI
SalI
XhoI
H2.8A
X.6
R14.4E1
E4.1
(=Mp64.9)
J13
H1.2
E3.1
J44
E4.3
H1.8
BssHII
SacII
BssHII
XhoI
XhoI
II
III
IV
V
VI
VII
VIII
IX
X
T6/20
XhoI
NruI
XhoI
FspI
XI
XII
XIII
XIV
XV
XVI
XVII
XVIII
XIX
XX
SfiI
SfiI
XXI
XXII
XXIII
XXIV
D7S8
(7qter)

Figure 2.

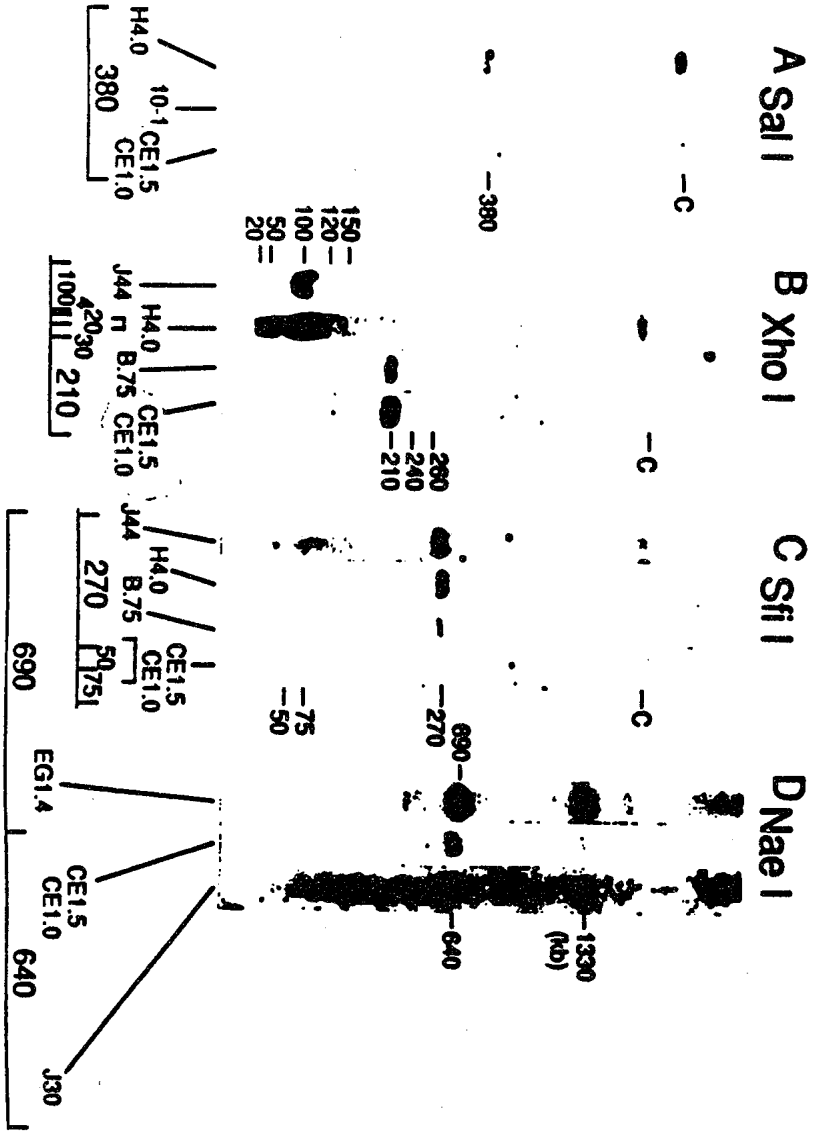

A Sal I
B Xho I
C Sfi I
D Nae I
-C
-380
-260
-240
-210
-270
-75
-50
690
640
1330
(kb)
H4.0
10-1
CE1.5
CE1.0
380
150
120
100
50
20
J44
n
B.75
210
270
690
EG1.4
J30

Figure 3 A-D

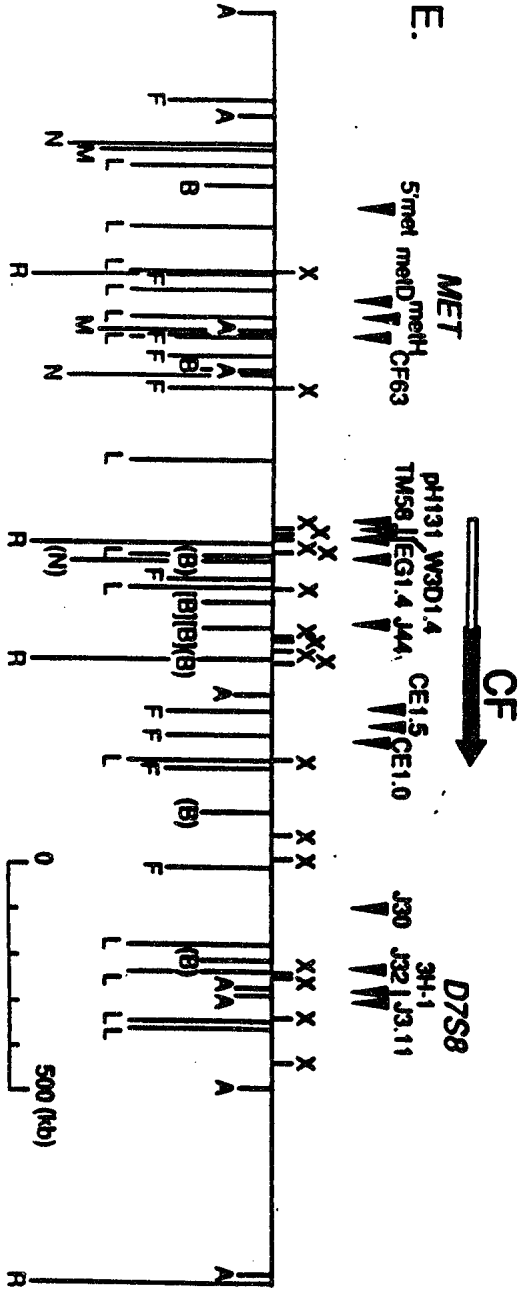
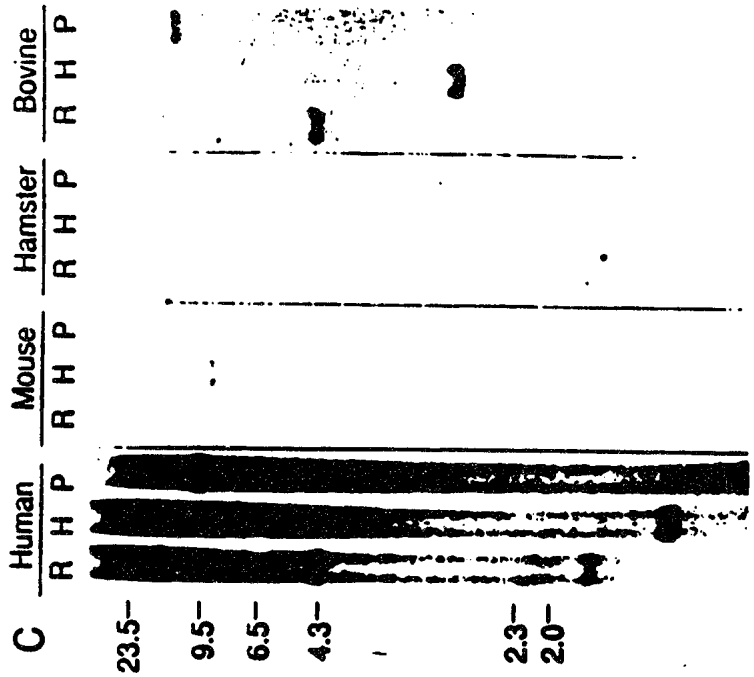

E.
MET
5'met
metD
metH
CF63
pH131
W3D1.4
TM58
EG1.4
J44
CE1.5
CE1.0
CF
D7S8
3H-1
J30
J32
J3.11
0
500 (kb)

Figure 3E.

Figure 4A.


A
Human
Bovine
Mouse
Chicken
R H P
R H P
R H P
R H P
23.5
9.5
6.6
4.3
2.3
2.0
0.56

Figure 4B, 4C


C
Human
Mouse
Hamster
Bovine
R H P R H P R H P R H P
23.5
9.5
6.5
4.3
2.3
2.0

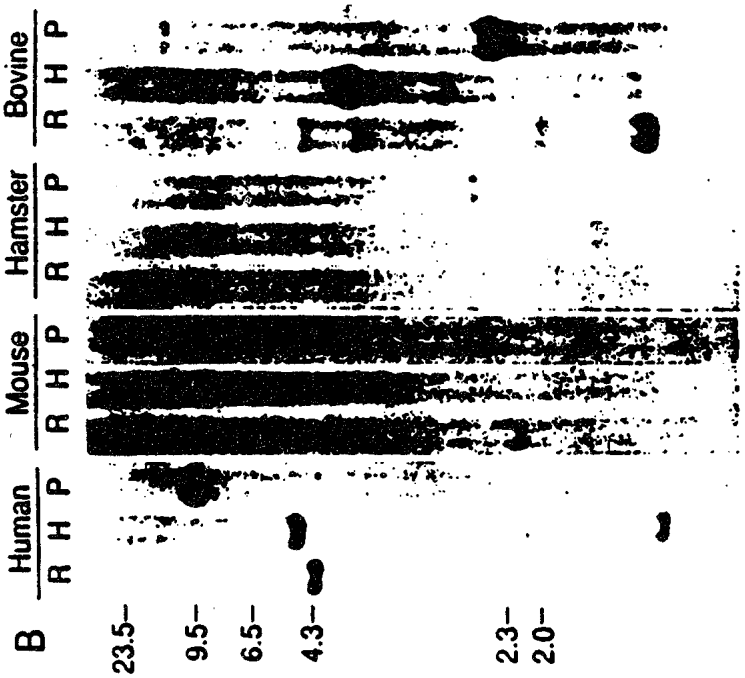

B
Human
Mouse
Hamster
Bovine
R H P R H P R H P R H P
23.5
9.5
6.5
4.3
2.3
2.0

Figure 4D

D.

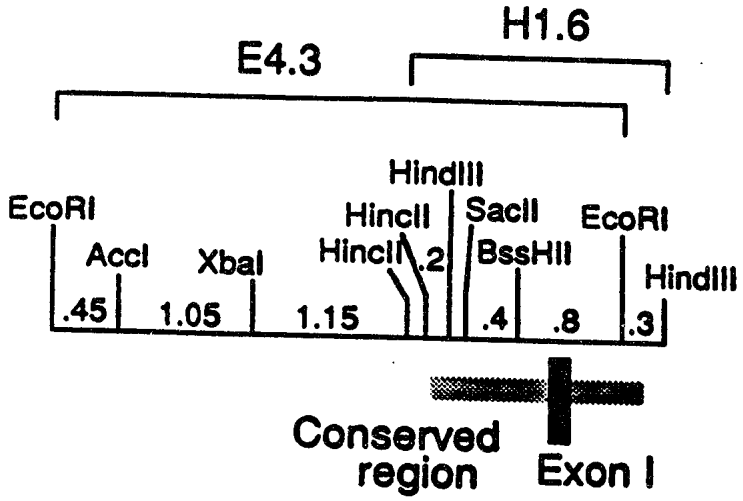

H1.6
E4.3
HindIII
EcoRI
HincII
SacII
EcoRI
AccI
XbaI
HincII
.2
BssHII
HindIII
.45
1.05
1.15
.4
.8
.3
Conserved region
Exon I

Figure 5

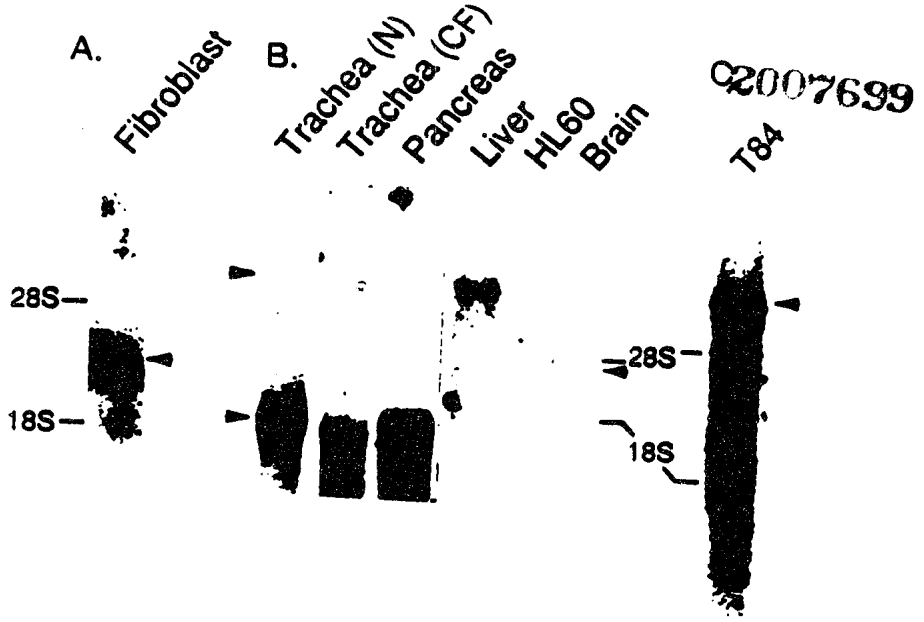
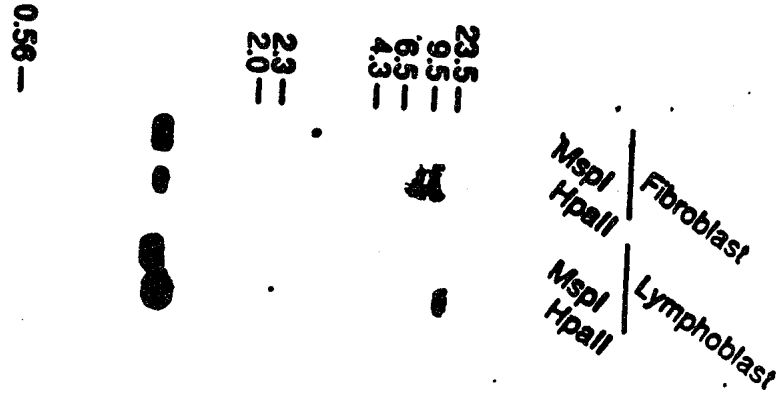
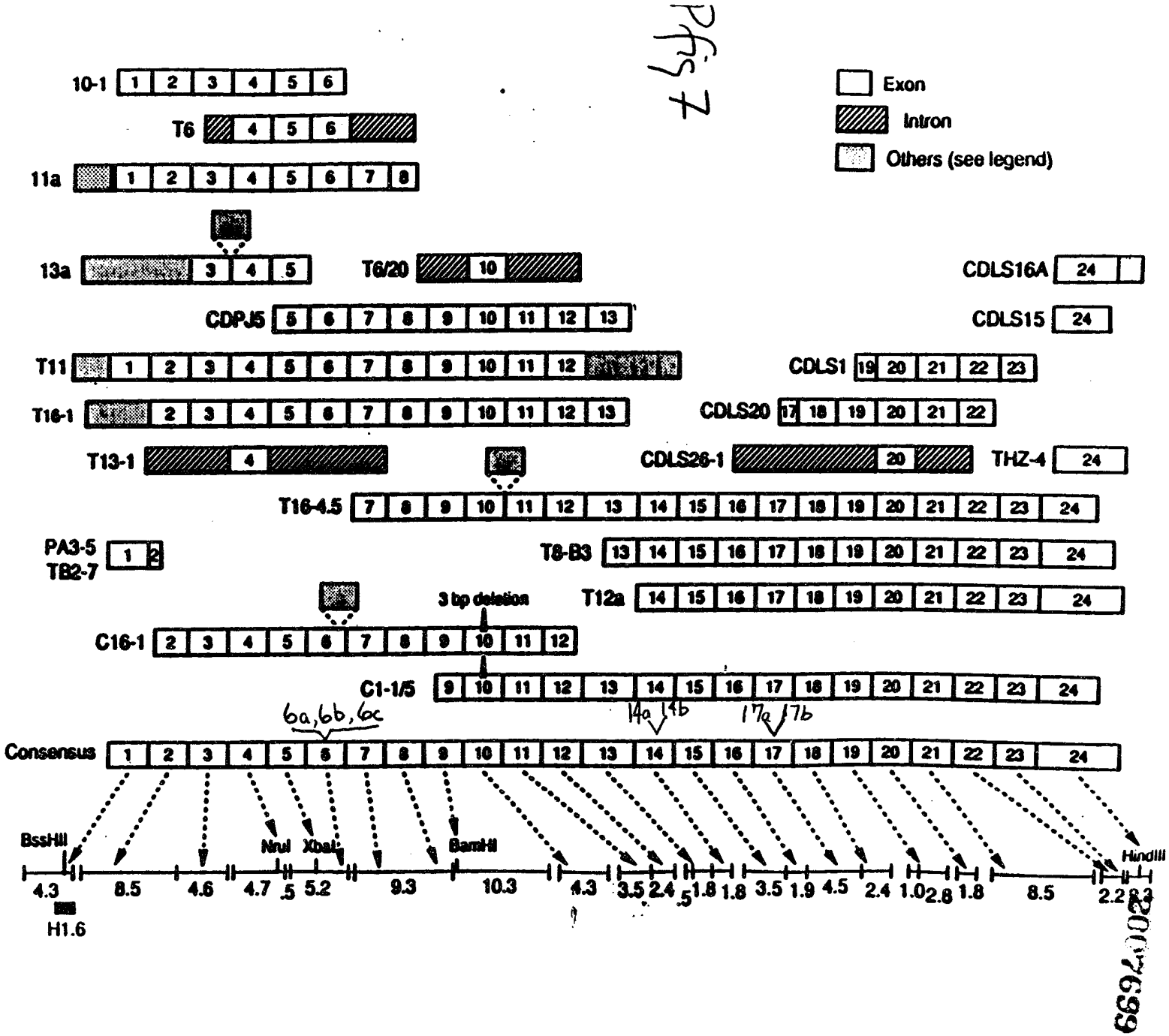


