Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes

Rachel E. Ellsworth*, D. Curtis Jamison*, Jeffrey W. Touchman†, Stephanie L. Chissoe‡, Valerie V. Braden Maduro*, Gerard G. Bouffard†, Nicole L. Dietrich†, Stephen M. Beckstrom-Sternberg†, Leslie M. Iyer*, Lauren A. Weintraub*, Marc Cotton‡, Laura Courtney‡, Jennifer Edwards‡, Rachel Maupin‡, Philip Ozersky‡, Theresa Rohlfing‡, Patricia Wohldmann‡, Tracie Miner‡, Kimberley Kemp‡, Jason Kramer‡, Ian Korf‡, Kimberlie Pepin‡, Lucinda Antonacci-Fulton‡, Robert S. Fulton‡, Patrick Minx‡, LaDeana W. Hillier‡, Richard K. Wilson‡, Robert H. Waterston‡, Webb Miller§, and Eric D. Green*†¶

*Genome Technology Branch and †National Institutes of Health Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; ‡Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110; and §Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802

Communicated by Francis S. Collins, National Institutes of Health, Bethesda, MD, November 24, 1999 (received for review September 17, 1999)

The identification of the cystic fibrosis transmembrane conductance regulator gene (*CFTR*) in 1989 represents a landmark accomplishment in human genetics. Since that time, there have been numerous advances in elucidating the function of the encoded protein and the physiological basis of cystic fibrosis. However, numerous areas of cystic fibrosis biology require additional investigation, some of which would be facilitated by information about the long-range sequence context of the *CFTR* gene. For example, the latter might provide clues about the sequence elements responsible for the temporal and spatial regulation of *CFTR* expression. We thus sought to establish the sequence of the chromosomal segments encompassing the human *CFTR* and mouse *Cftr* genes, with the hope of identifying conserved regions of biologic interest by sequence comparison. Bacterial clone-based physical maps of the relevant human and mouse genomic regions were constructed, and minimally overlapping sets of clones were selected and sequenced, eventually yielding ≈1.6 Mb and ≈358 kb of contiguous human and mouse sequence, respectively. These efforts have produced the complete sequence of the ≈189-kb and ≈152-kb segments containing the human *CFTR* and mouse *Cftr* genes, respectively, as well as significant amounts of flanking DNA. Analyses of the resulting data provide insights about the organization of the *CFTR*/*Cftr* genes and potential sequence elements regulating their expression. Furthermore, the generated sequence reveals the precise architecture of genes residing near *CFTR*/*Cftr*, including one known gene (*WNT2*/*Wnt2*) and two previously unknown genes that immediately flank *CFTR*/*Cftr*.

Cystic fibrosis (CF) is one of the most common inherited disorders in individuals of northern European descent (1). With an autosomal recessive pattern of inheritance, CF has a carrier frequency of ≈1 in 30 Caucasians, with ≈1,000 affected individuals born in the United States each year. The disease is associated with pancreatic insufficiency, repeated pulmonary infections, intestinal blockages, elevated sweat chloride levels, and male infertility. Despite major improvements in therapeutic approaches, CF remains a major and challenging health problem.

One decade ago, investigators elucidated the underlying genetic defect responsible for CF with the identification of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene (2–4). This discovery represents one of the most important triumphs in contemporary human genetics, demonstrating the efficacy of positional cloning as a strategy for identifying genes associated with human disease (5, 6), even in the absence of cytogenetic rearrangements that assist the search. Residing on chromosome 7q31.3 (7) and consisting of 27 exons, the human *CFTR* gene encodes a 6,129-bp transcript that directs the synthesis of a 1,480-aa protein (2, 3) shown to function as a chloride channel (8–10).

In the decade since the identification of the *CFTR* gene, there have been various advances in understanding the function of the encoded protein and the pathogenesis of CF (11). These have included the development of animal models for the disease (12–17) as well as new avenues for drug- (18, 19) and gene therapy- (20) based treatment modalities. Less clear are the mechanisms responsible for the characteristic tissue-specific expression of *CFTR* (e.g., in pancreas, lung, sweat glands, intestine, and liver). A number of studies have pointed to possible *CFTR* promoter sequences and regulatory elements (21–32), including regions with high GC content, Sp1 and AP-1 binding sites, and DNase I hypersensitivity sites. These sequences, some of which constitute the basal *CFTR* promoter ≈250 bp upstream of the translation start site, resemble regulatory elements more typically found in housekeeping as opposed to tissue-specific genes. Furthermore, studies in other animal species (e.g., mouse) have revealed slightly different expression patterns of the gene (33). In short, the regulatory mechanism(s) responsible for the tissue-specific expression of *CFTR* remains unclear.

