Identification of the Cystic Fibrosis Gene: Chromosome Walking and Jumping

JOHANNA M. ROMMENS, MICHAEL C. IANNUZZI, BAT-SHEVA KEREM, MITCHELL L. DRUMM, GEORG MELMER, MICHAEL DEAN, RICHARD ROZMAHEL, JEFFERY L. COLE, DARA KENNEDY, NORIKO HIDAKA, MARTHA ZSIGA, MANUEL BUCHWALD, JOHN R. RIORDAN, LAP-CHEE TSUI, FRANCIS S. COLLINS

An understanding of the basic defect in the inherited disorder cystic fibrosis requires cloning of the cystic fibrosis gene and definition of its protein product. In the absence of direct functional information, chromosomal map position is a guide for locating the gene. Chromosome walking and jumping and complementary DNA hybridization were used to isolate DNA sequences, encompassing more than 500,000 base pairs, from the cystic fibrosis region on the long arm of human chromosome 7. Several transcribed sequences and conserved segments were identified in this cloned region. One of these corresponds to the cystic fibrosis gene and spans approximately 250,000 base pairs of genomic DNA.

CYSTIC FIBROSIS (CF) IS REGARDED AS THE MOST COMMON severe autosomal recessive disorder in the Caucasian population, with a disease frequency of 1 in 2000 live births and a calculated carrier frequency of about 5 percent (*1*). The major clinical symptoms and signs include chronic pulmonary disease, pancreatic exocrine insufficiency, and an increase in the concentration of sweat electrolytes. Although recent advances have been made in the analysis of ion transport across the apical membrane in CF epithelium (*2*), it is not clear that the abnormal regulation of chloride channels represents the primary lesion in the disease. Apart from these electrophysiological studies, an alternative approach has 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 (*3, 4*).

Linkage analysis based on a large number of polymorphic DNA markers has unambiguously assigned the CF locus (*CF*) to the long arm of chromosome 7, band q31 (*3–5*). The identification of closely linked flanking markers, *MET* and *D7S8*, has made it possible to use various novel gene cloning strategies to pinpoint the CF gene. These methods include chromosome jumping from the flanking markers (*6*), cloning of DNA fragments from a defined physical region with the use of pulsed field gel electrophoresis (*7*), a combination of somatic cell hybrid and molecular cloning techniques designed to isolate DNA fragments from undermethylated CpG islands near *CF* (*8*), chromosome microdissection and cloning (*9*), and saturation cloning of a large number of DNA markers from the 7q31 region (*10*).

The saturation mapping approach, by systematic examination of DNA markers from a flow-sorted genomic DNA library specific to chromosome 7, allowed the identification of two additional DNA markers (*D7S122* and *D7S340*) closely linked to *CF* (*10*). Genetic and physical mapping studies indicated the order of the four markers to be *MET-D7S340-D7S122-D7S8*, with distance intervals of 500, 10, and 980 kilobase (kb) pairs, respectively (*11*). This distance estimate for the *MET-D7S8* interval agrees well with the data from previous genetic (*4, 5, 10*) and physical mapping (*12*) studies.

Chromosome walking and jumping. As the genetic data indicated that *D7S122* and *D7S340* were probably in close proximity to *CF*, and the physical map of the region was well defined, the next logical step was to clone a large amount of the surrounding DNA and search for candidate gene sequences. In addition to conventional chromosome walking methods, the chromosome jumping technique was used to accelerate the process, as a new bidirectional walk could be initiated from the end point of each jump. Furthermore, sequential walks halted by "unclonable" regions often encountered in the mammalian genome could be circumvented by chromosome jumping (see below). Parallel chromosome jumping experiments were also performed from *D7S8* toward *D7S122* and *D7S340* to narrow the region of interest (*13*).

Ten genomic libraries were constructed during the course of our experiments (*14*). The contiguous chromosome region covered by chromosome walking and jumping was about 280 kb (Fig. 1). This effort involved the isolation and characterization of 49 recombinant phage and cosmid clones and nine jumping clones. The ability to bias the direction of jumps by careful choice of probes (*6*) proved to be a useful feature of the strategy.

A restriction map of the cloned human DNA segments derived from chromosome walking and jumping was constructed (Fig. 1). As the two independently isolated DNA markers, *D7S122* (pH131)

J. M. Rommens, B. Kerem, G. Melmer, R. Rozmahel, D. Kennedy, M. Zsiga, M. Buchwald, and L.-C. Tsui are in the Department of Genetics, Research Institute, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. M. C. Iannuzzi, M. L. Drumm, J. L. Cole, N. Hidaka, and F. S. Collins are at the Howard Hughes Medical Institute and Departments of Internal Medicine and Human Genetics, University of Michigan, Ann Arbor, MI 48109. M. Buchwald and L.-C. Tsui are also faculty members of the Departments of Medical Genetics and Medical Biophysics, University of Toronto, Toronto M5S 1A8, Canada. M. Dean is at Program Resources, Inc., Frederick Cancer Research Facility, Frederick, MD 21701. J. R. Riordan is in the Department of Biochemistry, Research Institute, The Hospital for Sick Children, Toronto, and the Departments of Biochemistry and Clinical Biochemistry, University of Toronto, Toronto M5S 1A8, Canada.

and *D7S340* (TM58), were only ~10 kb apart (Fig. 1), 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 (Fig. 1), and with reference to the long-range physical map (*11, 12*). The pulsed field mapping data also revealed that the Not I site identified in our study (see Fig. 1, position 113 kb) corresponded to the one previously found associated with the *int*-related protein (IRP) locus (*IRP*) (*8*). As subsequent genetic studies showed that *CF* was most likely located between *IRP* and *D7S8* (*15, 16*), our walking and jumping effort, as described below, was directed exclusively toward cloning of this interval.