Figure 6



P. fig 6


Exon
Intron
Others (see legend)
10-1
T6
11a
13a
T6/20
CDPJ5
T11
T16-1
T13-1
T16-4.5
PA3-5
TB2-7
C16-1
C1-1/5
CDLS16A
CDLS15
CDLS1
CDLS20
CDLS26-1
THZ-4
T8-B3
T12a
3 bp deletion
6a,6b,6c
14a 14b
17a 17b
Consensus
BssHII
NruI
XbaI
BamHI
HindIII
4.3
8.5
4.6
4.7
.5
5.2
9.3
10.3
4.3
3.5
2.4
.5
1.8
1.8
3.5
1.9
4.5
2.4
1.0
2.8
1.8
8.5
2.2
.3
H1.6

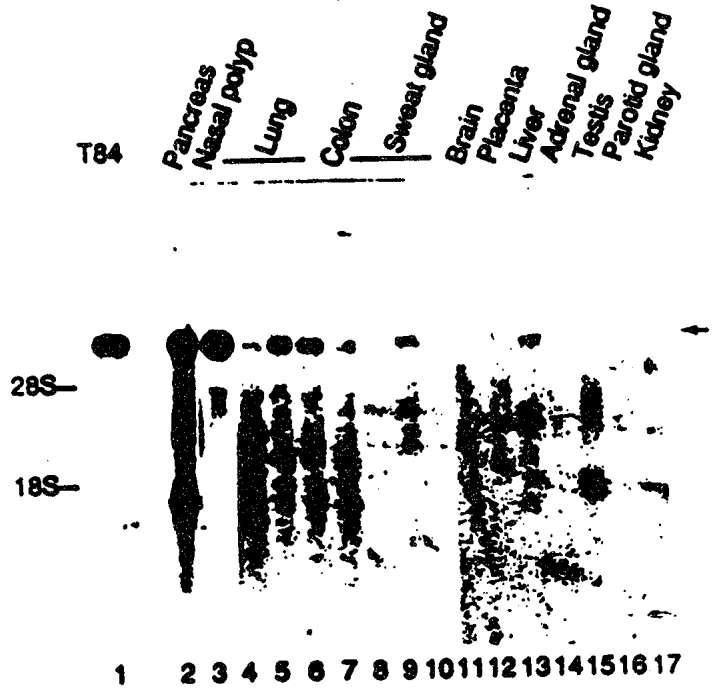

T84
Pancreas
Nasal polyp
Lung
Colon
Sweat gland
Brain
Placenta
Liver
Adrenal gland
Testis
Parotid gland
Kidney
28S
18S
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Figure 8.

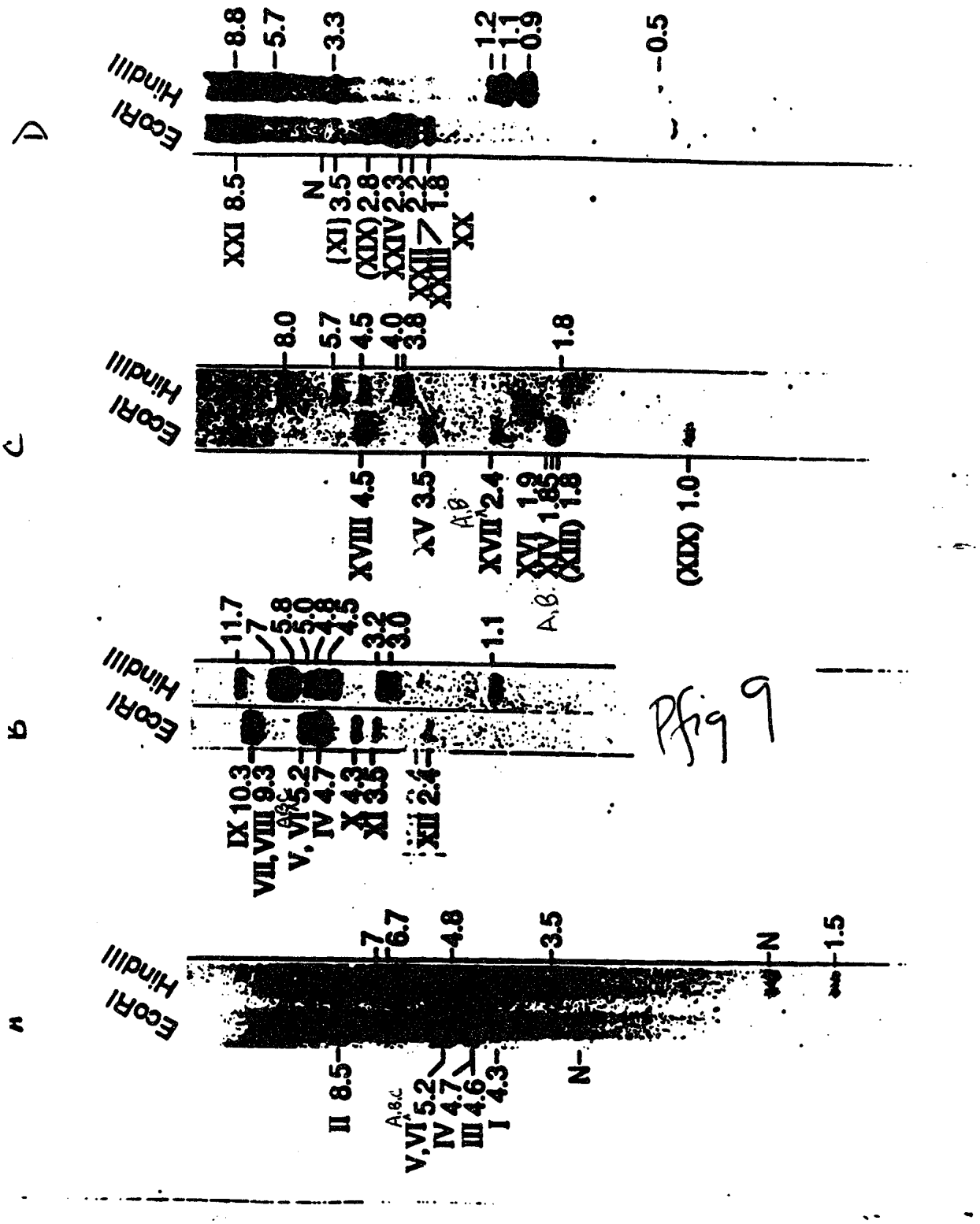

a
EcoRI
HindIII
II 8.5
V,VI 5.2
IV 4.7
III 4.6
I 4.3
N
7
6.7
4.8
3.5
N
1.5
b
EcoRI
HindIII
IX 10.3
VII,VIII 9.3
V, VI 5.2
IV 4.7
X 4.3
XI 3.5
XII 2.4
11.7
7
5.8
5.0
4.8
4.5
3.2
3.0
1.1
A,B
c
EcoRI
HindIII
XVIII 4.5
XV 3.5
XVII 2.4
XVI 1.9
(XIII) 1.8
(XIX) 1.0
8.0
5.7
4.5
4.0
3.8
1.8
A,B
d
EcoRI
HindIII
XXI 8.5
N
(XI) 3.5
(XII) 2.8
XXIV 2.3
XXIII 1.8
XX
8.8
5.7
3.3
1.2
1.1
0.9
0.5

Fig 9

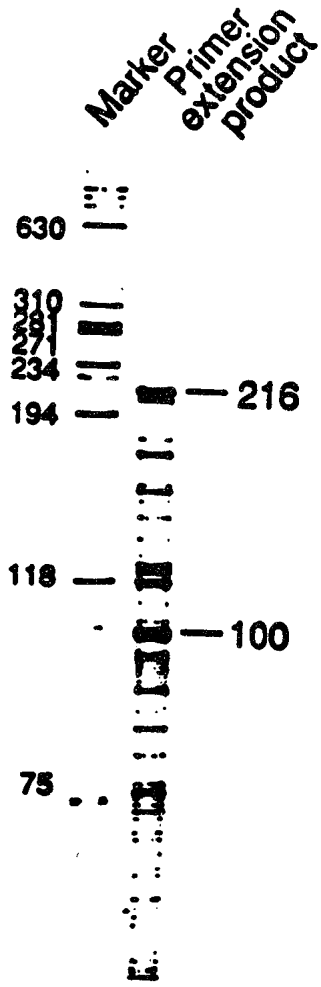

Marker
Primer extension product
630
310
281
271
234
194
118
75
216
100

Fig 10A

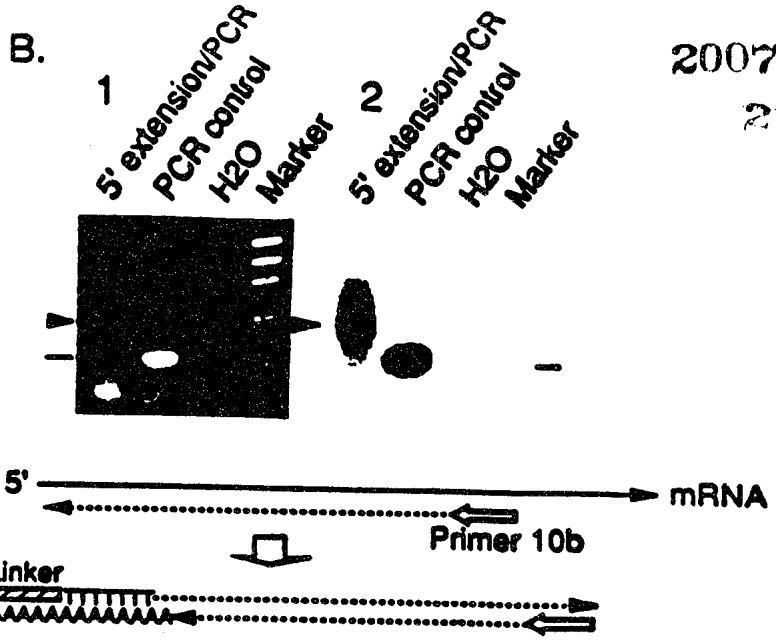


Figure 10B

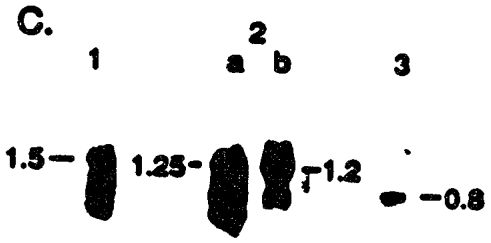
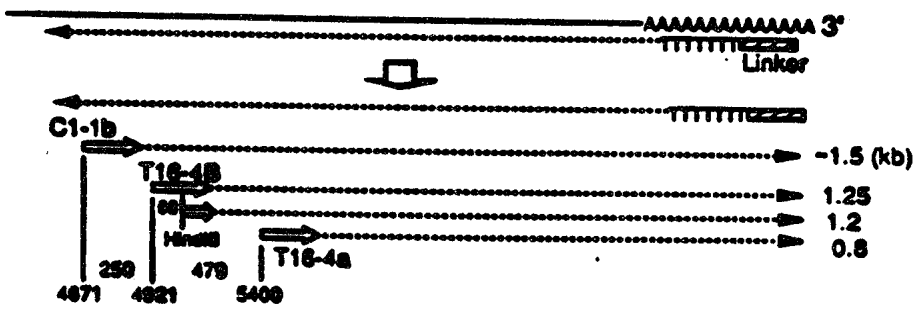