To facilitate studies of *CFTR* structure and function, we sought to establish the complete sequence of the genomic segments containing the human and mouse genes. Here we report the high-resolution mapping and systematic sequencing of an ≈1.6-Mb segment of human chromosome 7 containing *CFTR*, an effort performed under the auspices of the ongoing Human Genome Project (34, 35). In addition, we performed parallel analyses of the corresponding *Cftr*-containing region of mouse chromosome 6, producing ≈358 kb of contiguous mouse sequence. The availability of homologous human and mouse sequence has allowed important long-range, cross-species sequence comparisons to be performed.

Abbreviations: CF, cystic fibrosis; BAC, bacterial artificial chromosome; PAC, P1-derived artificial chromosome; PIP, percent identity plot; STS, sequence-tagged site.

Data deposition: The sequences reported in the paper have been deposited in the GenBank database (accession nos. AC002542, AC003987, AC006326, AC002465, AC003045, AC000111, AC000061, AC004240, AC002431, AC003084, AC004536, AC006389, AC007874, AC004029, AC006926, AC002529, and AF162137).

¶To whom reprint requests should be addressed at: Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, 49 Convent Drive, Building 49, Room 2A08, Bethesda, MD 20892. E-mail: egreen@nhgri.nih.gov.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

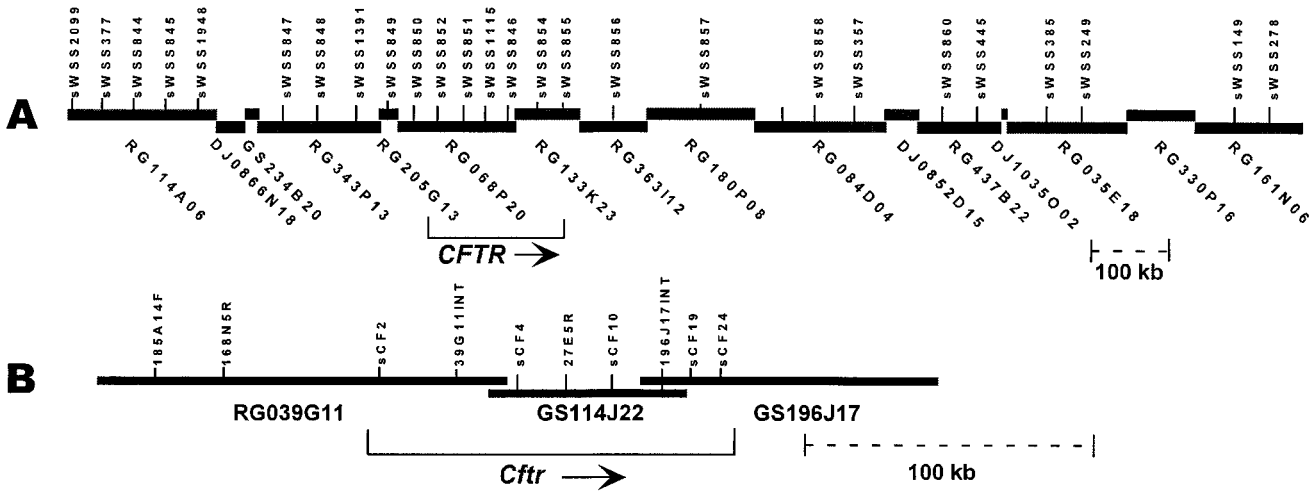

Fig. 1. Sequence maps of the genomic segments encompassing the human *CFTR* and mouse *Cftr* genes. High-resolution BAC/PAC-based physical maps of the human and mouse *CFTR*/*Cftr* regions were assembled, with the complete contig maps available at http://genome.nhgri.nih.gov/chr7/cftr. From each map, minimal overlapping sets of ordered clones were selected and completely sequenced. (*A*) The sequence map of the *CFTR* region on human chromosome 7q31.3 (7) consists of the indicated 16 ordered clones that together span ≈1.6 Mb. Numerous STSs have been mapped to this region (50, 58) (also see http://genome.nhgri.nih.gov/chr7), with a small subset indicated relative to their content in particular BACs/PACs. The *CFTR* gene resides in BACs RG068P20 and RG133K23. Note that the human clones are depicted as barely overlapping, reflecting the actual sequence records in GenBank. Before submission, the sequence generated from each human clone was trimmed to yield the nonredundant sequence from that clone flanked by very small amounts of sequence in common with adjacent clones. Thus, the actual overlaps between adjacent human clones are typically much larger than that reflected by the sequence in their GenBank records. (*B*) The sequence map of the *Cftr* region on mouse chromosome 6 consists of the indicated three ordered clones that together span ≈358 kb. Representative STSs used to assemble the mouse contig map are depicted relative to their content in particular BACs. Note that the mouse clones are depicted based on their size and degree of overlap with one another; a single GenBank record (accession no. AF162137) contains one contiguous sequence assembled from all three clones. Information about the indicated human and mouse STSs is available in GenBank.

Materials and Methods