Three regions in the 280-kb segment were not readily recoverable in the amplified genomic libraries initially used (*14*). These less clonable regions were located near the DNA segments H2.3A and X.6 and just beyond cosmid cW44, at positions 75 to 100 kb, 205 to 225 kb, and 275 to 285 kb, respectively (Fig. 1). The recombinant clones near H2.3A were unstable, and underwent dramatic rearrangements after only a few passages of bacterial culture. To fill in the resulting gaps, we constructed primary walking libraries with special host-vector systems that allow propagation of unstable sequences (*14, 17*). Although the region near cosmid cW44 has not yet been recovered, the region near X.6 was successfully rescued with these libraries. Mammalian DNA segments with unusual secondary structure or repetitive elements are unstable in bacterial cells (*17*), but the nature of the less clonable sequences encountered in our study remain to be determined. It is of interest that potential recombination hot spots have been identified near H2.3A and the end of cW44 (*16*).

Alignment of cloned regions with genomic DNA. Together with the genomic DNA sequences isolated with the overlapping cDNA clones described by Riordan *et al.* (*18*), the entire region cloned in our study extended >500 kb. To ensure that the DNA segments isolated by the chromosome walking and jumping procedures were colinear with the genomic sequence, we examined each segment by (i) hybridization analysis with human-rodent somatic hybrid cell lines to confirm localization on chromosome 7 (*10, 19*); (ii) pulsed field gel electrophoresis; and (iii) 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 were used as probes in each of these hybridization analyses (*20, 21*).

Although most 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. Because this cloning strategy requires the ligation of the two ends of a large genomic segment (*6*), tandem ligations of unrelated molecules can give rise to anomalous jumping clones. One of the jump clones was not located on chromosome 7 and was discarded.

Further confirmation of the overall physical map of the overlapping clones was obtained by long-range restriction mapping with the use of pulsed field gel electrophoresis (*11, 12*). A preliminary long-range map of this region describing *D7S122* and *D7S340* was previously published (*11*). The more recent walk-jump clones and cDNA clones corresponding to the CF locus generated a more extensive pulsed field restriction map, which was in complete concordance with that derived from chromosome walking (Fig. 2). Many of the recognition sites for rare-cutting restriction enzymes in this region, such as Not I and Bss HII, were resistant to digestion in the human-rodent cell hybrid (*19*), presumably due to DNA methylation. These sites were less resistant to digestion, however, in other human cell lines (Fig. 2).

The result of the long-range restriction mapping study showed that the entire CF locus was contained on a 380-kb Sal I fragment (Fig. 2). 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 locus was about ~250 kb.

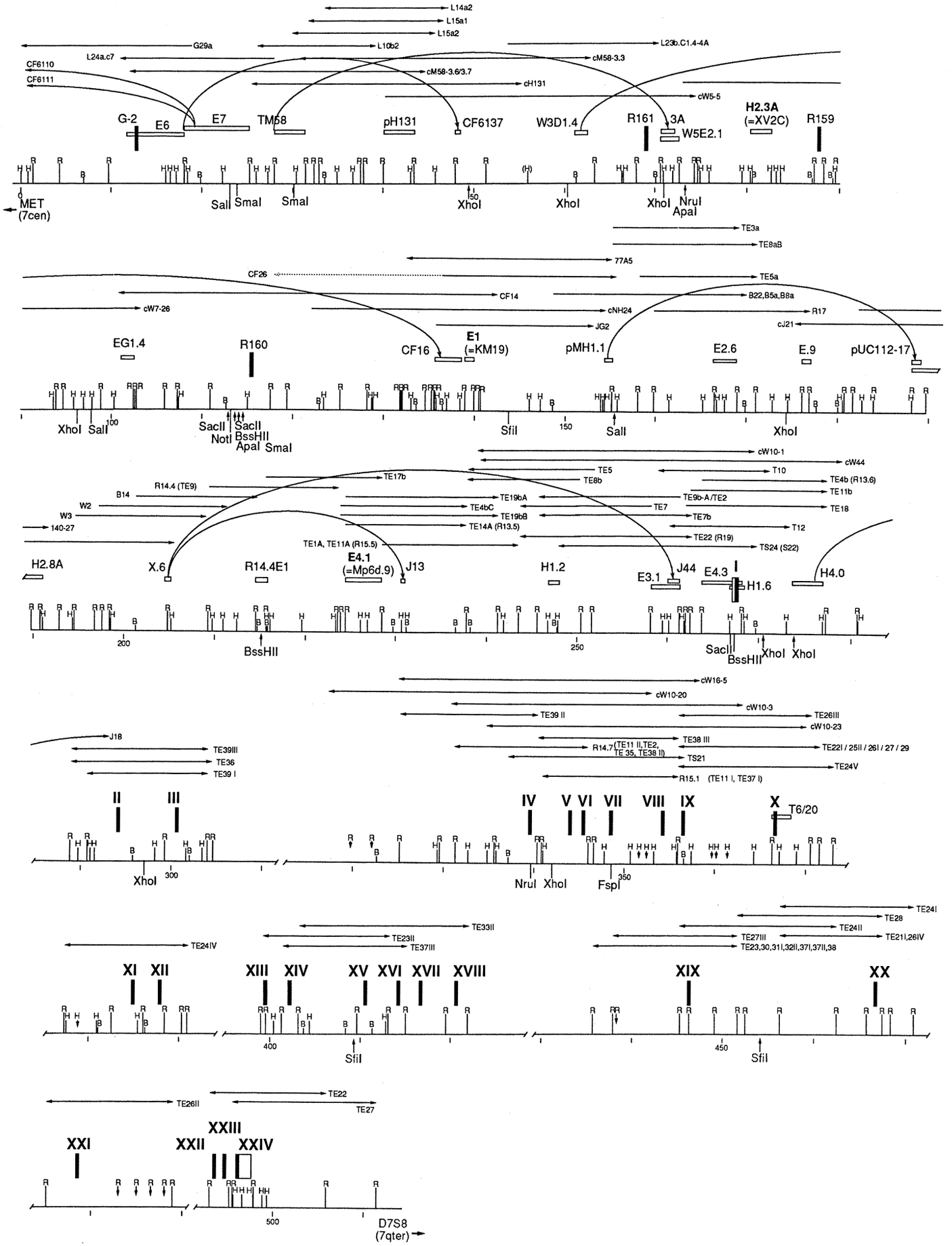
The most informative restriction enzyme that served to align the map of the cloned DNA fragments and the long-range restriction map was Xho I; all of the nine Xho I sites identified with the recombinant DNA clones appeared to be susceptible to at least partial cleavage in genomic DNA (compare maps in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