Figure 10C



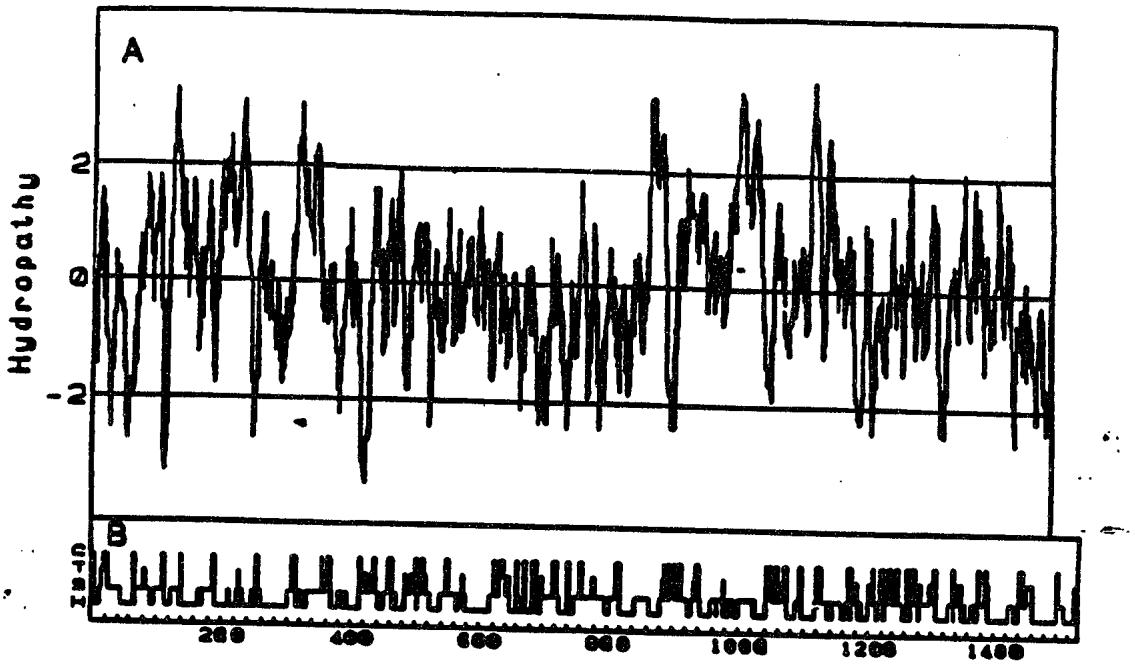


Fig 11

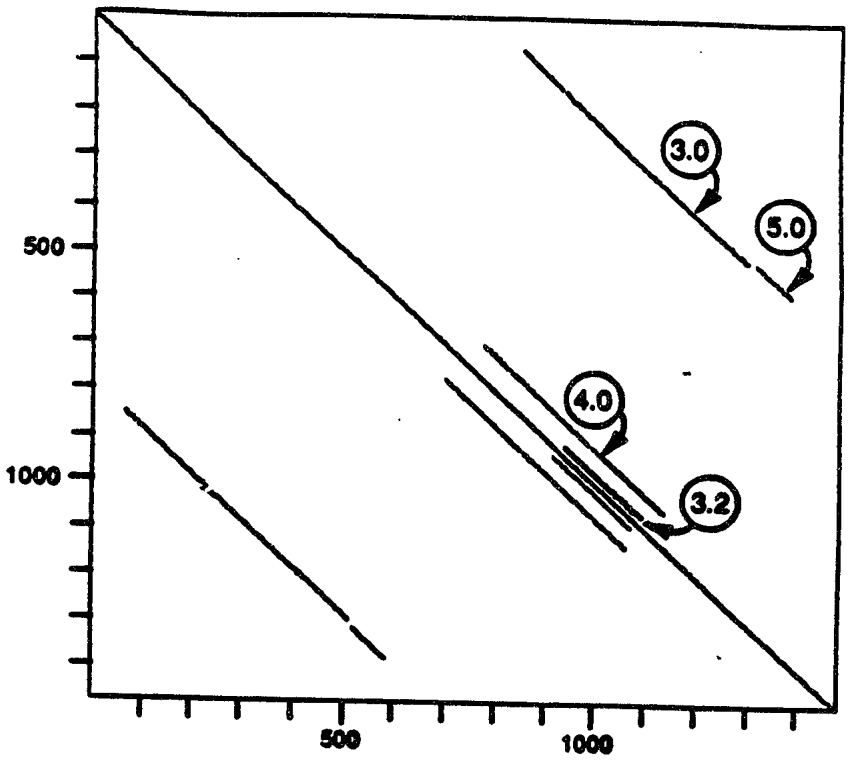


fig 12

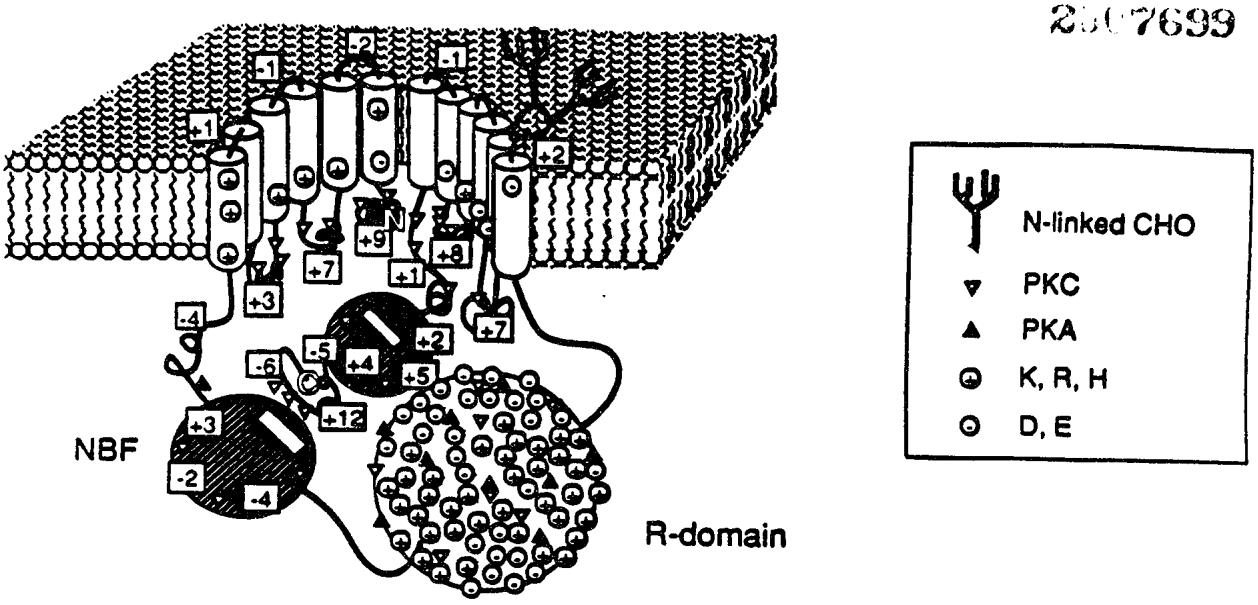


Figure 13.

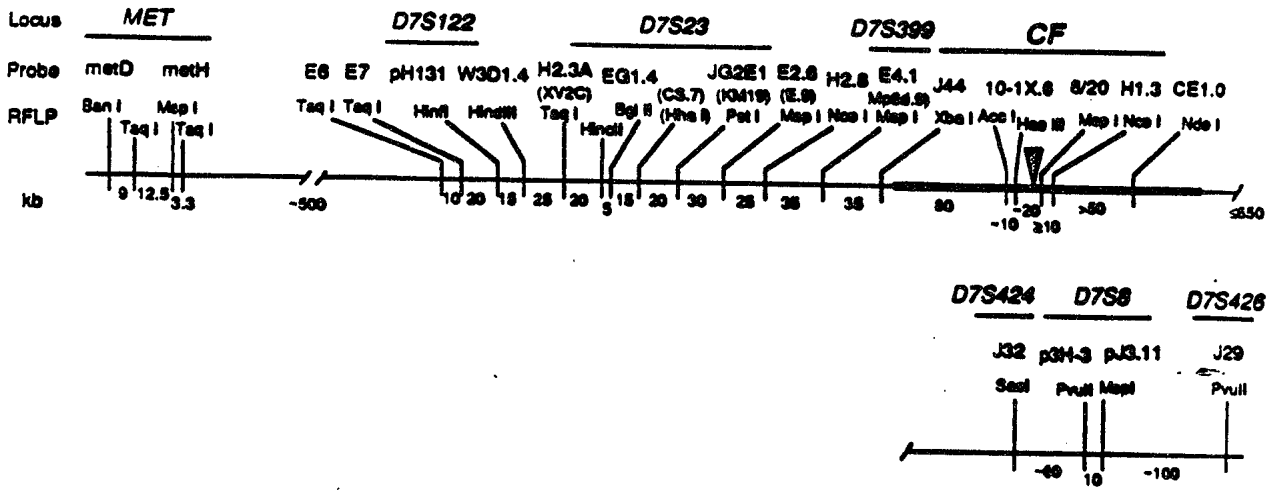

Locus
MET
D7S122
D7S23
D7S399
CF
Probe
metD
metH
E6
E7
pH131
W3D1.4
H2.3A
(XV2C)
EG1.4
JG2E1
E2.8
H2.8
E4.1
J44
10-1X.6
8/20
H1.3
CE1.0
RFLP
Ban I
Msp I
Taq I
Taq I
Taq I
Taq I
Hinf I
HindIII
Taq I
HindII
Bgl II
(CS.7)
(Hha I)
(KM19)
Pst I
(E.8)
Msp I
Nae I
MboII.8)
Msp I
Xba I
Ace I
Hae III
Msp I
Nae I
Nde I
kb
9
12.5
3.3
~500
10
20
15
25
20
5
15
20
30
25
35
35
80
~10
~20
≥10
>50
≤550
D7S424
D7S8
D7S426
J32
p3H-3
pJ3.11
J29
SasI
PvuII
MaeI
PvuII
~60
10
~100

Figure 14.

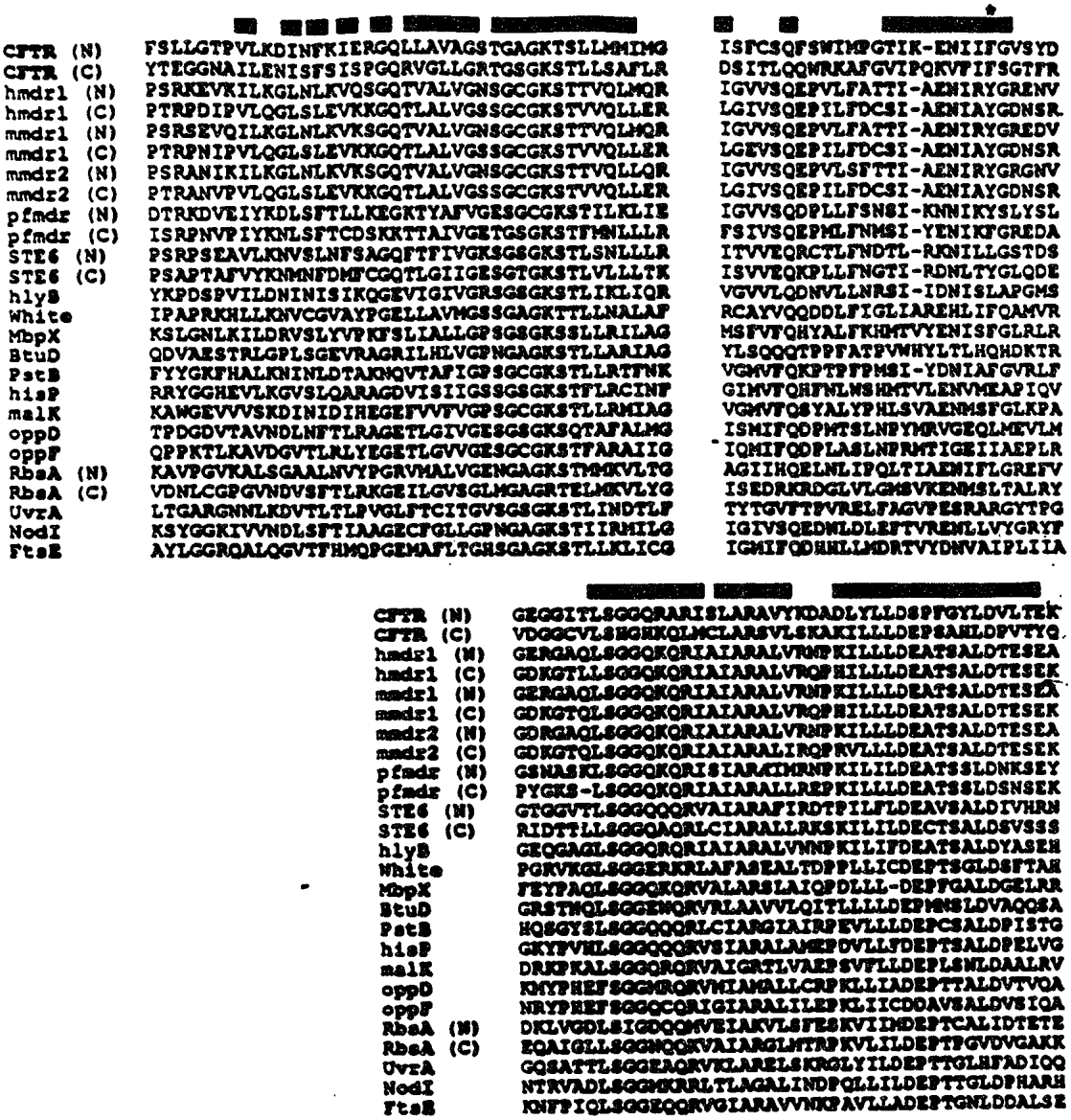
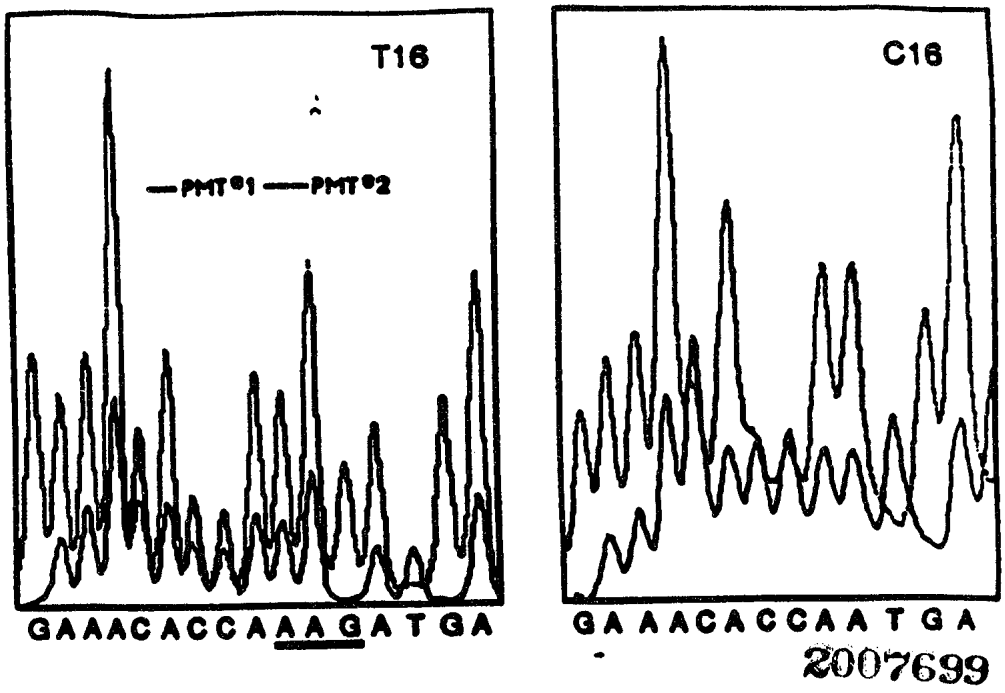