Construction of a Human Bacterial Clone-Based Physical Map. Human DNA-containing bacterial clones [bacterial artificial chromosomes (BACs) and P1-derived artificial chromosomes (PACs)] were isolated, analyzed, and assembled in contigs as described (36, 37). Clones are named with a prefix reflecting their library of origin [RG, Research Genetics human BAC library (Huntsville, AL); GS, Genome Systems human BAC library (St. Louis); DJ, Roswell Park Cancer Institute PAC library (Buffalo, NY)].

Construction of a Mouse BAC-Based Physical Map. Purified inserts from three cDNA clones containing the mouse *Cftr* coding region (38, 39) (see http://www.atcc.org, nos. 63165, 63166, and 63167) were radiolabeled and hybridized to membrane filters containing high-density arrays of mouse BAC clones [mouse C57BL/6 library from Genome Systems (GS); mouse ES/129Sv library from Research Genetics (RG)]. Positive clones were colony purified, were reanalyzed by hybridization to confirm the presence of probe sequences, and were subjected to restriction enzyme digest-based fingerprint analysis (40).

BAC/PAC Clone Sequencing. The high-accuracy sequence of all BAC/PAC clones was established by a shotgun sequencing strategy. Human clones were sequenced at the Washington University Genome Sequencing Center by well established methods (41–43) (also see http://genome.wustl.edu/gsc). The sequence of each human clone was submitted to GenBank as a separate record. Mouse clones were sequenced at the NIH Intramural Sequencing Center (see http://www.nisc.nih.gov) by similar methods. Individual sequences were edited and assembled with the PHRED/PHRAP/CONSED suite of programs (44–46). The sequence of the three overlapping mouse BACs was assembled into one contiguous block and was submitted to GenBank as a single record (GenBank accession no. AF162137). The approaches used for sequencing both human and mouse DNA should produce error rates of less than 1 in 10^4 bp, something confirmed by recent quality assessment experiments (47).

Human–Mouse Sequence Comparison. Repetitive sequences were masked with the REPEATMASKER program (A. F. A. Smit and P. Green, see http://www.genome.washington.edu/UWGC/analysistools/repeatmask.htm), and the resulting sequence was aligned with a modified version of the SIM program (48) by using the default parameters (+1 for a match, −1 for a mismatch, and $-6 - 0.2k$ for a gap length of k). For another view of the alignment, regions between successive gaps were converted into segments of percent identity relative to positions in the human sequence, with the resulting data then drawn as a percent identity plot (PIP) by using the program LAPS (49). Only segments with an identity of 50% or greater were plotted, so regions that match poorly appear blank. Gaps within an alignment appear as discontinuities between adjacent horizontal lines. We also analyzed the sequences by using the web site http://globin.cse.psu.edu/pipmaker, which utilizes another alignment program with a distinct substitution matrix and permits matching regions within the sequences to be at different positions and orientations. No significant differences were seen between the results of the two analyses.