Search for gene sequences. A positive result based on one or more of the following criteria suggested that a cloned DNA segment might contain candidate gene sequences: (i) detection of cross-hybridizing sequences in other species (as many genes show evolutionary conservation) (*22*), (ii) identification of CpG islands, which often mark the 5′ end of vertebrate genes (*23*), (iii) examination of possible mRNA transcripts in tissues affected in CF patients, (iv) isolation of corresponding cDNA sequences, and (v) identification of open reading frames by direct sequencing of cloned DNA segments. The ongoing genetic analysis (*15, 16, 24*) strongly influenced how extensively a region was examined for possible gene sequences. All the methods have potential inherent limitations; because the DNA hybridization method for detecting conserved DNA sequences across species was relatively straightforward and has been successful in detection of other loci (*22*), it was generally used as the first step.

In some of the cross-species hybridization experiments, it was possible to use entire phage or cosmid clones containing human sequences as probes without removal of the repetitive elements because these sequences are in general not shared between distantly related species. Distinct cross-hybridization signals were detected with probes from four regions in the 280-kb span (Fig. 3).

Conserved region 1 was defined by the DNA segment G-2 (position 13 in Fig. 1); region 2 was detected by the cosmid CF14 (positions 100 to 142); region 3 was defined by the probe R14.4E1

Fig. 1 (facing page). Restriction map of the region of chromosome 7 containing *CF*. The map proceeds from left to right in six tiers with the direction of ends toward 7cen and 7qter as indicated. The restriction map for the enzymes Eco RI (R), Hind III (H), and Bam HI (B) is shown above the solid line, spanning the entire cloned region. Restriction sites indicated with arrows rather than vertical lines indicate sites that have not been unequivocally positioned. Additional restriction sites for other enzymes are shown below the line. The scale is in kilobases. Gaps in the cloned region are indicated by a gap in the solid line (//). These occur only in the portion detected by cDNA clones of the CF transcript and, on the basis of pulsed field mapping of the region (Fig. 2), are unlikely to be large. Chromosome jumps are indicated by the arcs. Walking clones are indicated by horizontal arrows above the map, with the direction of the arrow indicating the walking progress obtained with each clone. Cosmid clones begin with C or 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 in this and accompanying papers (*16, 18*). Three of the probes represent independent subcloning of fragments previously identified to detect polymorphisms in this region: H2.3A corresponds to probe XV2C (*8*), probe E1 corresponds to KM19 (*8*), and E4.1 corresponds to Mp6d.9 (*37*). G-2 is a subfragment of E6 that detects a transcribed sequence (see Fig. 4); R161, R159, and R160 are synthetic oligonucleotides constructed from the IRP locus sequence (*26*), indicating the location of this transcript on the genomic map.


L14a2
L15a1
L15a2
G29a
L10b2
L23b.C1.4-4A
L24a.c7
cM58-3.3
CF6110
cM58-3.6/3.7
CF6111
cH131
cW5-5
G-2
E6
E7
TM58
pH131
CF6137
W3D1.4
R161
3A
W5E2.1
H2.3A
(=XV2C)
R159
MET
(7cen)
SalI
SmaI
SmaI
XhoI
XhoI
XhoI
NruI
ApaI
TE3a
TE8aB
77A5
CF26
TE5a
CF14
B22,B5a,B8a
cW7-26
cNH24
R17
JG2
cJ21
EG1.4
R160
CF16
E1
(=KM19)
pMH1.1
E2.6
E.9
pUC112-17
XhoI
SalI
SacII
NotI
BssHII
ApaI
SacII
SmaI
SfiI
SalI
XhoI
cW10-1
cW44
TE5
T10
TE17b
TE8b
TE4b (R13.6)
R14.4 (TE9)
TE19bA
TE9b-A/TE2
TE11b
B14
W2
TE4bC
TE7
TE18
W3
TE19bB
TE7b
140-27
TE14A (R13.5)
T12
TE22 (R19)
TE1A, TE11A (R15.5)
TS24 (S22)
H2.8A
X.6
R14.4E1
E4.1
(=Mp6d.9)
J13
H1.2
E3.1
J44
E4.3
H1.6
H4.0
BssHII
SacII
BssHII
XhoI
XhoI
cW16-5
cW10-20
cW10-3
TE39 II
TE26III
cW10-23
J18
TE38 III
TE39III
R14.7 (TE11 II,TE2, TE 35, TE38 II)
TE22I / 25II / 26I / 27 / 29
TE36
TS21
TE39 I
TE24V
R15.1 (TE11 I, TE37 I)
II
III
IV
V
VI
VII
VIII
IX
X
T6/20
XhoI
NruI
XhoI
FspI
TE24I
TE28
TE33II
TE24II
TE23II
TE27III
TE21I,26IV
TE24IV
TE37III
TE23,30,31I,32II,37I,37II,38
XI
XII
XIII
XIV
XV
XVI
XVII
XVIII
XIX
XX
SfiI
SfiI
TE22
TE26II
TE27
XXIII
XXI
XXII
XXIV
D7S8
(7qter)

(position 215); and region 4 was initially recognized by the probes E4.3 and H1.6 (positions 264 to 268). The DNA segments that revealed sequence conservation were then tested for RNA hybridization and used to screen cDNA libraries of tissues affected in CF. Only a brief description of regions 1 to 3 is given below; region 4 corresponds to the 5′ end of the CF locus.