CFTR (N) FSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMG ISFCSQFSWIMPGTIK-ENIIFGVSYD
CFTR (C) YTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLR DSITLQQWRKAFGVIPQKVFIFSGTFR
hmdr1 (N) PSRKEVKILKGLNLKVQSGQTVALVGNSGCGKSTTVQLMQR IGVVSQEPVLFATTI-AENIRYGRENV
hmdr1 (C) PTRPDIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER LGIVSQEPILFDCSI-AENIAYGDNSR
mmdr1 (N) PSRSEVQILKGLNLKVKSGQTVALVGNSGCGKSTTVQLMQR IGVVSQEPVLFATTI-AENIRYGREDV
mmdr1 (C) PTRPNIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER LGEVSQEPILFDCSI-AENIAYGDNSR
mmdr2 (N) PSRANIKILKGLNLKVKSGQTVALVGNSGCGKSTTVQLLQR IGVVSQEPVLSFTTI-AENIRYGRGNV
mmdr2 (C) PTRANVPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER LGIVSQEPILFDCSI-AENIAYGDNSR
pfmdr (N) DTRKDVEIYKDLSFTLLKEGKTYAFVGESGCGKSTILKLIE IGVVSQDPLLFSNSI-KNNIKYSLYSL
pfmdr (C) ISRPNVPIYKNLSFTCDSKKTTAIVGETGSGKSTFMNLLLR FSIVSQEPMLFNMSI-YENIKFGREDA
STE6 (N) PSRPSEAVLKNVSLNFSAGQFTFIVGKSGSGKSTLSNLLLR ITVVEQRCTLFNDTL-RKNILLGSTDS
STE6 (C) PSAPTAFVYKNMNFDMFCGQTLGIIGESGTGKSTLVLLLTK ISVVEQKPLLFNGTI-RDNLTYGLQDE
hlyB YKPDSPVILDNINISIKQGEVIGIVGRSGSGKSTLIKLIQR VGVVLQDNVLLNRSI-IDNISLAPGMS
White IPAPRKHLLKNVCGVAYPGELLAVMGSSGAGKTTLLNALAF RCAYVQQDDLFIGLIAREHLIFQAMVR
MbpX KSLGNLKILDRVSLYVPKFSLIALLGPSGSGKSSLLRILAG MSFVFQHYALFKHMTVYENISFGLRLR
BtuD QDVAESTRLGPLSGEVRAGRILHLVGPNGAGKSTLLARIAG YLSQGQTPPFATPVWHYLTLHQHDKTR
PstB FYYGKFHALKNINLDTAKNQVTAFIGPSGCGKSTLLRTFNK VGMVFQKPTPFPMSI-YDNIAFGVRLF
hisP RRYGGHEVLKGVSLQARAGDVISIIGSSGSGKSTFLRCINF GIMVFQHFNLNSHMTVLENVMEAPIQV
malK KAWGEVVVSKDINIDIHEGEFVVFVGPSGCGKSTLLRMIAG VGMVFQSYALYPHLSVAENMSFGLKPA
oppD TPDGDVTAVNDLNFTLRAGETLGIVGESGSGKSQTAFALMG ISMIFQDPMTSLNPYMRVGEQLMEVLM
oppF QPPKTLKAVDGVTLRLYEGETLGVVGESGCGKSTFARAIIG IQMIFQDPLASLNPRMTIGEIIAEPLR
RbsA (N) KAVPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLTG AGIIHQELNLIPQLTIAENIFLGREFV
RbsA (C) VDNLCGPGVNDVSFTLRKGEILGVSGLMGAGRTELMKVLYG ISEDRKRDGLVLGMSVKENMSLTALRY
UvrA LTGARGNNLKDVTLTLPVGLFTCITGVSGSGKSTLINDTLF TYTGVFTPVRELFAGVPESRARGYTPG
NodI KSYGGKIVVNDLSFTIAAGECFGLLGPNGAGKSTIIRMILG IGIVSQEDMLDLEFTVRENLLVYGRYF
FtsE AYLGGRQALQGVTFHMQPGEMAFLTGHSGAGKSTLLKLICG IGMIFQDHHLLMDRTVYDNVAIPLIIA
CFTR (N) GEGGITLSGGQRARISLARAVYKDADLYLLDSPFGYLDVLTEK
CFTR (C) VDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVTYQ
hmdr1 (N) GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
hmdr1 (C) GDKGTLLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr1 (N) GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr1 (C) GDKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr2 (N) GDRGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr2 (C) GDKGTQLSGGQKQRIAIARALIRQPRVLLLDEATSALDTESEK
pfmdr (N) GSNASKLSGGQKQRISIARAIMRNPKILILDEATSSLDNKSEY
pfmdr (C) PYGKS-LSGGQKQRIAIARALLREPKILLLDEATSSLDSNSEK
STE6 (N) GTGGVTLSGGQQQRVAIARAFIRDTPILFLDEAVSALDIVHRN
STE6 (C) RIDTTLLSGGQAQRLCIARALLRKSKILILDECTSALDSVSSS
hlyB GEQGAGLSGGQRQRIAIARALVNNPKILIFDEATSALDYASEH
White PGRVKGLSGGERKRLAFASEALTDPPLLICDEPTSGLDSFTAH
MbpX FEYPAQLSGGQKQRVALARSLAIQPDLLL-DEPFGALDGELRR
BtuD GRSTHQLSGGEWQRVRLAAVVLQITLLLLDEPMNSLDVAQQSA
PstB HQSGYSLSGGQQQRLCIARGIAIRPEVLLLDEPCSALDPISTG
hisP GKYPVHLSGGQQQRVSIARALAMEPDVLLFDEPTSALDPELVG
malK DRKPKALSGGQRQRVAIGRTLVAEPSVFLLDEPLSNLDAALRV
oppD KMYPHEFSGGMRQRVMIAMALLCRPKLLIADEPTTALDVTVQA
oppF NRYPHEFSGGQCQRIGIARALILEPKLIICDEAVSALDVSIQA
RbsA (N) DKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTCALIDTETE
RbsA (C) EQAIGLLSGGHQQKVAIARGLMTRPKVLILDEPTFGVDVGAKK
UvrA GQSATTLSGGEAQRVKLARELSKRGLYILDEPTTGLHFADIQQ
NodI NTRVADLSGGMKRRLTLAGALINDPQLLILDEPTTGLDPHARH
FtsE KNFPIQLSGGEQQRVGIARAVVNKPAVLLADEPTGNLDDALSE

Figure 15

fig. 16



2007699

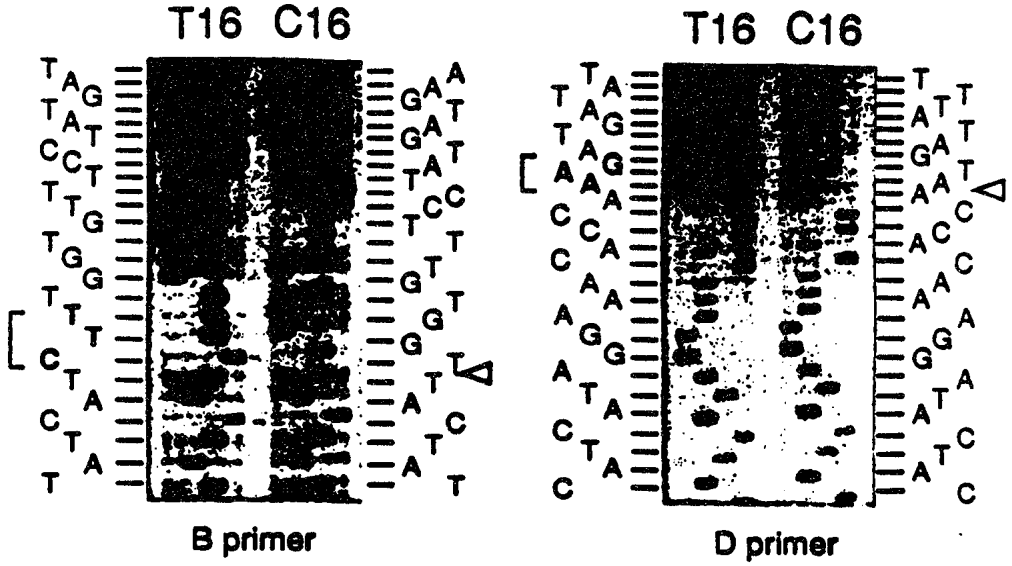


Figure 17

EXON: 4

GENOMIC CLONE: R14.7E4.7

FIG. 18

2007699

```
  1 AAAATACCTCATATGTAAACTTGTCTCCCACTGTTGCTATAACAAATCCCAAGTCTTATT
 61 TCAAAGTACCAAGATATTGAAAATAGTGCTAAGAGTTTCACATATGGTATGACCCTCTAT
121 ATAAACTCATTTTAAGTCTCCTCTAAAGATGAAAAGTCTTGTGTTGAAATTCTCAGGGTA
                                                    |  -------
181 TTTTATGAGAAATAAATGAAATTTAATTTCTCTGTTTTTCCCCTTTGTAG|GAAGTCACCA
    ----------> 10G                                  |
241 AAGCAGTACAGCCTCTCTTACTGGGAAGAATCATAGCTTTCCTATGACCCGGCTAACAAG
301 GAGGAACGCTCTATCGCGATTTATCTAGGCATAGGCTTATGCTTCTCTTTATTGTGAGGA
                                                        10L <-
361 CACTGCTCCTACACCCAGCCATTTTTGGCCTTCATCACATTGGAATGCAGATGAGAATAG
    ------------------
                        |
421 CTATGTTTAGTTTGATTTATAAGAAG|GTAATACTTCCTTGCACAGGCCCCATGGCACATA
                        |
481 TATTCTGTATCGTACATGTTTTAATGTCATAAATTAGGTAGTGAGCTGGTACAAGTAAGG
541 GATAAATGCTGAAATTAATTTAATATGCCTATTAAATAAATGGCAGGAATAATTAATGCT
601 CTTAATTATCCTTGATAATTT
```

EXON 4

2007699

EXON: 6A, 6B & 6C

GENOMIC CLONE: TE11IIE5.2

FIG. 18 (cont'd)

```
   1  CTTAAGATGTCCAATCTTGATTCCGACTGAATAAAATATGCTTAAAAATGCACTGACTTG
  61  GAAATTTGTTTTTTGGGAAAACCGATTCTATGTGTAGAATGTTTAAGCACATTGCTATGT
 121  GCTCCATGTAATGATTACCTAGATTTTAGTGTGCTCAGAACCACGAAGTGTTTGATCATA
 181  TAAGCTCCTTTTACTTGCTTTCTTTCATATATGATTGTTAGTTTCTAGGGGTGGAAGATA
                                            |
 241  CAATACACCTGTTTTTGCTGTGCTTTTATTTTCCAG|GGACTTGCATTGGCACATTTCGTG      |
                                            |                               |
                                            ------------------> Ex6-K2     |
 301  TGGATCGCTCCTTTGCAAGTGGCACTCCTNATGGGGCTAATCTGGGAGTTGTTACAGGCG        |
                                                      10K <------           |
                                                                            |
 361  TCTGCCTTCTGTGGACTTGGTTTCCTGATAGTCCTTGCCCTTTTTCAGGCTGGGCTAGGG  EXON 6A |
      -------------                                                         |
                           |
 421  AGAATGATGATGAAGTACAG|GTAGCAACCTATTTTCATAACTTGAAAGTTTTAAAAATTA
                           |
 481  TGTTTTCAAAAAGCCCACTTTAGTAAAACCAGGACTGCTCTAGCATAGAACAGTGATCTT
 541  CAGTGTCATTAAATTTTTTTTTTTTTTTTTTTTGAGACAGAGTCTAGATCTGTCACCCAG
 601  GCTGGAGTGCAGTGGCACGATCTTGGCTCACTGCACTGCAACTTCTGCCTCCCAGGCTCA
 661  AGCAATTCTCCTGCCTCAGCCTCCGGAGTAGCTGGGATTAGAGGCGCATNACCACACCCA
                    gap      |
 721  GCTAATTTTNNNNNNNNNNNNNN|AGATCAGAGAGCTGGGAAGATCAGTGAAAGACTTGTG    |
                             |                                          |
                |                                               EXON 6B |
                       gap                                              |
 781  ATTACCTCAG|NNNNNNNNNNNNNNNNNNNNNNTTTGAGTCAAATAGCACTTTGTTTGAAT       |
                |                                                       |
 841  CCCAGATTGCAGTCTTACTAGTTATGTGACCTTAGTCAAGCACTTCACCTCACTGAGTCT
 901  TTGCTTTTTTCATCTCTAAAATAGAGATACCCACCGCTCATAGGCTGTCATAAGGATAGA
 961  GATAGCATATGGAATGAGTCTGTACAGCGTCTGGCACATAGGAGGCATTTACCAAACAGC
1021  AGTTATTATTTTTGTTACCATCTATTTGATAATAAAATAATGCCCATCTGTTGAANAAAA
1081  GAAATATGACTTAAAACCTTGAGCAGTTCTTAATAGATAATTTGACTTGTTTTTACTATT
1141  AGATTGATTGATTGATTGATTGATTGATTTACAGAGATCAGAGAGCTGGGAAGATCAGTG
                              |           CCGGAAT-----------------------
1201  AAAGACTTGTGATTACCTCAG|AAATGATTGAAAATATCCAATCTGTTAAGGCATACTGCT    |
                              |                              EX6-F2 <   |
      > 10F                                                     EXON 6C |
                                                   |                    |
1261  GGGAAGAAGCAATGGAAAAAATGATTGAAAACTTAAGACA|GTAAGTTGTTCCNAATAATT      |
      ----------------                             |                    |
1321  TCAATATTGTTAGTAATTCTGTCCTTAATTTTTTAAAAATATGTTTATCATGGTAGACTT
1381  CCACCTCATATTTGATGTTTGTGACAATCAAATGATTGCATTTAAGTTCTGTCAATATTC
1441  ATGCATTAGTTGCACAAATTCACTTTCATGGGCTGTAGTTTATGTAGTTGGTCCAGGGT
```