Results

Bacterial Clone-Based Physical Mapping of the Human and Mouse *CFTR*/*Cftr* Regions. To assemble a BAC/PAC-based contig map of the human *CFTR* region, mapped sequence-tagged sites (STSs) residing within an ≈2-Mb region encompassing the gene (50) (see also http://genome.nhgri.nih.gov/chr7) were used to isolate corresponding BACs and PACs. Analysis of the isolated clones and subsequent contig expansion resulted in the assembly of a >1.6-Mb contig consisting of ≈169 clones (the complete

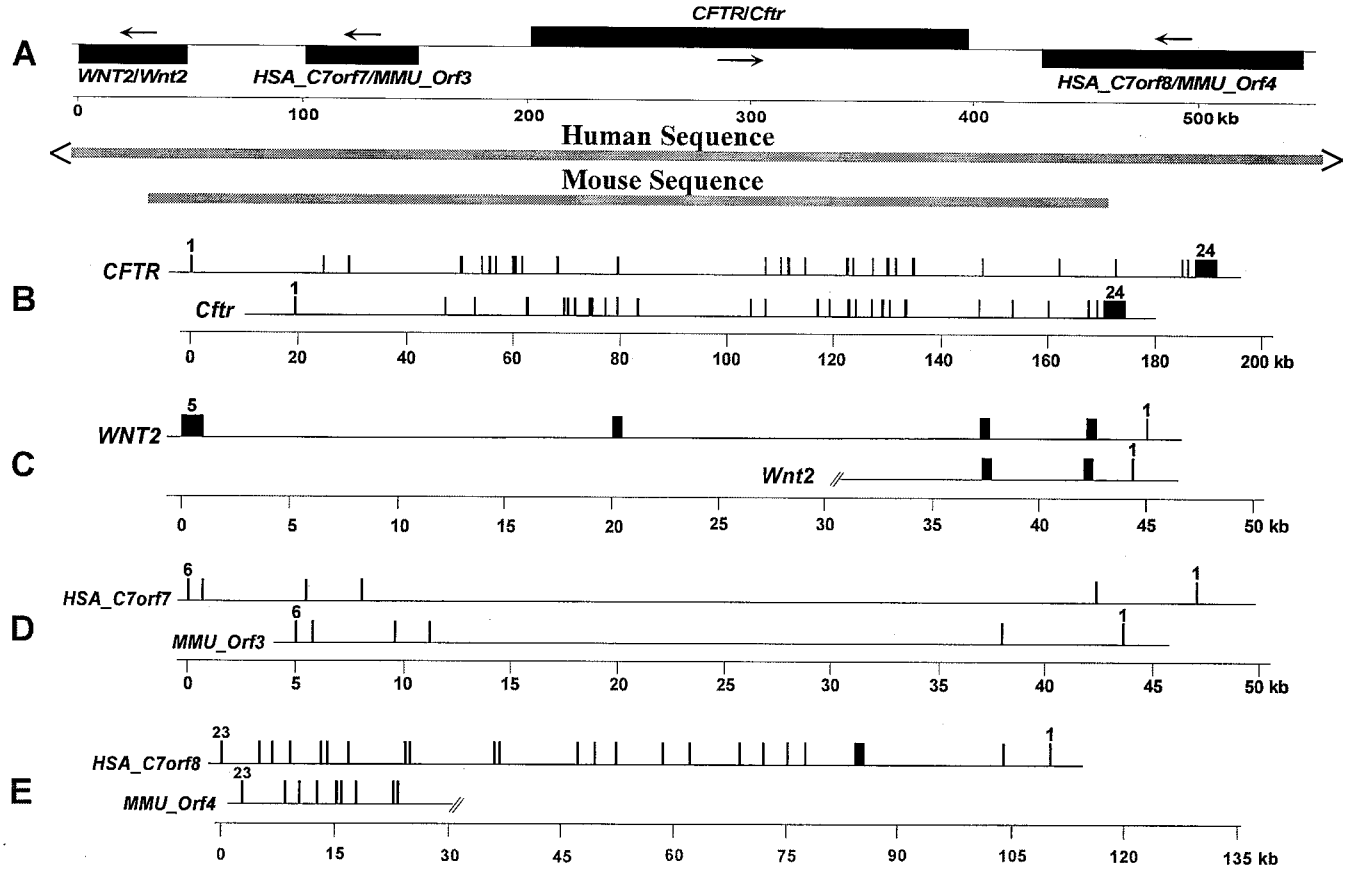

Fig. 2. Long-range organization of the greater human and mouse *CFTR*/*Cftr* regions. (*A*) Schematic overview of the location of genes identified in both the human and mouse genomic sequence: *WNT2*/*Wnt2*, *HSA_C7orf7*/*MMU_Orf3*, *CFTR*/*Cftr*, and *HSA_C7orf8*/*MMU_Orf4*. Arrows indicate the direction of transcription. The available human sequence spans all four genes as well as extensive amounts of flanking DNA whereas the generated mouse sequence ends in the middle of the *Wnt2* gene on the centromeric side and in the middle of *MMU_Orf4* on the telomeric side. Higher-resolution comparative depictions of the intron/exon organization of human and mouse *CFTR*/*Cftr* (*B*), *WNT2*/*Wnt2* (*C*), *HSA_C7orf7*/*MMU_Orf3* (*D*), and *HSA_C7orf8*/*MMU_Orf4* (*E*) are also provided (in each case, with the human gene drawn above the mouse gene). The exon sizes are identical between species whereas intron sizes vary. Note that the *CFTR*/*Cftr* exons are numbered as originally designated (2), even though the gene is now known to contain 27 exons (with exons 6A, 6B, 14A, 14B, 17A, and 17B).

contig map is available at http://genome.nhgri.nih.gov/chr7/cftr). A minimal overlapping set of 16 clones was selected and sequenced (Fig. 1*A*). These efforts yielded ≈1.6 Mb of contiguous sequence. A more focused effort was applied to the mapping and sequencing of the mouse *Cftr* region, resulting in the assembly of a BAC contig encompassing the gene (see http://genome.nhgri.nih.gov/chr7/cftr). Three clones (from two mouse strains) were selected and sequenced (Fig. 1*B*), yielding ≈358 kb of contiguous sequence.