The probe G-2, one of the first segments tested, detected a 3.7-kb transcript in simian virus 40 (SV40)–transformed human fibroblasts (Fig. 4A). When this fragment was used to screen a human fibroblast and a human lung cDNA library, three independent clones were isolated (*25*). The overlapping cDNA clones spanned a length of 1.8 kb, and nucleotide sequence analysis revealed a potential open reading frame corresponding to the 3′ end of a coding region. Alignment of the cDNA sequence with that of the genomic DNA showed perfect sequence identity as well as exon-intron structures. Because this gene could not be the CF gene on the basis of genetic data (*16*), characterization studies were not continued.

Region 2 was identified by the cosmid clone CF14, which revealed strong cross-species hybridization signals in mouse, chicken, and bovine DNA (Fig. 3A). Restriction mapping of the genomic DNA showed that part of this region corresponded to the previously reported *IRP* (*8, 26*). The extent of this locus was subsequently confirmed by hybridization with oligonucleotide probes made to the IRP sequence of Wainwright *et al.* (*26*) (Fig. 1). As family studies indicated that *CF* maps to the D7S8 side of *IRP* (*15, 16*), chromosome walking and jumping experiments were continued in this direction.

The first region that revealed a transcript at a location on the D7S8 side of the IRP gene was identified by the probe CF16 (Fig. 1, positions 135 to 140). This probe detected RNA transcripts of different sizes in various tissues; a 2-kb species was observed in tracheal epithelium and pancreas, a less abundant 4-kb mRNA was seen in the brain, and a 9-kb transcript was observed in the liver (Fig. 4). When this probe was used to screen cDNA libraries made from human lung and cultured epithelial cells from sweat glands, more than ten clones were isolated. Restriction enzyme analyses of a subset of these cDNA clones revealed significant differences. Nucleotide sequence analyses of representative cDNA clones and the genomic DNA revealed that they shared a high degree (more than 85 percent) of sequence similarity but that none of the cDNA clones showed perfect identity with the genomic DNA sequence. Furthermore, neither the genomic DNA nor any of the cDNA clones contained an open reading frame. Screening a sequence databank (GenBank) showed that these clones share remarkable sequence similarity with a region in the β-globin locus (between ε and $^{G}\gamma$), suggesting that these sequences correspond to a transcribed repetitive DNA family that is distinct from the LINE-1 (long interspersed element) sequence (*27*).

Region 3 (position 215) contained a high proportion of CpG nucleotide residues, as determined by sequencing the 1-kb Eco RI fragment of genomic DNA. Open reading frames were also detected; however, neither RNA transcripts nor cDNA clones were detected with this probe. This could indicate that this transcript is restricted in tissue or developmental specificity, or that the notably weaker hybridization signals observed in other mammalian DNA's

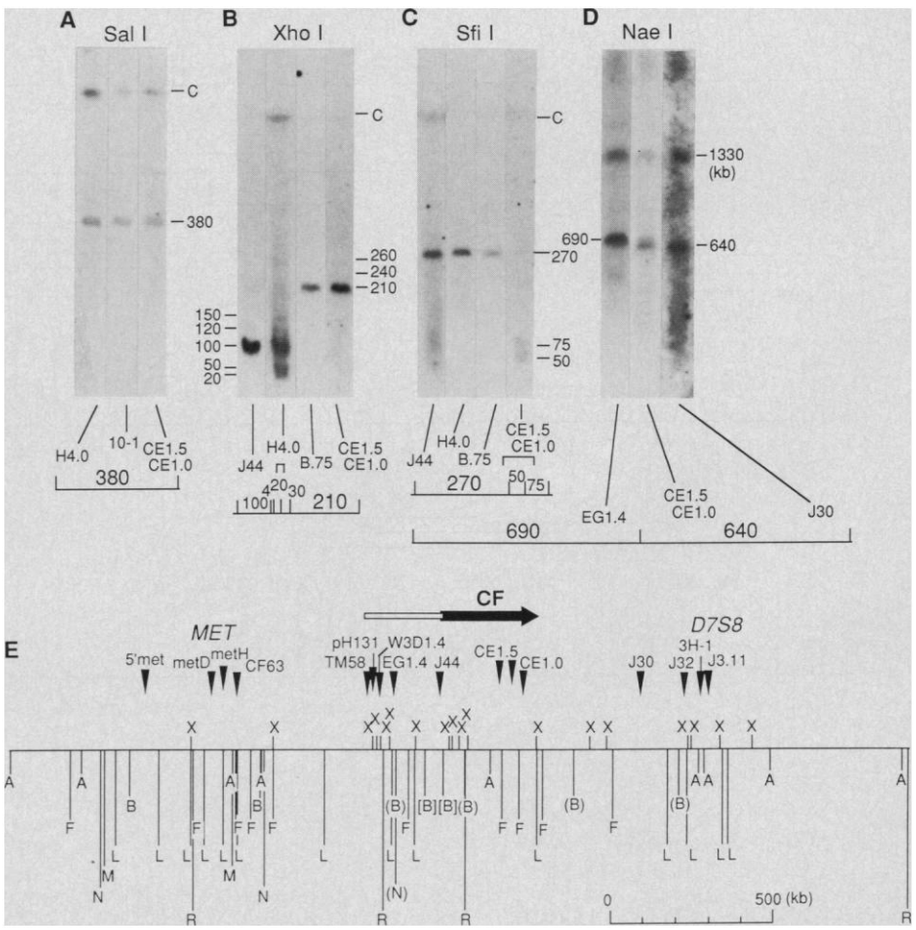
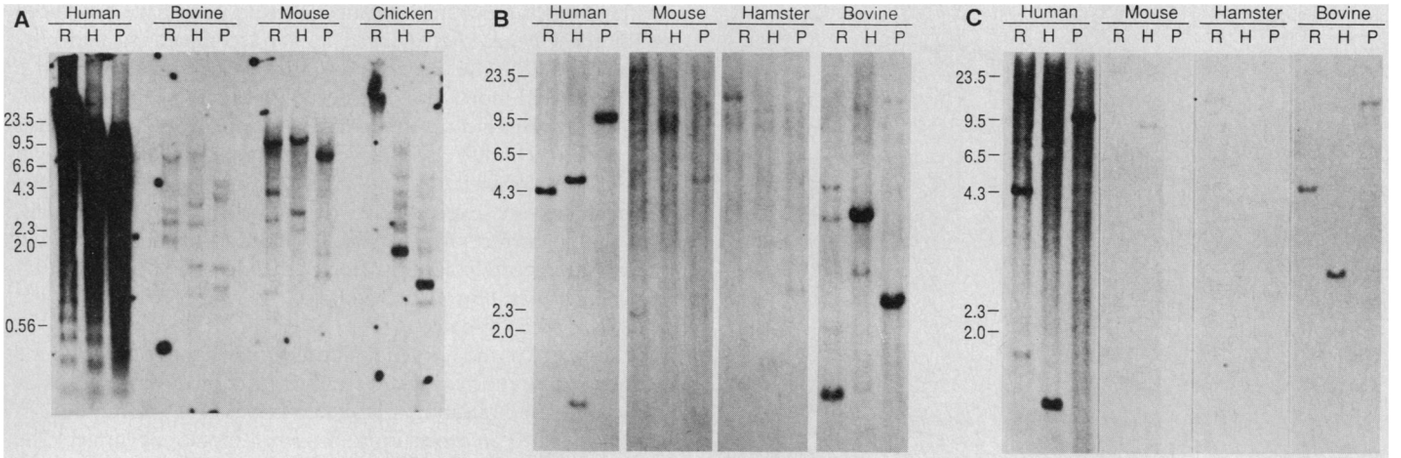