1501 GTTGTTTTATGCTGCAAGTATATTATACTGATACGTTATTAAAGAATTTCCTACATATGT

1561 TCACTGCTGCTCAATACATTTATTTCGTTAAAAACAA

2007699

FIG. 18 (cont'd)

2007699

FIG. 18 (contd)

EXON: 7

GENOMIC CLONE: R15.1H3.0

```
  1  CAACTGGTACTTTCATTGTTATCTTTTCATATAAGGTAACTGAGGCCCAGAGAGATTAAA
 61  TAACATGCCCAAGGTCACACAGGTCATATGATGTGGAGCCAGGTTAAAAATATAGGCAGA
121  AAGACTCTAGAGACCATGCTCAGTCTTCCATTCCAAGATCCCTGATATTTGAAAAATAAA
181  ATAACATCCTGAATTTTATTGTTATTGTTTTTTATAGAACAG|AACTGAAACTGACTCGGA
241  AGGCAGCCTATGTGAGATACTTCAATAGCTCAGCCTTCTTCTCAGGGTTCTTTGTGGTGT
301  TTTTATCTGTGCTTCCCTATGCACTAATCAAAGGAATCATCCTCCGGAAAATATTCACCA
       11D <----------------

             ---------------------> 11G                          EXON 7
361  CCATCTCATTCTGCATTGTTCTGCGCATGGCGGTCACTCGGCAATTTCCCTGGGCTGTAC
421  AAACATGGTATGACTCTCTTGGAGCAATAAACAAAATACAG|GTAATGTACCATAATGCTG
481  CATTATATACTATGATTTAAATAATCAGTCAATAGATCAGTTCTAATGAACTTTGCAAAA
                                    7i-3 <----------------------
541  ATGTGCGAAAAGATAGAAAAAGAAATTTCCTTCACTAGGAAGTTATAAAAGTTGCCAGCT
601  AATACTAGGAATGTTCACCTTAAACTTTTCCTAGCATTTCTCTGGACAGTATGATGGATG
661  AGAGTGGCATTTATGCAAATTACCTTAAAA
```

EXON: 8

GENOMIC CLONE: A15.1H3.0

FIG. 18 (cont'd)

```
  1  GGTTGCTTTGTAAATTCATCACTAAGGTTAGCATGTAATAGTACAAGGAAGAATCAGTTG
 61  TATGTTAAATCTAATGTATAAAAAGTTTTATAAAATATCATATGTTTAGAGAGTATATTT
               ------------------------> 8i-5
121  CAAATATGATGAATCCTAGTGCTTGGCAAATTAACTTTAGTTCACTAATAAAATTATTTT
181  ATTAAGAAATAATTACTATTTCATTATTAAAATTCATATATAAGATGTAGCACAATGAGA
241  GTATAAAGTAGATGTAATAATGCATTAATGCTATTCTGATTCTATAATATGTTTTTGCTC
                   |
301  TCTTTTATAAATAG|GATTTCTTACAAAAGCAAGAANATAAGACATTGGAATATAACTTAA      |
                   |                                                    |
     ---------------------> T16H                                        |
361  CGACTACAGAAGTAGTGATGGAGAATGTAACAGCCTTCTGGGAGGAG|GTCAGAATTTTTA  EXON 8 |
                    T16G <-----------------------                       |
421  ANNAATTGTTTGCTCTAAACACCTAACTGTTTTCTTCTTTGTGAATATGGATTTCATCCT
                                                    8i-3 <-------------
481  AATGGCGAATAAAATTAGAATGATGATATAACTGGTAGAACTGGAAGGAGGATCACTCAC
     -----------
541  TTATTTTCTAGATTAAGAAGTAGAGGAATGGCCAGGTGCTCATGGTTGTAATCCAGCACT
601  TTNGGAGACCAAGGCGGGTGGATCACCTGAGGTCAGGAGTTCAAGACCAGCCTGCCAACA
661  TGGTAAANCCCNCGGTCTCTACTAAAAATACAAAAAATTAACTG
```

FIG. 18 (cont'd)

2007699

EXON: 9

GENOMIC CLONE: W10DE1.0

```
   1  GGGTAGTGACTTTAAAGCTGTGTGACTTTAGTCATTTAACTGCTGAGTCACAGTCTACAG

  61  CTTTGAAAGAGGAGGATTATAAAATCTATCTCATGTTAATGCTGAAGATTAAATAATAGT

 121  GTTTATGTACCCCGCTTATAGGAGAAGAGGGTGTGTGTGTGTGTGTGTGTGTGTGTGT

 181  GTGTATGTGTATGTATACATGTATGTATTCAGTCTTTGCTGAAATTAAAAATCTTTAAC

 241  TTGATAATGGGCAAATATCTTAGTTTTAGATCATGTCCTCTAGAAACCGTATGCTATATA

                                          ----------> 9i-5
 301  ATTATGTACTATAAAGTAATAATGTATACAGTGTAATGGATCATGGGCCATGTGCTTTTC

 361  AAACTAATTGTACATAAAACAAGCATCTATTGAAAATATCTGACAAACTCATCTTTTATT

 421  TTTGATGTGTGTGTGTGTGTGTGTGTGTGTTTTTTTTAACAG|GGATTTGGGGAATTATTTG

 481  AGAAAGCAAAACAAAACAATAACAATAGAAAAACTTCTAATGGTGATGACAGCCTCTTCT

                          ----------> X98
 541  TCAGTAATTTCTCACTTCTTGGTACTCCTGTCCTGAAAGATATTAATTTCAAGATAGAAA

 601  AGAGGACAGTTGTTGGCGGTTGCTGGATCCACTGGAGCAGGCAAG|GTAGTTCTTTTGTTC
        T16L <----------                            |

 661  TTCACTATTAAGAACTTAATTTGGTGTCCATGTCTCTTTTTTTTTCTAGTTTGTAGTGCT

 721  GGAAGGTATTTTTGGAGAAATTCTTACATGAGCATTAGGAGAATGTATGGGTGTAGTGTC

 781  TTGTATAATAGAAATTGTTCCACTGATAATTTACTCTAGTTTTTTATTTCCTCATATTAT

 841  TTTCAGTGGCTTTTTCTTCCACATCTTTATATTTTGCACCACATTCAACACTGTATCTTG
                                  9i-3   <----------

 901  CACATGGCGCGCATTCAATAACTTTATTGAATAAACAAATCATCCATTTTATCCATTCTT

 961  AACCAGAACAGACATTTTTTCAGAGCTGGTCCAGGAAAATCATGACTTACATTTTGCCTT

1021  AGTAACCTCATAAACAAAAAGTCTCCATTTTTGTTGACCTCGAGGGGG
```

EXON 9

EXON: 10

CLONE: T6-20

```
  1 CCTCTTCTCCTTTTTTCATTATTTTTAACAACCCTATTGATATATGATTCACATACCACA
 61 CACTCCTTCATTATTAACACAGGGGACTATGCTATAAACCCTTCAGAAAAATAGACTATG
121 TCTTATTCAATTTGGTATTCCCAGGACCTAGCACAGTGTTCAGAAATTAGTAAATGCTCA
181 TTTTGAGAATGAGAGACCCACAGTACTAAATTTATTACTAGTAATATATTTAAGCTATAT
241 TTTGGCTAAATATACTTTTAAAGGTTGACTTGAGGAAATCAAACACATTTATTTAATTCT
301 ATGGAAAACACATACTTAACGCAAAAAAAAANNNNCCGCTTCTCTGTGAACCTCTATCATA
361 ATACTTGTCACACTGTATTGTAATTGTCTCTTTTACTTTCCCTTGTATCTTTTGTCGATA
    ----------->  10i-5
421 GCAGAGTACCTGAAACAGGAAGTATTTTAAATATTTTGAATCAAATGAGTTAATAGAATC
481 TTTACAAATAAGAATATACACTTCTGCTTAGGATGATAATTGGAGGCAAGTGAATCCTGA
                                                    |
541 GCGTGATTTGATAATGACCTAATAATGATGGGTTTTATTTCCAG|ACTTCACTTCTAATGA
                                                    |
601 TGATTATGGGAGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCAT
          ------------------------> C16B                          EXON 10
661 TCTGTTCTCAGTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTG
                                    C16C  <-------------------
                                                    |
721 TTTCCTATGATGAATATAGATACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAG|GTAA
                    C16D  <-----------------------          |
781 GAAACTATGTGAAAACTTTTTTGATTATGCATATGAACCCTTCACACTACCCAAATTATA
841 TATTTGGCTCCATATTCAATCGGTTAGTCTACATATATTTATGTTTCCTCTATGGGTAAG
                                                  10i-3 <-------
901 CTACTGTGAATGGATCAATTAATAAAACACATGACCTATGCTTTAAGAAGCTTGCAAACA
    ----------
961 CATGAAATAAATGCAATTTATTTTTTAAATAATGGGTTCATTTGATCACAATAAATGCAT
1021 TTTATGAAATGGTGAGAATTTTGTTCACTCATTAGTGAGACAANNNNNGAATGGTATAAG
1081 TGTGAGTGTAAAGAAATTTGCTGATTGCTTTATTAAGAAAAGCTGAAAGTCAAAAGGTAT
1141 CATTTAAAGCTAATAAATAAAGTAATAGAAGCATAAGCAGATTTAACAATACAAAGATAA
1201 ATCTGAAAAAAGATAATACTACTGACTAAAACTGAGTAGAAGGAAAGGAGGTAGCAGAGG
1261 AAGAAAAAGCACTGATTTTATTTATTTATTTATTTATTTATTTATTTATTTAGAGACAGA
1321 GTCTCACTCTGTCACCCAGGCTAGAGTGCAGTGGCGCGATCTCGGCTCACTGCAAGTTCT
1381 GCCTCCTGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACGACGGCAC
1441 CCGCCATCACGTCTGGCT
```

EXON: 11

GENOMIC CLONE: TE24IVE3.5

```
  1  ATATACCCATAAATATACACATATTTAATTTTGGTATTNTTATAATTATTATTTAATGAT

 61  CATTCATGACATTTTAAAAATTACAGGAAAAATTTACATCTAAAATTTCAGCAATGTTGT

                                                              -----
121  TTTGACCAACTAAATAAATTGCATTTGAAATAATGGAGATCAATGTTCAAAATTTCAACT

     ------------------> 11i-5
181  GTGGTTAAAGCAATAGTGTGATATATGATTACATTAGAAGGAAGATGTGCCTTTCAAATT
                                                     |
241  CAGATTGAGCATACTAAAAGTGACTCTCTAATTTTCTATTTTTGGTAATAG|GACATCTCC
                                                     |
                                                     ------------
301  AAGTTTGCAGAGAAAGACAATATAGTTCTTGGAGAAGGTGGAATCACACTGAGTGGAGGT   EXON 11
                                        C16E <----------------------
     →  T16J                  |
361  CAACGAGCAAGAATTTCTTTAGCAAG|GTGAATAACTAATTATTGGTCTAGCAAGCATTTG
                               |
421  CTGTAAATGTCATTCATGTAAAAAAATTACAGACATTTCTCTATTGCTTTATATTCTGTT

481  TCTGGAATTGAAAAAATCCTGGGGTTTTATGGCTAGTGGGTTAAGAATCACATTTAAGAA

541  CTATAAATAATGGTATAGTATCCAGATTTGGTAGAGATTATGGTTACTCAGAATCTGTGC
                                   11i-3  <-----------------------

601  CCGTATCAAGG
```

FIG. 18 (contd)

2007699

EXON: 12

GENOMIC CLONE: TE24IVE2.4

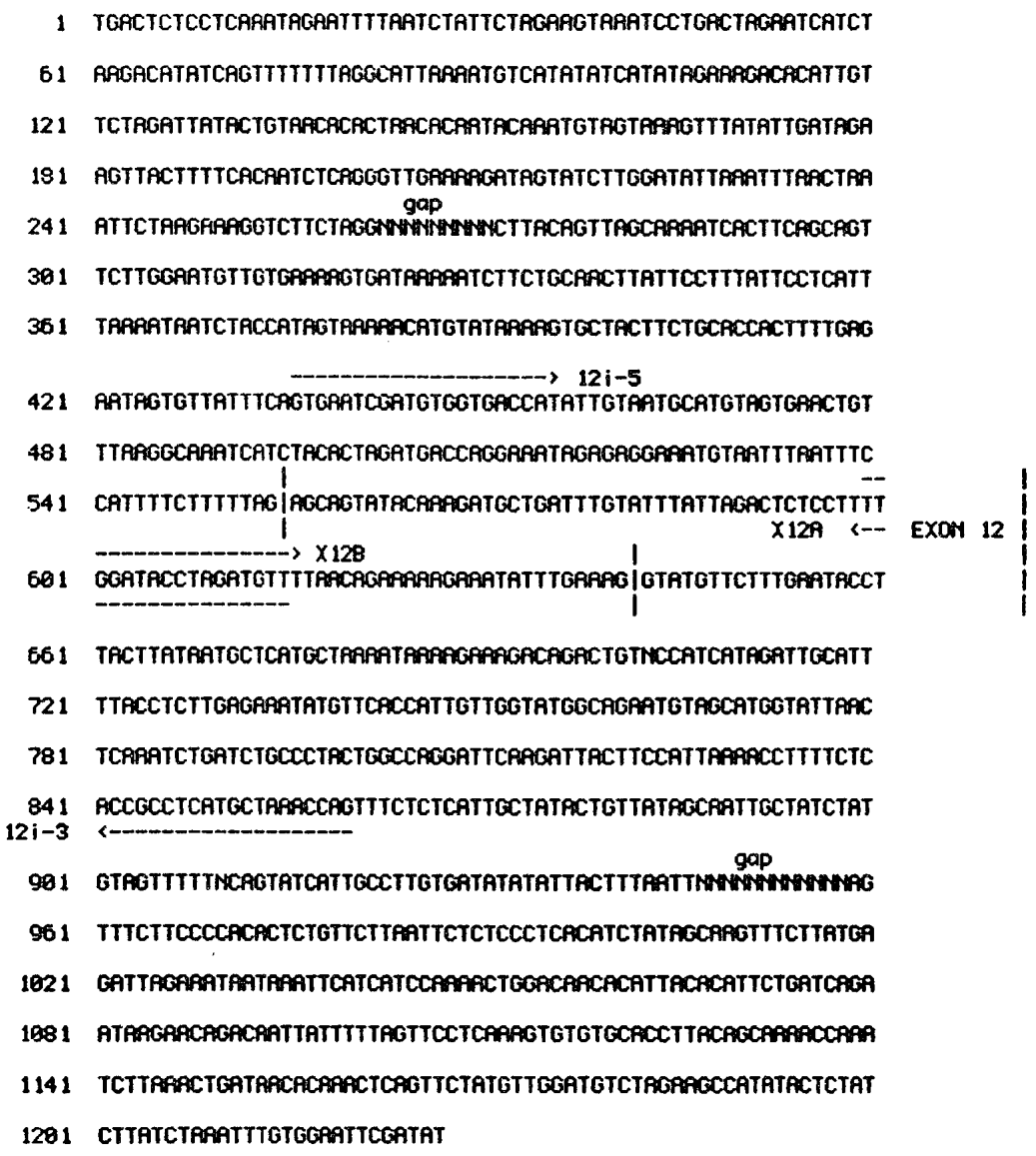


FIG 18(cont'd)