***CFTR*/*Cftr* Gene Structures.** The availability of large blocks of homologous human and mouse sequence allowed detailed analysis and comparisons of the *CFTR* and *Cftr* genes. Human *CFTR* spans ≈189 kb, ≈60 kb less than initially estimated (2). Comparison of the genomic sequence with the original long-range restriction map of the region (2) suggests that this discrepancy most likely reflects the previous overestimation of the sizes of large restriction fragments by pulsed-field gel analysis. Comparison of the *CFTR* genomic and cDNA sequences confirms the presence of 27 exons (Fig. 2*B*). Each intron is flanked by the consensus GT-AG splice-site sequence (51), as previously reported (52). Although exon sizes are consistent with previous reports, intron sizes range from 599 bp (intron 22) to 28.1 kb (intron 10), notably different than earlier estimates (52).

Mouse *Cftr* spans ≈152 kb, considerably less than its human counterpart. The REPEATMASKER program detected 33,142 bp of interspersed repeats within the *Cftr* introns, compared with 58,168 bp in *CFTR* introns. All of the 27 exons are highly similar between human and mouse at a sequence level (see below and http://genome.nhgri.nih.gov/chr7/cftr). In addition, the intron/exon structures of the mouse and human genes are mostly the same, with splice sites occurring at identical positions in both species (Fig. 2*B*). As with human *CFTR*, each mouse *Cftr* intron is flanked by the consensus GT-AG splice-site sequence. Of note, the polymorphic polyT tract located upstream of exon 9, which is implicated in the variable in-frame skipping of exon 9 during transcription (53), is absent in the mouse. Finally, although most of the introns are larger in the human gene, three mouse *Cftr* introns are notably larger than their human counterparts; specifically, introns 1, 4, and 12 are 27.6 vs. 24.1 kb, 6.3 vs. 3.2 kb, and 9.4 vs. 1.5 kb, respectively (in mouse vs. human).

Neighboring Genes. The human and mouse genomic sequences were aligned and subjected to comparative analyses. To obtain a graphical overview of the resulting local alignments, we used a PIP in which each gap-free section of an alignment is represented as a horizontal line that indicates the position in the human sequence and the percent nucleotide identity (a more comprehensive summary of the PIP analysis is presented as an electronic supplement to this paper at http://genome.nhgri.