Fig. 2. Pulsed field gel electrophoresis mapping of the cloned region. DNA from the human-hamster cell line 4AF/102/K015 was digested with the enzymes (**A**) Sal I, (**B**) Xho I, (**C**) Sfi I, and (**D**) Nae I, and the fragments were separated by pulsed field gel electrophoresis and transferred to Zetaprobe (Bio-Rad). For each enzyme, a single blot was sequentially hybridized with the probes indicated below each panel, with stripping of the blot between hybridizations. DNA preparation, restriction enzyme digestion, and crossed field gel electrophoresis methods were as described (*11*). Electrophoresis was as follows: in 0.5× TBE (tris, borate, and EDTA) at 7 V/cm for 20 hours with switching linearly ramped from 10 to 40 for (A), (B), and (C), and at 8 V/cm for 20 hours with switching ramped linearly from 50 to 150 for (D). C corresponds to the compression zone region of the gel. Schematic interpretations of the hybridization pattern are given below each panel. Fragment lengths are in kilobases and were sized by comparison to oligomerized bacteriophage λ DNA and *Saccharomyces cerevisiae* chromosomes. Alignment of individual enzyme maps was facilitated by reference to previously described maps (*6, 11*). H4.0, J44, and EG1.4 are genomic probes generated from the walking and jumping experiments (see Fig. 1). J30 was isolated by four consecutive jumps from *D7S8* (*6, 13*). 10-1, B.75, and CE1.5 and CE1.0 together are cDNA probes that cover different regions of the CF transcript: 10-1 contains exons I to VI, B.75 contains exons V to XII, and CE1.5 and CE1.0 together contain exons XII to XXIV. Shown in (**E**) is a composite map of the entire *MET-D7S8* interval. The open box indicates the segment cloned by walking and jumping, and the arrow indicates the region covered by the CF transcript. The CpG-rich region associated with *D7S23* (*8*) is at the Not I site shown in parentheses. This and other sites shown in parentheses or square brackets are not cut in 4AF/102/K015 but have been observed in human lymphoblast cell lines. The symbols for each enzyme 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 I.



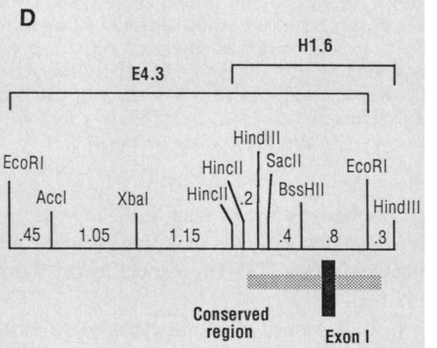


Fig. 3. Detection of conserved sequences by cross-species hybridization. Human, bovine, mouse, hamster, and chicken genomic DNA's were digested with Eco RI (R), Hind III (H), and Pst I (P), and the fragments were subjected to electrophoresis and blotted to Zetabind (Bio-Rad) as described (*10*). The hybridization procedures were also as described (*10*) with the most stringent washing being at 55°C in 0.2× SSC (standard saline citrate) and 0.1 percent SDS. Probes for hybridization (Fig. 1) included: (**A**) entire cosmid CF14, (**B**) E4.3, and (**C**) H1.6. The schematic in (**D**) shows a detailed restriction map of the overlapping segments E4.3 and H1.6. The shaded region indicates the area of cross-species conservation. Sizes are in kilobases.

were due to nonspecific hybridization of the CpG-rich sequence.

The CF locus. The next region of interest was first noted by the strong sequence conservation between human and bovine DNA with the probes E4.3 and H1.6 (Fig. 3, B and C); only weak hybridization was detected in the mouse and hamster DNA with the human probe. The fact that different subsets of bands were detected in bovine DNA with these two overlapping DNA segments suggested that the conserved sequences were located at the boundaries of the overlapped region (Fig. 3D). When these DNA segments were used to detect RNA transcripts from various tissues, no hybridization signal was detected. In an attempt to understand the cross-hybridizing region and to identify possible open reading frames, we determined the DNA sequences of the entire H1.6 and part of the E4.3 fragment. 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 (*23*), there were only short open reading frames, which could not easily explain the strong cross-species hybridization signals.

Undermethylated CpG islands have been associated with the 5′ ends of most housekeeping genes and a number of tissue-specific genes (*23*). To examine the methylation status of the 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.