2007699

EXON: 13

GENOMIC CLONE: TE23IIE0.5 and TE23IIE1.8A

```
   1  ACAGAGTACTTATAGAATCATTTAAAATATAATAAAATTGTATGATAGAGATTATNTNCA

                        ----------------------> 13i-5
  61  ATAAAACATTAACAAAATGCTAAAATACGAGACATATTGCAATAAAGTATTTATAAAATT
                                    |
 121  GATATTTATATGTTTTTATATCTTAAAG|CTGTGTCTGTAAACTGATGGCTAACAAAACTA
                                    |
                                                              ---
 181  GGATTTTGGTCACTTCTAAAATGGAACATTTAAAAGAAAGCTGACAAAATATTAATTTTG

      --------------> T16E                CCGGAATT-------------------
 241  CATGAAGGTAGCAGCTATTTTTATGGGACATTTTCAGAACTCCAAAATCTACAGCCAGAC

      -----> T16F
 301  TTTAGCTCAAAACTCATGGGATGTGATTCTTTCGACCAATTTAGTGCAGAAAGAAGAAAT

 361  TCAATCCTAACTGAGACCTTACACCGTTTCTCATTAGAAGGAGATGCTCCTGTCTCCTGG
             T16D <-----------------

 421  ACAGAAACAAAAAAACAATCTTTTAAACAGACTGGAGAGTTTGGGGNAAAAAGGAAGAAT

 481  TCTATNCTCAATCCAATCAACTCTATACGAAAATTTTCCATTGTGCAAAAGACTCCCTTA
                                                                    EXON 13
                                  ------------------>  C1-1L
 541  CAAATGAATGGCATCGAAGAGGATTCTGATGAGCCTTTAGAGAGAAGGCTGTCCTTAGTA
                                                     C1-1M <-----------

 601  CCAGATTCTGAGCAGGGAGAGGCGATACTGCCTCGCATCAGCGTGATCAGCACTGGCCCC
      ------

 661  ACGCTTCAGGCACGAAGGAGGCAGTCTGTCCTGAACCTGATGACACACTCAGTTAACCAA

 721  GGTCAGAACATTCACCGAAAGACAACAGCATCCACACGAAAAGTGTCACTGGCCCCTCAG

 781  GCAAACTTGACTGAACTGGATATATATTCAAGAAGGTTATCTCAAGAAACTGGCTTGGAA

             -----------------> C1-1C |
 841  ATAAGTGAAGAAATTAACGAAGAAGACTTAAAG|GTAGGTATACNTCGCTTGGGGGTATTT
   C1-1K <-----------------------       |

 901  CACCCCACAGAATGCAATTGAGTAGAATGCAATATGTAGCATGTAACAAAATTTACTAAA
                13i-3 <-----------------------

 961  ATCATAGGATTAGGATAAGGTGTATCTTAAAACTCAGAAAGTATGAAGTTCATTAATTAT

1021  ACAAGCAACGTTAAAATGTAAAATAACAAATGATTTTCTTTTTGCAATGGACATATCTCT

1081  TCCCATAANATGGGAAAGGATTTAGTTTTTGGTCCTCTACTAAGCCAGTGATAACTGTGA
                                                          gap
1141  CTATAGTTAGAAAGCATTTGCTTTATTACCATCTTGAACCCTCTGTGNNNNNNNGAGATT

1201  ACAGGCATGCACCACCATGCGAGCTAATTTTTTGTATTTTTTAGTAGAGAAGGGGTTTCA

1261  TCATGTTGACCAGGCTGGTCTTGAATCTTCGACCTTGTGATCCACCCACCTCAGCCTCCC

1321  AAAGTGCTGGTATTACAGGCGTGTGCCACCACGTCCAGCCTGAGCCACTGCGCCCAGCCC

1381  ATCTATATAGTTTAATATCAATCTAAATGAATTTCTCAGTCCTGAGCCTAAAAATTTAGT
```

EXON 13 (cont'd) FIG 18 (cont'd)

1441 TGTAAAGAATGATATCCTTGACTAATAATAGTTTCTATTAATGGATTCGATCTAGTGCTA

1501 GGTGGCATATATTTAGTCCCCACAACTACCCTGGAAGGTATTTAAAATTTTTCACATTTG

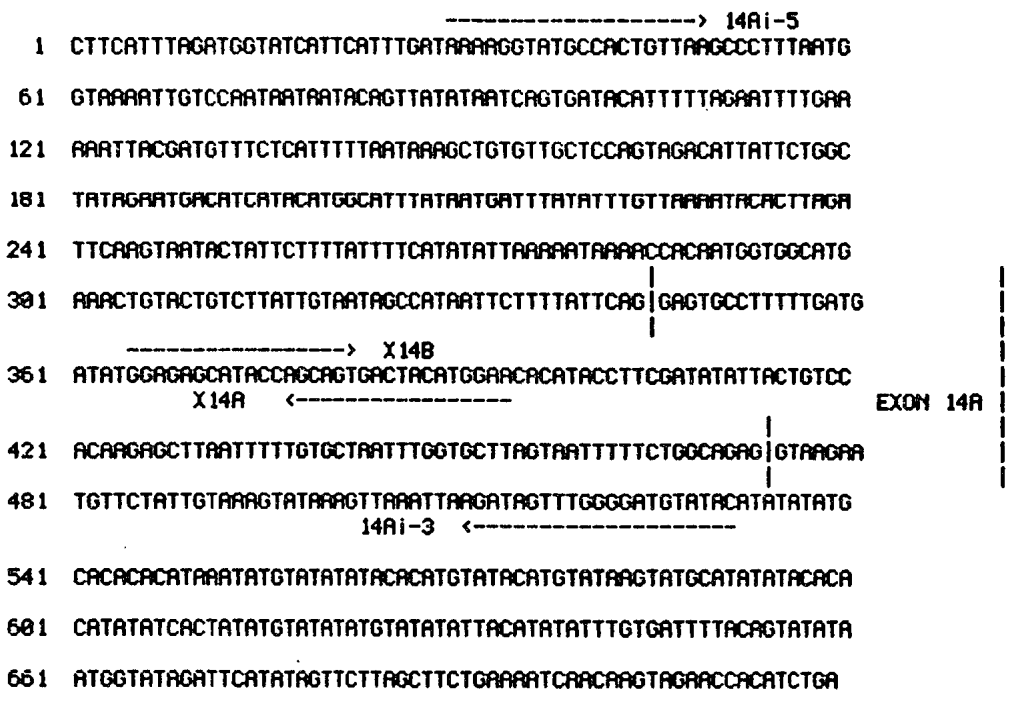
1561 CAGATAAGGAAACTAAAGTTCAGAGTTCGGCAA

FIG 18 (contd)

2007699

EXON: 14A

GENOMIC CLONE: TE23IIE1.8D




```
                                    ----------> 14Ai-5
  1 CTTCATTTAGATGGTATCATTCATTTGATAAAAGGTATGCCACTGTTAAGCCCTTTAATG

 61 GTAAAATTGTCCAATAATAATACAGTTATATAATCAGTGATACATTTTTAGAATTTTGAA

121 AAATTACGATGTTTCTCATTTTTAATAAAGCTGTGTTGCTCCAGTAGACATTATTCTGGC

181 TATAGAATGACATCATACATGGCATTTATAATGATTTATATTTGTTAAAATACACTTAGA

241 TTCAAGTAATACTATTCTTTTATTTTCATATATTAAAAATAAAACCACAATGGTGGCATG
                                                  |
301 AAACTGTACTGTCTTATTGTAATAGCCATAATTCTTTTATTCAG|GAGTGCCTTTTTGATG
                                                  |
    ----------> X14B
361 ATATGGAGAGCATACCAGCAGTGACTACATGGAACACATACCTTCGATATATTACTGTCC
          X14A  <----------                                  |
                                                             |
421 ACAAGAGCTTAATTTTTGTGCTAATTTGGTGCTTAGTAATTTTTCTGGCAGAG|GTAAGAA
                                                             |
481 TGTTCTATTGTAAAGTATAAAGTTAAATTAAGATAGTTTGGGGATGTATACATATATATG
                      14Ai-3  <----------

541 CACACACATAAATATGTATATATACACATGTATACATGTATAAGTATGCATATATACACA

601 CATATATCACTATATGTATATATGTATATATTACATATATTTGTGATTTTACAGTATATA

661 ATGGTATAGATTCATATAGTTCTTAGCTTCTGAAAATCAACAAGTAGAACCACATCTGA
```

FIG. 18 (contd)

2007699

EXON: 14B

GENOMIC CLONE: TE23IIE3.5

```
                                                          ----------
  1  GAATTCCATTAACTTAATGTGGTCTCATCACAAATAATAGTACTTAGAACACCTAGTACA

     -----> 14Bi-5
 61  GCTGCTGGACCCAGGAACACAAAGCAAAGGAAGATGAAATTGTGTGTACCTTGATATTGG

121  TACACACATCAAATGGTGTGATGTGAATTTAGATGTGGGCATGGAGGAATAGGTGAAGA

                      5-10 nt  |------------------> X14BB                    |
181  TGTTAGAAAAAAAATCAANNNNNNNNN|GTGGCTGCTTCTTTGGTTGTGCTGTGGCTCCTT          |
                                |                X14BA <-------------        |
       | 5-10 nt                                                  EXON 14B  |
241  GGAAA|NNNNNNNNNGTGAGTATTCCATGTCCTATTGTGTAGATTGTGTTTTATTTCTGTT          |
     -----|                                                                 |

301  GATTAAATATTGTAATCCACTATGTTTGTATGTATTGTAATCCACTTTGTTTCATTTCTC

361  CCAAGCATTATGGTAGTGGAAAGATAAGGTTTTTTGTTTAAATGATGACCATTAGTTGGG

421  TGAGGAGACACATTCCTGTAGTCCTAGCTCCTCCACAGGCTGACGCAGGAGGATCACTTG
                                                    14Bi-3 <-------

481  AGCCCAGGAGTTCAGGGCTGTAGTGTTGTATCATTGTGAGTAGCCACC
     ------------
```

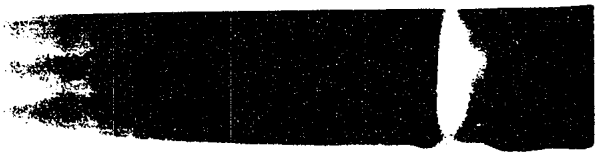


FIG. 18 (contd)

2007699

EXON: 15

GENOMIC CLONE: TE23IIE3.5

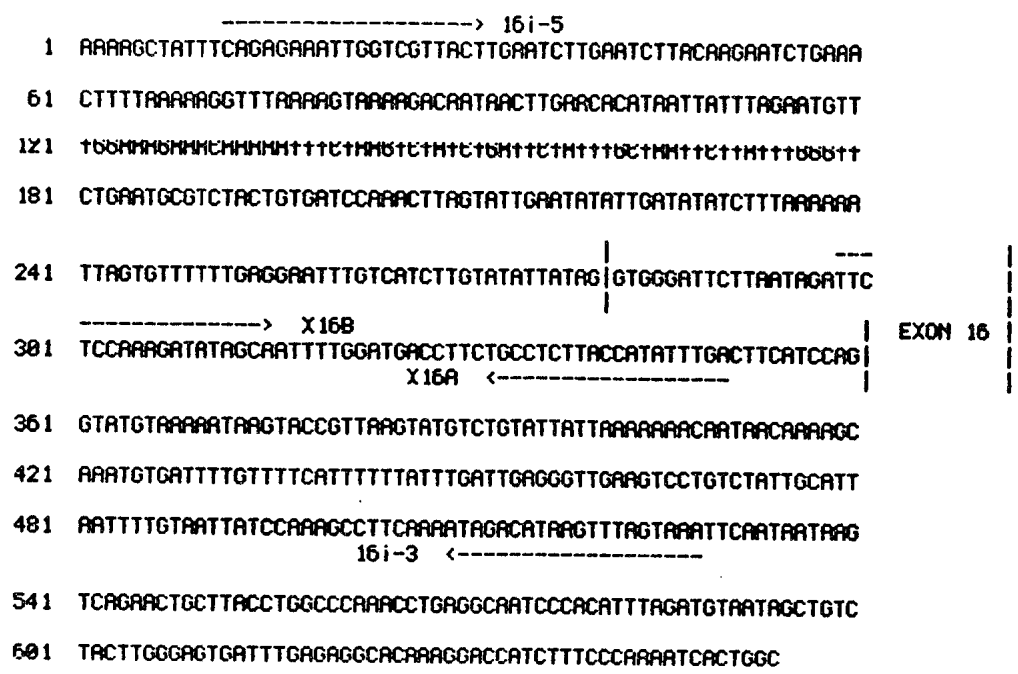
```
  1  TCCTATATCTAAATAAATAAATAAATGAATAAATTGTGAGCATGTGCAGCTCCTGCAGTT

 61  TCTAAAGATATAGTTCTGTTCAGTTTCTGTGAAACACAATAAAAATATTTGAAATAACAT

121  TACATATTTAGGGTTTTCTTCAAATTTTTTAATTTAATAAAGAACAACTCAATCTCTATC
                                                                ---
181  AATAGTGAGAAAACATATCTATTTTATTGCAATAATAGTATGATTTTGAGGTTAAGGGTG
     ---------------> 15i-5
241  CATGCTCTTCTAATGCAAAATATTGTATTTATTTAGACTCAAGTTTAGTTCCATTTACAT

301  GTATTGGAAATTCAGTAAGTAACTTTGGCTGCCAAATAACGATTTCCTATTTGCTTTACA
      |                  ------------------> C1-1N
361  G|CACTCCTCTTCAAGACAAAGGGAATAGTACTCATAGTAGAAATAACAGCTATGCAGTGA
C1-10 |<------------------

421  TTATCACCAGCACCAGTTCGTATTATGTGTTTTATATTTACGTGGGAGTAGCCGACACTT

481  TGCTTGCTATGGGATTCTTCAGAGGTCTACCACTGGTGCATACTCTAATCACAGTGTCGA

541  AAATTTTACACCACAAAATGTTACATTCTGTTCTTCAAGCACCTATGTCAACCCTCAACA
                                     C1-1J <------------------
                    |
601  CGTTGAAAGCAG|GTACTTTACTAGGTCTAAGAAATGAAACTGCTGATCCACCATCAATAG
                    |
661  GGCCTGTGGTTTTGTTGGTTTTCTAATGGCTGTGCTGGCTTTTGCACAGAGGCATGTGCC
                                          15i-3 <-----------------

721  TTTGTT
     --
```