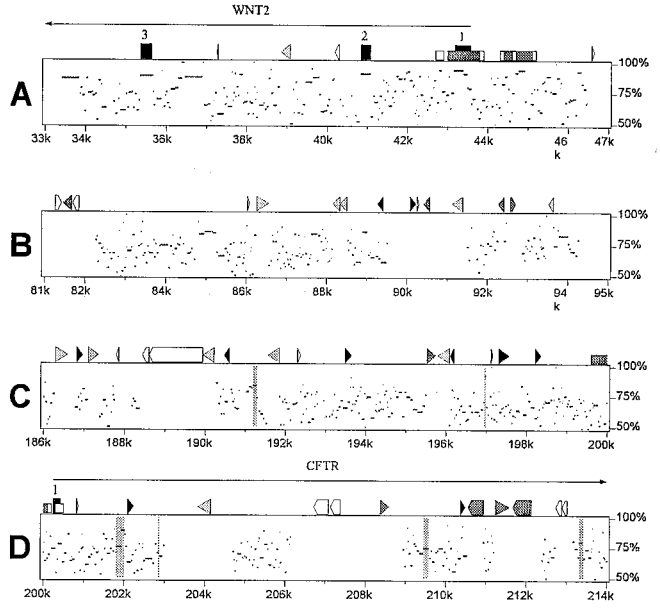

Fig. 3. Percent identity plots (PIPs) for human and mouse genomic sequences. Percent identity plots (60) (see http://globin.cse.psu.edu/pipmaker) for four 14-kb segments are provided: (*A*) a region containing exons 1–3 of *WNT2*/*Wnt2* (nucleotides 33,000–47,000 of the human sequence in GenBank accession no. AC002465); (*B*) a region residing between the *WNT2*/*Wnt2* gene and *HSA_C7orf7*/*MMU_Orf3* with no known functional elements (nucleotides 81,000–95,000 in GenBank accession no. AC002465); (*C*) a region immediately upstream of *CFTR*/*Cftr* exon 1 (nucleotides 5,425–19,425 in GenBank accession no. AC000111); and (*D*) a region containing the proximal portion of *CFTR*/*Cftr* intron 1 (nucleotides 19,425–33,425 in GenBank accession no. AC000111). In *C* and *D*, the vertical stripes are used to highlight the gap-free regions in an ≈28-kb interval encompassing *CFTR*/*Cftr* exon 1 that have a higher percent identity than other gap-free regions in that interval of the same or larger length. Features in the PIP: tall black rectangle, exon; white pointed box, L1-type repeat; dark gray pointed box, LTR repeat; black triangle, MIR-type repeat; light gray triangles, other SINE-type repeat; dark gray triangles, all other interspersed repeats; short white rectangle, CpG island where $0.6 \leq$ CpG/GpC < 0.75; short dark gray rectangle, CpG island where CpG/GpC ≥ 0.75.

nih.gov/chr7/cftr). For instance, the alignment shown in Fig. 3*B* begins as follows (with the human sequence aligned on top of the corresponding mouse sequence):

```
82270 TAATCAGGTGAAATAATCAGAGTAGGAAGAGACTTATGCTCTAGATGA
      |||| | || ||  |||||||||| ||||| |||| -||  |||||||--
28578 TAATAAAGTCAACAAATCAGAGTGGGAAGGGACT ATAGTCTAGAT
```

The first gap-free section covers human positions 82,270–82,303 at 79% nucleotide identity whereas the second covers 82,305–82,315 at 82% identity. This gives the first two tiny horizontal lines in Fig. 3*B*.

The region between the first and last aligned nucleotides spanned 422,311 and 356,976 bp in the human and mouse sequence, respectively, including ≈167 and ≈138 kb upstream of the *CFTR* and *Cftr* genes, respectively. Within this region, much of the sequence could be aligned. By using the approach described by Endrizzi *et al.* (54) to assess the degree of conservation between human and mouse DNA, 49.9% of the unique, noncoding human sequence from the *CFTR* region aligns with the corresponding mouse sequence. Interestingly, this makes the *CFTR*/*Cftr* region more highly conserved than 8 of the 10 regions surveyed by Endrizzi *et al.* (54), which ranged from 6.4 to 78.1%.

Comparative sequence analysis also revealed important insights about genes neighboring *CFTR*/*Cftr.* Complete sequence was generated for human *WNT2*, a gene first identified in 1987 during the search for the CF gene (55) (GenBank accession no. NM_003391). This gene contains five exons, spans ≈45 kb, and is located on the antisense strand relative to *CFTR* (Fig. 2 *A* and *C*). The generated mouse sequence extends through the first three exons of the mouse *Wnt2* gene (Fig. 2*C*). The sequence similarity between the human and mouse genes is high (e.g., see positions 35.5, 41, and 43 kb in Fig. 3*A*, which corresponds to exons 3, 2, and 1, respectively). In addition, sequences flanking exon 3 at 33.5 and 36.5 kb are highly conserved (Fig. 3*A*), perhaps indicating the presence of functional intronic elements (see below).