Exhaustive screening of multiple cDNA libraries with the DNA segment H1.6 eventually yielded a single isolate (clone 10-1) carrying a 920-bp insert from a cDNA library constructed from cultured sweat gland cells of a non-CF individual (*18*). Nucleotide sequence analysis indicated that only 113 bp at the 5′ end of this clone aligned with sequences in H1.6 and thus provided a partial explanation for the poor hybridization signals observed in cDNA library screening. Use of the 10-1 cDNA as a probe revealed a 6.5-kb transcript in RNA from the T84 colon cancer cell line (Fig. 4). Results of further cDNA cloning experiments, sequencing, and genetic analysis suggested that H1.6 corresponds to the 5′ end of the gene most likely to be responsible for cystic fibrosis (*16, 18*). With the use of several additional overlapping cDNA clones, a number of genomic DNA segments were isolated from the recombinant phage and cosmid libraries. Alignment of these cloned genomic DNA segments with the long-range restriction map revealed that the locus spans ~250 kb (see Fig. 1). DNA sequencing and gel-blot hybridization demonstrated that this gene locus contains a minimum of 24 exons. Pulsed field analysis of this region from CF patients with a variety of haplotypes (*16*) gave no evidence for any visible genomic rearrangements within this interval (*28*). A detailed description of the coding region of this gene is now available (*18*).

Lessons from the search. A detailed analysis of 280 kb of contiguous DNA isolated by chromosome jumping and walking has permitted the cloning of the locus responsible for cystic fibrosis without prior knowledge of the basic defect. A major difficulty in identifying the CF locus has been the lack of chromosome rearrangements or deletions, which greatly facilitated all previous successes in the cloning of human disease genes by knowledge of map position (*29*). The strategy used in our study may therefore serve as an example for other similar disease gene cloning studies for which no gross genetic alteration has been demonstrated, although the task will be more challenging for disorders that are rare or where diagnosis is difficult.

As discussed above, the use of various molecular cloning techniques has led to the identification of DNA markers closely linked to *CF* (*6–10*). The positioning of these markers relative to each other was facilitated by somatic cell hybrid mapping (*10, 30*), linkage analysis (*3–5, 31*), and long-range restriction mapping with pulsed field gel electrophoresis (*8, 11, 12*). Through the cooperation of patient families, clinicians, and CF researchers throughout the world, many families, especially those in whom recombination

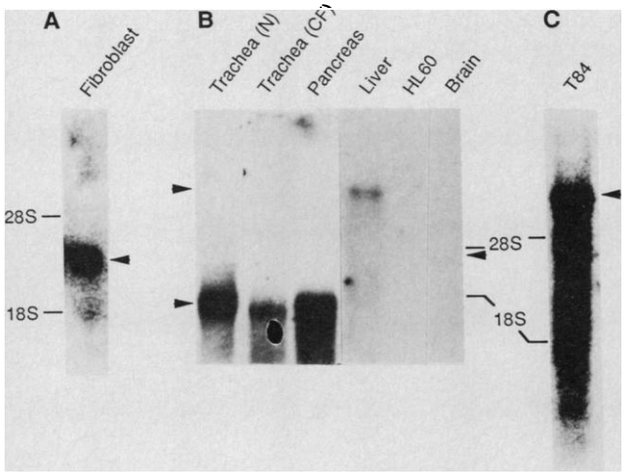


Fig. 4. RNA gel-blot hybridization analysis. RNA hybridization results are shown for the genomic probes G-2 (**A**) and CF16 (**B**). The cDNA clone 10-1 is the probe in (**C**). Approximately 10 μg of total RNA from each human tissue as indicated was separated on a 1 percent formaldehyde gel (*18*). Positions of the 28*S* and 18*S* ribosomal RNA bands are indicated, and arrows indicate the positions of transcripts. HL60 is a human promyelocytic leukemia cell line (*38*), and T84 is a human colon cancer cell line (*39*). Normal (N) and CF trachea are shown.

events had occurred near *CF* (*15, 24*), were identified and made available for genetic mapping studies, which led to an accurate localization of *CF* with respect to the flanking DNA markers (*4, 5, 10, 16*).

In the absence of useful cytogenetic landmarks to pinpoint *CF*, a systematic search of gene sequences within the entire region suggested by genetic data was required. Cloning from pulsed field gels and isolation of undermethylated CpG-rich regions (*7, 8*) served to identify further regions of interest, but cloning of a large contiguous stretch of DNA, as described in this article, allowed a more thorough examination of the region for candidate gene sequences. In this regard, the combination of chromosome walking and jumping appeared to be a highly productive strategy in covering the *CF* region. The jumping technique was particularly useful in bypassing "unclonable" regions, which are estimated to constitute 5 percent of the human genome (*17*). An alternative to this strategy would be the use of yeast artificial chromosome (YAC) vectors which allow cloning of large DNA fragments in the size ranges of 100 to 1000 kb (*32*); however, the construction of sublibraries with phage or cosmid vectors will probably still be required in order to generate a complete restriction map and identify candidate gene sequences from the YAC clones.

The challenge of identifying all gene sequences in a large DNA segment was also formidable, as no single method was guaranteed to succeed. In view of the experience described above, it would be advisable to attempt a combination of all available methods. Searching for sequence conservation by cross-hybridization is rapid, and the ability to use entire phage or cosmid clones represents a substantial simplification of the screening procedure. However, not all large segments of human DNA could be used in this way; simple repetitive sequences (for example, CA repeats) that are highly abundant in other animal species (*33*) can interfere with hybridization analysis. The evolutionary conservation of the E4.3 and H1.6 fragments, for example, was only apparent when these segments were isolated away from neighboring repeats.

The ultimate task in this type of "reverse genetics" approach (*34*) is to prove the identity of a candidate gene as, by definition, the basic biochemical defect of the disease is unknown. Appropriate tissue distribution and predicted properties of the gene product provide strong supporting evidence; these criteria have now been met for the CF gene, as detailed in the accompanying paper by Riordan *et al.* (*18*). The identification of a specific mutation which is found in affected individuals but never appears in normal chromosomes is much more compelling, and this evidence is presented for CF by Kerem *et al.* (*16*), who have now defined the most common CF mutation. Identification of other CF mutations will provide additional support. Expression of the normal cDNA in CF cells, which should correct the phenotypic chloride channel defect, will represent an important confirmation of the identity of the gene (*35*) and will be useful in the elucidation of the precise molecular pathology of the CF defect.