EXON 15

FIG. 18 (cont'd)

EXON: 16

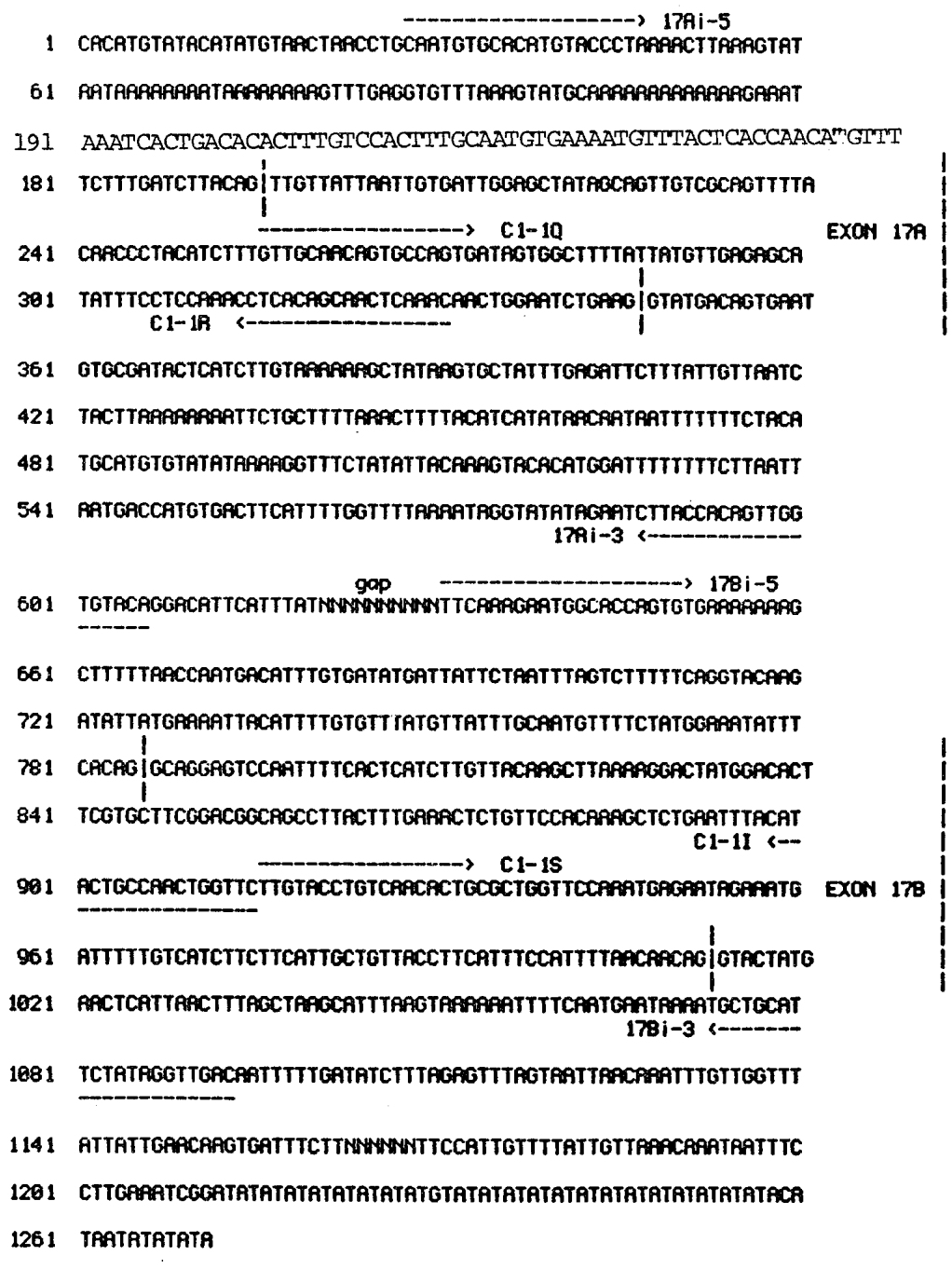
GENOMIC CLONE: TE33IIE1.9




```
                  ------------------> 16i-5
  1 AAAAGCTATTTCAGAGAAATTGGTCGTTACTTGAATCTTGAATCTTACAAGAATCTGAAA

 61 CTTTTAAAAAGGTTTAAAAGTAAAAGACAATAACTTGAACACATAATTATTTAGAATGTT

121 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

181 CTGAATGCGTCTACTGTGATCCAAACTTAGTATTGAATATATTGATATATCTTTAAAAAA

                                                  ---
241 TTAGTGTTTTTTGAGGAATTTGTCATCTTGTATATTATAG|GTGGGATTCTTAATAGATTC

    ---------------> X16B                                          EXON 16
381 TCCAAAGATATAGCAATTTTGGATGACCTTCTGCCTCTTACCATATTTGACTTCATCCAG
                              X16A <------------------

361 GTATGTAAAAATAAGTACCGTTAAGTATGTCTGTATTATTAAAAAAACAATAACAAAAGC

421 AAATGTGATTTTGTTTTCATTTTTTATTTGATTGAGGGTTGAAGTCCTGTCTATTGCATT

481 AATTTTGTAATTATCCAAAGCCTTCAAAATAGACATAAGTTTAGTAAATTCAATAATAAG
                        16i-3 <------------------

541 TCAGAACTGCTTACCTGGCCCAAACCTGAGGCAATCCCACATTTAGATGTAATAGCTGTC

601 TACTTGGGAGTGATTTGAGAGGCACAAAGGACCATCTTTCCCAAAATCACTGGC
```

EXON: 17A and 17B

GENOMIC CLONE: TE33IIE4.5




```
                                  ------------------> 17Ai-5
   1 CACATGTATACATATGTAACTAACCTGCAATGTGCACATGTACCCTAAAACTTAAAGTAT

  61 AATAAAAAAAATAAAAAAAAGTTTGAGGTGTTTAAAGTATGCAAAAAAAAAAAAGAAAT

 191 AAATCACTGACACACTTTGTCCACTTTGCAATGTGAAAATGTTTACTCACCAACATGTTT
                    |                                              |
 181 TCTTTGATCTTACAG|TTGTTATTAATTGTGATTGGAGCTATAGCAGTTGTCGCAGTTTTA  |
                    |                                              |
                    ------------------> C1-1Q                      EXON 17A |
 241 CAACCCTACATCTTTGTTGCAACAGTGCCAGTGATAGTGGCTTTTATTATGTTGAGAGCA  |
                                                   |              |
 301 TATTTCCTCCAAACCTCACAGCAACTCAAACAACTGGAATCTGAAG|GTATGACAGTGAAT  |
       C1-1R  <------------------                  |              |

 361 GTGCGATACTCATCTTGTAAAAAAGCTATAAGTGCTATTTGAGATTCTTTATTGTTAATC

 421 TACTTAAAAAAAATTCTGCTTTTAAACTTTTACATCATATAACAATAATTTTTTTCTACA

 481 TGCATGTGTATATAAAAGGTTTCTATATTACAAAGTACACATGGATTTTTTTTCTTAATT

 541 AATGACCATGTGACTTCATTTTGGTTTTAAAATAGGTATATAGAATCTTACCACAGTTGG
                                          17Ai-3 <-------------

                           gap    --------------------> 17Bi-5
 601 TGTACAGGACATTCATTTATNNNNNNNNNNTTCAAAGAATGGCACCAGTGTGAAAAAAAG
     ------

 661 CTTTTTAACCAATGACATTTGTGATATGATTATTCTAATTTAGTCTTTTTCAGGTACAAG

 721 ATATTATGAAAATTACATTTTGTGTTTATGTTATTTGCAATGTTTTCTATGGAAATATTT
       |                                                          |
 781 CACAG|GCAGGAGTCCAATTTTCACTCATCTTGTTACAAGCTTAAAAGGACTATGGACACT  |
       |                                                          |
 841 TCGTGCTTCGGACGGCAGCCTTACTTTGAAACTCTGTTCCACAAAGCTCTGAATTTACAT  |
                                                   C1-1I <--      |
                    ------------------> C1-1S                     |
 901 ACTGCCAACTGGTTCTTGTACCTGTCAACACTGCGCTGGTTCCAAATGAGAATAGAAATG  EXON 17B |
     ---------------                                              |
                                                   |              |
 961 ATTTTTGTCATCTTCTTCATTGCTGTTACCTTCATTTCCATTTTAACAACAG|GTACTATG  |
                                                   |              |
1021 AACTCATTAACTTTAGCTAAGCATTTAAGTAAAAAATTTTCAATGAATAAAATGCTGCAT
                                          17Bi-3 <-------

1081 TCTATAGGTTGACAATTTTTGATATCTTTAGAGTTTAGTAATTAACAAATTTGTTGGTTT
     -------------

1141 ATTATTGAACAAGTGATTTCTTNNNNNNTTCCATTGTTTTATTGTTAAACAAATAATTTC

1201 CTTGAAATCGGATATATATATATATATATGTATATATATATATATATATATATATATACA

1261 TAATATATATA
```

GENOMIC CLONE: TE33IIE2.4

EXON: 18

```
  1  TTATTACTTATAGAATAATAGTAGAAGAGACAAATATGGTACCTACCCATTACCAACAAC

 61  ACCTCCAATACCAGTAACATTTTTTAAAAAGGGCAACACTTTCCTAATATTCAATCGCTC

                                                                 ------
121  TTTGATTTAAAATCCTGGTTGAATACTTACTATATGCAGAGCATTATTCTATTAGTAGAT

     ---------------> 18i-5
181  GCTGTGATGAACTGAGATTTAAAAATTGTTAAAATTAGCATAAAATTGAAATGTAAATTT

241  AATGTGATATGTGCCCTAGGAGAAGTGTGAATAAAGTCGTTCACAGAAGAGAGAAATAAC

                                          |                  -----
301  ATGAGGTTCATTTACGTCTTTTGTGCTTCTATAG|GAGAAGGAGAAGGAAGAGTTGGTATT
                                          |
     ---------------> X18B
361  ATCCTGACTTTAGCCATGAATATCATGAGTACATTGCAGTGGGCTGTAAACTCCAGCAT
 X18A <-----------------                                          EXON 18
                      |
421  AGATGTGGATAGCTTG|GTAAGTCTTATCATCTTTTTAACTTTTATGAAAAAAATTCAGAC
                      |
481  AAGTAACAAAGTATGAGTAATAGCATGAGGAAGAATATATACCGTATATTGAGCTTAAGA

541  AATAAAACATTACAGATAAATTGAGGGTCACTGTGTATCTGTCATTAAATCCTTATCTCT

601  TCTTTCCTTCTCATAGATAGCCACTATGAAGATCTAATACAGCAGTGAGCATTCTTTCAC
     18i-3 <--------------------

661  CTGTTTCCTTATTCAGGATTTTCTAGGAGAAATACCTAGGGGTTGTATTGCTGGGTCATA

721  GGATTCACCCATGCTTAAC
```

EXON: 19

GENOMIC CLONE: TE24IIE1.0 and TE24IIE2.8

```
   1  TCTTTTTATTCATTTATATTTATTTACTCATTAGTATATTCATTCCTTTATTCATGTATT

  61  GTTCAAATATATATTGGGTACTTATTATATGCCAAGTTGTTTTTAAAATCACATTCCAAA

 121  TTCCCGTAAGTCATAATTATTCAGAGATGTATGTTTTTTTTTAAAAAAAATTGAACACCT

 181  TTAAAAAAATTTATCAAGTCCTTTTATTTCTGTATGCATTAAAGATAAACTTTACTAAATG
                                                   gap
 241  TTACATGAATAGATTTATAAAGCAGATAAAATATTTATTTNNNNNNNNNNTTCTCTTCAG

                                                     ----------------
 301  TTAACTTTAATTATATCCAATTATTTCCTGTTAGTTCATTGAAAAGCCCGACAAATAACC

      ------> 19i-5
 361  AAGTGACAAATAGCAAGAGTTGCATTTTACAAGTTATTTTTTAGGAAGCATCAAACTAAT
                                                     |                 |
 421  TGTGAAATTGTCTGCCATTCTTAAAACAAAATGTTGTTATTTTTATTTCAG|ATGCGATCT    |
                                                     |                 |
        ------------------> C1-1H                                --    |
 481  GTGAGCCGAGTCTTTAAGTTCATTGACATGCCAACAGAAGGTAAACCTACCAAGTCAACC    |
                                           X19A  <-----------------    |
      ---------------> C1-1D                                   EXON 19 |
 541  AAACCATACAAGAATGGCCAACTCTCGAAAGTTATGATTATTGAGAATTCACACGTGAAG    |
                                                                       |
 601  AAAGATGACATCTGGCCCTCAGGGGGCAAATGACTGTCAAAGATCTCACAGCAAAATACA    |
                                                               |       |
 661  CAGAAGGTGGAAATGCCATATTAGAGAACATTTCCTTCTCAATAAGTCCTGGCCAGAGG|G    |
                                                               |       |
 721  TGAGATTTGAACACTGCTTGCTTTGTTAGACTGTGTTCAGTAAGTGAATCCCAGTAGCCT
                                                       19i-3 <----

 781  GAAGCAATGTGTTAGCAGAATCTATTTGTAACATTATTATTGTACAGTAGAATCAATATT
      ----------------

 841  AAACACACATGTTTTATTATATGGAGTCATTATTTTTAATATGAAATTTAATTTGCAGAG
                         gap
 901  TCTGAACTATATATNNNNNNNNNNNNGAAACTTTCTGTCCCAGGACATGATAATCTAAAA

 961  GGGAAAACGTAAGATCCACTGAAACCTGAGGCAGATTTATTGTGGCAATAACAAAGCTTA

1021  AGTTTCACAGACCTTCATTTGCCTGAGCCAACTTTGAAGGCCATGTATCTAATTTTGTTT

1081  TTATAATTCTATAATCTTTATTCTTGAAAAGAGCCCTCCCTCNAAATTTACAAGCTTTGG

1141  GCCCCNAAAATCCTTGAAATGCCCTTGAATAAGAGATATCCAGGTAAATGCTATGG
```