The large size of the CF gene came as somewhat of a surprise; the absence of apparent genomic rearrangements in CF chromosomes and the evidence indicating a limited number of CF mutations (*4*) might have led to an expectation of a small mutational target. The discovery that the most common CF abnormality gives rise to the loss of a single amino acid residue in a functional domain suggests, however, that the phenotype of CF is likely not due to complete loss of function of the gene product. In this regard, CF may be similar to the sickling disorders, where a very specific subset of mutations in the β-globin gene (β^S and β^C) give rise to an altered protein with unusual behavior (*36*). Complete absence of function of the β-globin gene gives rise to a different phenotype (β^0-thalassemia); similarly, homozygous loss of function of the CF protein product may lead to a different phenotype.

In summary, the application of genetic and molecular cloning strategies has allowed the cloning of the cystic fibrosis locus on the basis of its chromosomal location, even without the benefit of genomic rearrangements to point the way. Further improvements in "reverse genetics" technology should facilitate the identification of many more genetic loci of biological and medical importance.

REFERENCES AND NOTES

1. T. F. Boat, M. J. Welsh, A. L. Beaudet, in *The Metabolic Basis of Inherited Disease*, C. L. Scriver, A. L. Beaudet, W. S. Sly, D. Valle, Eds. (McGraw-Hill, New York, ed. 6, 1989), pp. 2649–2680.
2. M. Knowles, J. Gatzy, R. Boucher, *J. Clin. Invest.* **71**, 1410 (1983); P. M. Quinton and J. Bijman, *N. Engl. J. Med.* **308**, 1185 (1983); R. A. Frizzell, G. Rechkemmer, R. L. Shoemaker, *Science* **233**, 558 (1986); M. J. Welsh and C. M. Liedtke, *Nature* **322**, 467 (1986); R. A. Schoumacher *et al.*, *ibid.* **330**, 752 (1987); M. Li *et al.*, *Science* **244**, 1353 (1989); T.-C. Hwang *et al.*, *ibid.*, p. 1351; D. W. Landry *et al.*, *ibid.*, p. 1469.
3. L.-C. Tsui *et al.*, *Science* **230**, 1054 (1985); R. G. Knowlton *et al.*, *Nature* **318**, 380 (1985); R. White *et al.*, *ibid.*, p. 382; B. J. Wainwright *et al.*, *ibid.*, p. 384.
4. For reviews, see M. A. Spence and L.-C. Tsui, *Cytogenet. Cell Genet.* **46**, 170 (1988); M. Dean, *Genomics* **3**, 93 (1988); L.-C. Tsui, *Am. J. Hum. Genet.* **44**, 303 (1989).
5. A. Beaudet *et al.*, *Am. J. Hum. Genet.* **39**, 681 (1986); G. M. Lathrop *et al.*, *ibid.* **42**, 38 (1988).
6. F. S. Collins *et al.*, *Science* **235**, 1046 (1987); M. C. Iannuzzi *et al.*, *Am. J. Hum. Genet.* **44**, 695 (1989).
7. F. Michiels, M. Burmeister, H. Lehrach, *Science* **236**, 1305 (1987).
8. X. Estivill *et al.*, *Nature* **326**, 840 (1987).
9. R. Kaiser *et al.*, *Mol. Biol. Rep.* **12**, 3 (1987).
10. J. M. Rommens *et al.*, *Am. J. Hum. Genet.* **43**, 4 (1988).
11. J. M. Rommens *et al.*, *ibid.*, in press.
12. A.-M. Poustka, H. Lehrach, R. Williamson, G. Bates, *Genomics* **2**, 337 (1988); M. L. Drumm *et al.*, *ibid.*, p. 346.
13. M. Dean *et al.*, in preparation.
14. Twelve genomic DNA libraries were used in this study. This included eight phage libraries, one of which was provided by T. Maniatis [E. F. Fritsch, R. M. Lawn, T. Maniatis, *Cell* **19**, 959 (1980)]; the rest were constructed as part of this work according to procedures described in (*20*). Four phage libraries were cloned in λDASH (Stratagene) and three in λFIX (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) (*10*), and the other libraries from human peripheral blood or lymphoblastoid DNA after partial digestion with Sau 3A or total digestion with Bam HI or Eco RI . To avoid loss of unstable sequences, five of the phage libraries were propagated in the recombination-deficient hosts DB1316 (*recD*$^-$) (*17*), CES200 (*recBC*$^-$) (*17*), or TAP90 [T. A. Patterson and M. Dean, *Nucleic Acids Res.* **15**, 6298 (1987)]. Three cosmid libraries were constructed. In one, the vector pCV108 [Y. F. Lau and Y. W. Kan, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 5225 (1983)] was used to clone partially digested (Sau 3A) DNA from 4AF/102/K015 (*10*). A

second cosmid library was prepared by cloning partially digested (Mbo I) human lymphoblastoid DNA into the vector pWE-IL2R, prepared by inserting the Rous sarcoma virus (RSV) promoter-driven DNA for the interleukin-2 receptor α chain (supplied by M. Fordis and B. Howard) in place of the *neo*-resistance gene of pWE15 [G. M. Wahl *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2160 (1987)]. An additional partial Mbo I cosmid library was prepared in the vector pWE-IL2R-Sal, created by inserting a Sal I linker into the Bam HI cloning site of pWE-IL2R (M. L. Drumm, unpublished data); this allows the use of the partial fill-in technique to ligate Sal I and Mbo I ends, preventing tandem insertions [E. R. Zabarovsky and R. L. Allikmets, *Gene* **42**, 119 (1986)]. Cosmid libraries were propagated in *Escherichia coli* host strains DH1 or 490A [M. Steinmetz, A. Winoto, K. Minard, L. Hood, *Cell* **28**, 489 (1982)]. 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 (*20, 21*). 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 DNA gel-blot analysis to confirm its colinearity with the genome. The chromosome jumping library has been described (*6*). The original library was prepared from a preparative pulsed field gel and was intended to contain partial Eco RI fragments of 70 to 130 kb; subsequent experience with this library (including that reported here) indicates that smaller fragments are also represented, and jumps of 25 to 110 kb have been found. The library was plated on *sup*$^-$ host MC1061 and screened by standard techniques (*20*). Positive clones were subcloned into pBRΔAva, and the beginning and end of the jump were identified by Eco RI and Ava I digestion [F. S. Collins, in *Genome Analysis: A Practical Approach*, K. E. Davies, Ed. (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. About 10 percent of the clones in this library arise from noncircular ligations and thus give rise to anomalous jumps (F. S. Collins, unpublished).
15. M. Farrall *et al.*, *Am. J. Hum. Genet.* **43**, 471 (1988).
16. B. Kerem *et al.*, *Science* **245**, 1073 (1989); B. Kerem *et al.*, *Am. J. Hum. Genet.* **44**, 827 (1989).
17. A. R. Wyman, L. B. Wolfe, D. Botstein, *Proc. Natl. 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, *ibid.*, p. 263.
18. J. R. Riordan *et al.*, *Science* **245**, 1066 (1989).
19. S. M. Arfin, R. E. Circullo, V. F. X. Arredondo, M. Smith, *Somatic Cell Genet.* **9**, 517 (1983).
20. T. Maniatis, E. F. Fritsch, J. Sambrook, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982).
21. DNA probes that gave high background signals could often be used more successfully by annealing the boiled probe with sheared, denatured, placental DNA (250 μg/ml) for 60 minutes before the probe was added to the hybridization bag.
22. For example, localization of the Duchenne muscular dystrophy locus was facilitated by cross-species hybridization, as described in A. P. Monaco *et al.* [*Nature* **323**, 646 (1986)].
23. A. P. Bird, *Nature* **321**, 209 (1986); M. Gardiner-Garden and M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).
24. L.-C. Tsui *et al.*, *Am. J. Hum. Genet.* **39**, 720 (1986); R. White *et al.*, *ibid.*, p. 694; W. Berger *et al.*, *Hum. Genet.* 77, 197 (1987).
25. The SV40-transformed human fibroblast cDNA library, with 1.4×10^6 independent clones, was provided by P. Berg [H. Okayama and P. Berg, *Mol. Cell. Biol.* **3**, 280 (1983)].
26. B. J. Wainwright *et al.*, *EMBO J.* 7, 1743 (1988).
27. T. G. Fanning and M. F. Singer, *Biochim. Biophys. Acta* **910**, 203 (1987).
28. J. M. Rommens *et al.*, unpublished data.
29. B. Royer-Pokora *et al.*, *Nature* **322**, 32 (1986); M. Koenig *et al.*, *Cell* **50**, 509 (1987); S. H. Friend *et al.*, *Nature* **323**, 643 (1986); D. C. Page *et al.*, *Cell* **51**, 1091 (1987).
30. I. Bartels *et al.*, *Am. J. Hum. Genet.* **38**, 280 (1986); S. Zengerling *et al.*, *ibid.* **40**, 228 (1987).
31. D. Barker *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8006 (1987).
32. D. T. Burke, G. F. Carle, M. V. Olson, *Science* **236**, 806 (1987); B. H. Brownstein *et al.*, *ibid.* **244**, 1348 (1989).
33. R. Meisfeld, M. Krystal, N. Arnheim, *Nucleic Acids Res.* **9**, 5931 (1981); H. Hamada, M. G. Petrino, T. Kakunaga, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6465 (1982).
34. S. H. Orkin, *Cell* **47**, 845 (1986); F. Ruddle, *Am. J. Hum. Genet.* **36**, 944 (1984).
35. M. L. Drumm *et al.*, *Am. J. Hum. Genet.* **43**, A182 (1988); A. M. Jetten, J. R. Yankaskas, M. J. Stutts, N. J. Willumsen, R. C. Boucher, *Science* **244**, 1472 (1989).
36. H. F. Bunn and B. G. Forget, *Hemoglobin: Molecular, Genetic and Clinical Aspects* (Saunders, Philadelphia, 1986).
37. X. Estivill *et al.*, *Am. J. Hum. Genet.* **44**, 704 (1989).
38. S. J. Collins, F. W. Ruscetti, R. E. Gallagher, R. C. Gallo, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2458 (1978).
39. K. Dharmsathaphorn *et al.*, *Am. J. Physiol.* **246**, G204 (1984).
40. We thank N. Alon, M. Collins, B. Gerrard, J. Grim, Z. Grzelczak, S. Lok, D. Markiewicz, K. Murphy, L. Naismith, S. Norton, A. Perry, J. Pinnard, N. Plavsic, A. Saulino, C. Stewart, C. TomHon, Z. Weaver, and J. Zielenski for technical assistance; B. Sandri for preparing the manuscript; and J. Yankaskas and R. Boucher for RNA samples. Supported by NIH grants DK39690 (F.S.C.) and DK34944 (L.-C.T.); the Cystic Fibrosis Foundation and the Canadian Cystic Fibrosis Foundation; a postdoctoral fellowship of the Medical Research Council (MRC) of Canada (J.M.R.); an MRC Scientist award (L.-C.T.); the Deutsche Forschungsgemeinschaft (G.M.); the Sellers Foundation (M.B., J.R.R., and L.-C.T.); the North Dakota Cystic Fibrosis Association (L.-C.T. and M.B.); and federal funds from the Department of Health and Human Services under contract NO1-CO-74102 with Program Resources, Inc. (M.D.). F.S.C. is an Associate Investigator in the Howard Hughes Medical Institute.

7 August 1989; accepted 17 August 1989