EXON: 20

GENOMIC CLONE: TE24IE1.8

```
  1  AAAGGTCAGTGATAAAGAAGTCTGCATCAGGGTCCAATTCCTTATGGCCAGTTTCTCTAT

                                                  --------------------> 20i-5
 61  TCTGTTCCAAGGTTGTTTGTCTCCATATATCAACATTGGTCAGGATTGAAAGTGTGCAAC

121  AAGGTTTGAATGAATAAGTGAAAATCTTCCACTGGTGACAGGATAAAATATTCCAATGGT

181  TTTTATTGAAGTACAATACTGAATTATGTTTATGGCATGGTACCTATATGTCACAGAAGT

                             |                                         |
241  GATCCCATCACTTTTACCTTATAG|GTGGGCCTCTTGGGAAGAACTGGATCAGGGAAGAGT     |
                             |                                         |
     ------------> C1-1T                                       EXON 20 |
301  ACTTTGTTATCAGCTTTTTTGAGACTACTGAACACTGAAGGAGAAATCCAGATCGATGGT      |
                                                             |         |
361  GTGTCTTGGGATTCAATAACTTTGCAACAGTGGAGGAAAGCCTTTGGAGTGATACCACAG|GT   |
          C1-1U <-------------------                          |         |

421  GAGCAAAAGGACTTAGCCAGAAAAAAGGCAACTAATTTATATTTTTTACTGCTATTTG

481  ATACTTGTACGTCAAGAAATTCATATTACTCTGCAAAATATATTTGTTATGCGTTGCTGT

541  CTTTTTTTTCCAGTGCAGTTTTCTCATAGGCAGAAAAGATGTCTCTAAAAGTTTGGGAATT
       20i-3 <--------------------

601  CCC
```

FIG 18(contd)

EXON: 21

GENOMIC CLONE: TE261IE8.5

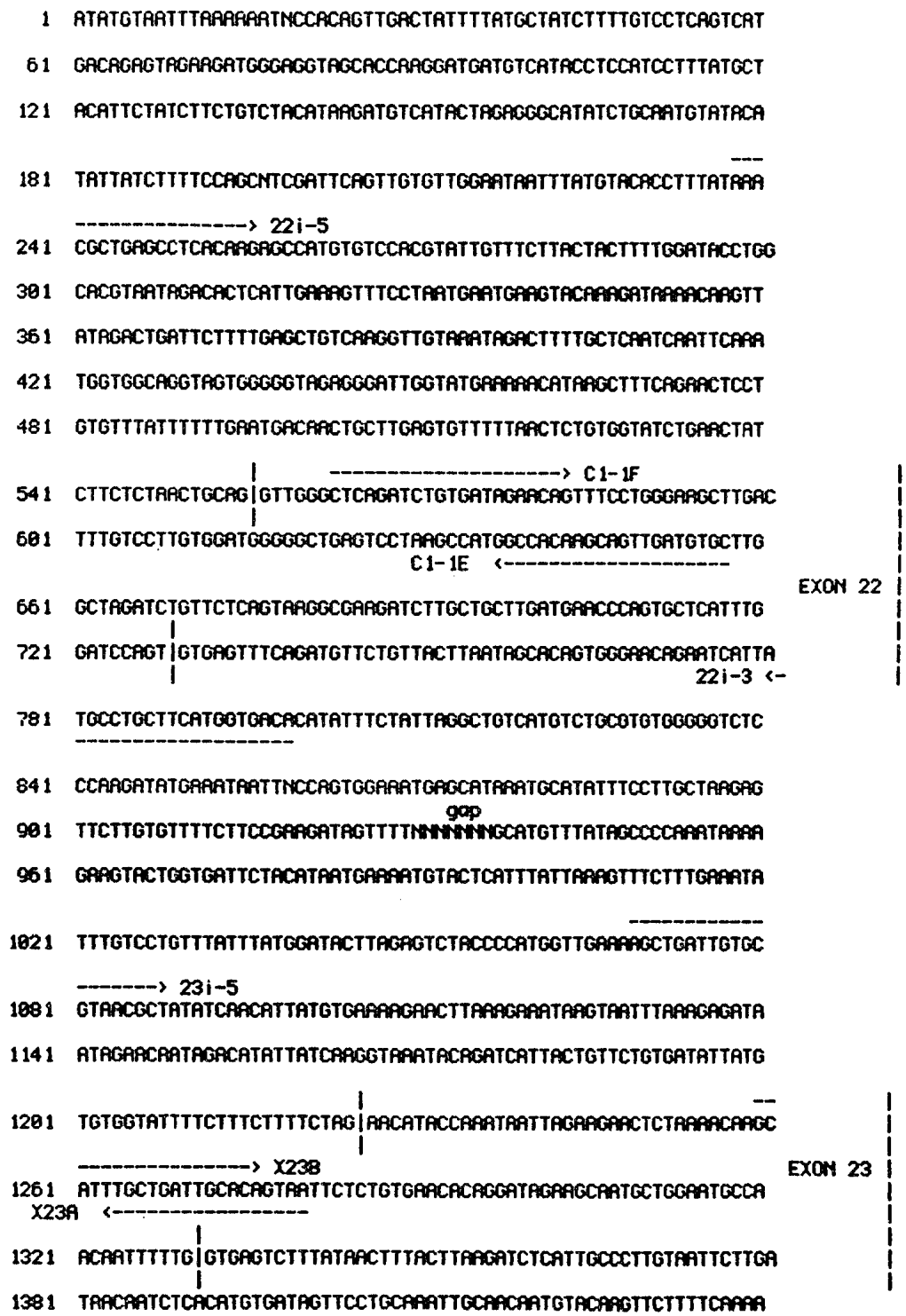
```
  1 TTTTTTAATATTCTACAATTAACAATTATCTCAATTTCTTTATTCTAAAGACATTGGATT

      ------------------> 21i-5
 61 AGAAAAATGTTCACAAGGACTCCAAATATTGCTGTAGTATTTGTTTCTTAAAAGAATGAT

121 ACAAAGCAGACATGATAAAATATTAAAATTTGAGAGAACTTGATGGTAAGTACATGGGTG

181 TTTCTTATTTTAAAATAATTTTTCTACTTGAAATATTTTACAATACAATAAGGGAAAAAT
                                                    |
241 AAAAAGTTATTTAAGTTATTCATACTTTCTTCTTCTTTTCTTTTTTGCTATAG|AAAGTAT
                                                    |
                                  ------------------> X21B
301 TTATTTTTTCTGGAACATTTAGAAAAAACTTGGATCCCTATGAACAGTGGAGTGATCAAG
                                       X21A  <-----------------
                                |
361 AAATATGGAAAGTTGCAGATGAG|GTAAGGCTGCTAACTGAAATGATTTTGAAAGGGGTAA
                                |
421 CTCATACCAACACAAATGGCTGATATAGCTGACATCATTCTACACACTTTGAGAGCATGT

481 ATGTGTGTGCACAACTTTAAAATGGAGTACCCTAACATACCTGGAGCAACAGGTACTTTG
                                       21i-3  <------------------

541 ACTGGACCTACCCCCTAACTGAAATGATTTTGAAAGAGGTAACTCATACCAACACAAATG

601 GTTGATATGGCTAAGATCATTCTACACACTTTGTGTGCATGTATTTCTGTGCACAACTTC

661 AAAATGGAGTACCTAAAATACCTGGCGCGACAAGTA
```

EXON 21

Fig. 18 (cont'd)

EXON: 22 and 23

GENOMIC CLONE: TE22E2.2



Exon 23 (contd)

FIG 18(contd)

2007699

```
1441  ATATGTATCATACAGCCATCCAGCTTTACTCAAAATAGCTGCACAAGTTTTTCACTTTGA
        23i-3  <--------------------

1501  TCTGAGCCATGTGGTGAGGTTGAAATATAGTAAATCTAAAATGGCAGCATATTACTAAGT

1561  TATGTTTATAAATAGGATATATATCTTTTGAGCCCTTTATTTGGGACCAAGTCATACAAA

1621  ATACTCTACAGTTTAAGATTTTAAAAAAGGTCCCTGTGATTCTTTCAATAACTAAATGTC

1681  CCATGATGTGGTCTGGACAGGCCTAGTTGTCTTACAGTCTGATTTATGGTATTAATGACA

1741  AAGTTGAGAGGCACATTTCATTTTTCTAGCCATGATTTGGGTTCAGGTAGTACCTTTCTC

1801  AACCACCTTCTCACTGTTCTTAAAAAAACTGTCACATGGCCAGGCACAGTGGCTTACATC

1861  TGTAATCCCAATACTTTGGGAGGCTGAAATGGGGGGATTACTTGAGGCCAGGAATTC
```

FIG 18 (contd)

EXON: 24

GENOMIC CLONE: TE27E2.3

```
   1  GAGGCTCAACTCTTTATGGTGCTGTAGACTTACNCTCATTTTCTAGGTAATTTATAAGGG
  61  ACCTAATATTTTGTTTTCAAAGCAACTTCAGTTCTACTAAACCTCCCTGAAGAATCTTCC
 121  AGCTGCTGAGTAGAAAATCACAACTAATTTCACAGATGGTAGAACCTCCTTAGAGCAAAA
 181  GGACACAGCAGTTAAATGTGACATACCTGATTGTTCAAATGCAAGGCTCTGGACATTGCA
 241  TTCTTTGACTTTTATTTTCCTTTGAGCCTGTGCCAGTTTCTGTCCCTGCTCTGGTCTGAC
 301  CTGCCTTCTGTCCCAGATCTCACTAACAGCCATTTCCTAGGTCATAGAAGAGAACAAAGC
        gap                                  ----------> 24i-5
 361  NNNNNNNAGATGGTAGACCTCCTAGAGCAAAAGGACACAGCAGTTAAATGTGACATACCT
 421  GATTGTTCAAAATGCAAGGCTCTGGACATTGCATTCTTTGACTTTTTATTTTTCCTTTGA
 481  GCCTGTGCCAGTTTCTGTCCCTGCTCTGGTCTGACCTGCCTTCTGTCCCAGATCTCACTA
                      |                          ----------
 541  ACAGCCATTTCCCTAG|GTCATAGAAGAGAACAAAGAGCGGCAGTACGATTCCATCCAGAA
                      |
      -----> C1-1G
 601  ACTGCTGAACGAGAGGAGCCTCTTCCGCAAGCCATCAGCCCCTCCGACAGGTGAAGCTCT
 661  TTCCCCACCGGAACTCAAGCAAGTGCAAGTCTAAGCCCCAGATTGCTGCTCTGAAAGAGG
                        C1-1A  <----------------------
 721  AGACAGAAGAAGAGGTGCAAGATACAAGGCTTTAGAGAGCAGCATAAATGTTGACATGGG
                                          ---
 781  ACATTTGCTCATGGAATTGGAGCTCGTGGGACAGTCACCTCATGGAATTGGAGCTCGTGG
                              ----------> C1-1B
 841  AACAGTTACCTCTGCCTCAGAAAACAAGGATGAATTAAGTTTTTTTTAAAAAAGAAACAT
 901  TTGGTAAGGGGAATTGAGGACACTGATATGGGTCTTGATAAATGGCTTCCTGGCAATAGT
                                        24i3  <----------
 961  CAAATTGTGTGAAAGGTACTTCAAATCCTTGAAGATTTACCACTTGTGTTTTGCAAGCCA
1021  GATTTTCCTGAAAACCCTTGCCATGTGCTAGTAATTGGAAAGGCAGCTCTAAATGTCAAT
                                                 T16-4B
                                          ---------->
1081  CAGCCTAGTTGATCAGCTTATTGTCTAGTGAAACTCGTTAATTTGTAGTGTTGGAGAAGA
1141  ACTGAAATCATACTTCTTAGGGTTATGATTAAGTAATGATAACTGGAACTCAGCGGTTTA
1201  TATAAGCTTGTATTCCTTTTTCTCTCCTCTCCCCATGATGTTTAGAAACACAACTATATT
1261  GTTTGCTAAGCATTCCAACTATCTCATTTCCAAGCAAGTATTAGAATACCACAGGAACCA
1321  CAAGACTGCACATCAAAATATGCCCCATTCAACATCTAGTGAGCAGTCAGGAAAGAGAAC
1381  TTCCAGATCCTGGAAATCAGGGTTAGTATTGTCCAGGTCTACCAAAAATCTCAATATTTC
1441  AGATAATCACAATACATCCCTTACCTGGGAAAGGGCTGTTATAATCTTTCACAGGGGACA
```

EXON 24

EXON 24 (cont'd) 2007699 FIG. 18 (cont'd)

gap

1501 GGATGGTTCCCTTNNNNNNNNNNNNTCGAGGTTGTTGCCCACAGCTGTATGATTCCC

1561 AGCCAGACACAGCCTCTTAGATGCAGTTCTGAAGAAGATGGTACCACCAGTCTGACTGTT

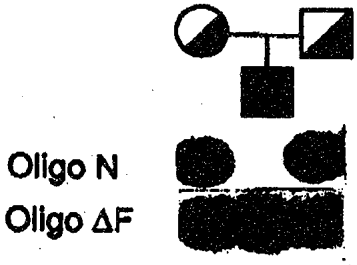
1621 TCCATCAAGGGTACACTGCCTTCTCAACTCCAAACTGACTCTTAAGAAGACTGCATTATA

1681 TTTATTACTGTAAGAAAATATCACTTGTCAATAAAATCCATACATTTGTGTG|AAACTTTG

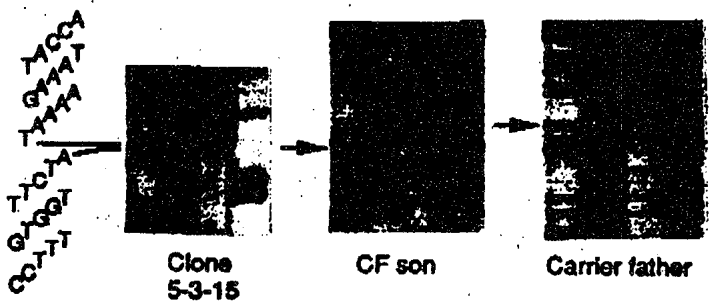
1741 TTGTTTTCAGATGCGTTCACTTGTCATGTTTCATCAGTCTCTCACTCCAATTTCTAAGCT

1801 TCATGGAACATGAAACACGAATCTGTCTTTTAGATATAGCCTC

A.



B.



C.

	501									510
	Thr	*Ile*	*Lys*	*Glu*	*Arg*	*Ile*	*Ile*	*Phe*	*Gly*	*Val* *Ser*
Normal	ACC	ATT	AAA	GAA	AAT	ATC	ATC	TTT	GGT	GTT TCC
	Thr	*Ile*	*Lys*	*Glu*	*Arg*	*Ile*		*Phe*	*Gly*	*Val* *Ser*
ΔI507	ACC	ATT	AAA	GAA	AAT	ATC		TTT	GGT	GTT TCC
	Thr	*Ile*	*Lys*	*Glu*	*Arg*	*Ile*	*Ile*		*Gly*	*Val* *Ser*
ΔF508	ACC	ATT	AAA	GAA	AAT	ATC	AT		TGG	TGT TTC C

FIG #19

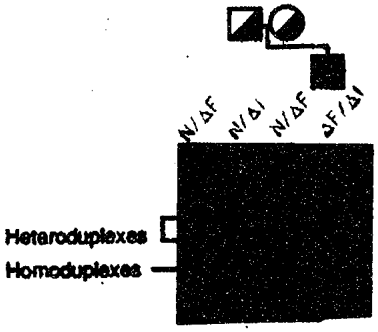


FIG 19 & 20