Several other regions with high human–mouse sequence conservation were identified between *WNT2*/*Wnt2* and *CFTR*/*Cftr* (Figs. 3*B* and 2*A*). Within this interval, the GENSCAN program (56) detected a series of putative exons on the antisense strand relative to *CFTR*/*Cftr* (Fig. 2*D*). There were no high-quality matches between these predicted exons and known genes or expressed sequence tags; however, BLAST analysis (57) of the predicted protein revealed homology to a *Caenorhabditis elegans* protein (GenBank accession no. U50071). Preliminary PCR-based studies have revealed that this gene (*HSA_C7orf7*/*MMU_Orf3*) is expressed in a number of human epithelial tissues, including prostate, colon, and mammary gland (J. C. Zenklusen and E.D.G., unpublished data). Interestingly, *HSA_C7orf7* likely corresponds to the conserved, CpG island-containing segment immediately upstream of *CFTR* that was evaluated as a candidate during the successful search for the CF gene (2).

Immediately downstream of *CFTR*, GENSCAN predicts a 23-exon human gene spanning ≈108 kb (Fig. 2 *A* and *E*). Some of the predicted exons have high-quality matches to available expressed sequence tags (e.g., GenBank accession nos. W24687, AA740322, and R51798), including those derived from fetal liver and infant brain. This gene (*HSA_C7orf8*), present on the antisense strand relative to *CFTR*, is predicted to encode a 1,692-aa protein. The generated mouse sequence (from clone GS196J17) contains the last nine predicted exons of the gene (*MMU_Orf4*). PIP analysis (see http://genome.nhgri.nih.gov/chr7/cftr) reveals strong interspecies sequence conservation for all of these exons. However, the amino acid sequence encoded by these conserved exons has no convincing match to any known protein. *HSA_C7orf8* was not recognized during the successful search for the CF gene because this region was not explored (2).

Potential *CFTR*/*Cftr* Regulatory Elements. We were particularly interested to see whether comparative sequence analysis would reveal conserved sequences that might play a role in regulating *CFTR*/*Cftr* expression. Particular attention was paid to segments within intron 1 (Fig. 3*D*) and ≈10 kb upstream of exon 1 (Fig. 3*C*). The latter includes the minimal promoter region (21, 22) as well as putative regulatory elements previously described in the human and mouse sequence (23, 24).

The most striking feature about the observed human–mouse sequence conservation in the 10 kb upstream of *CFTR*/*Cftr* exon 1 is its extreme uniformity (Fig. 3*C*). No segment stands out as markedly more conserved than others; thus, there are relatively few long horizontal lines. Even Fig. 3*B*, which depicts a region with no known or hypothesized functional elements, shows greater variation (e.g., the well conserved segment at 94 kb). A simple statistic for quantifying this phenomenon is the standard deviation of the lengths of the gap-free segments (calculated in base pairs). For the four intervals featured in Fig. 3 *A–D*, the respective computed values are 52, 33, 23, and 27 bp. Comparison of the last two values indicates that the sequence conser-

vation in intron 1 is slightly more irregular than the region immediately upstream of exon 1.

Taken together, these data indicate that there are regions of high sequence conservation present both upstream and within the *CFTR*/*Cftr* gene. Although these deserve more careful follow-up analyses, it is difficult to point to any one(s) as a notably strong candidate for being a regulatory element.

Discussion

The human genomic region containing the *CFTR* gene has been of long-standing interest for intensive study. Specifically, this chromosomal segment was first selected as a model for developing the paradigm of yeast artificial chromosome-based STS-content mapping, resulting in the assembly of a >2.5-Mb yeast artificial chromosome contig containing the *CFTR* gene (58, 59). The long-range physical map of this interval was later refined as a part of a global effort to map human chromosome 7 (50). We have now extended these studies and assembled a complete sequence map of the human *CFTR* region, reflected by the generation of high-accuracy sequence for a contiguous set of 16 BAC/PAC clones that together span ≈1.6 Mb and fully encompass the gene. In addition, we also assembled a similar ≈358-kb sequence map of the corresponding mouse *Cftr* region.

The sequence data reported here provides information about the precise genomic organization of a number of human genes, including *CFTR*, *WNT2*, and the previously unknown genes *HSA_C7orf7* and *HSA_C7orf8*. In the case of *CFTR*, this data gives important insight about the gene's long-range structure. For example, the size of the genomic interval from the *CFTR* translational start site to the end of the cDNA is ≈189 kb, which is ≈60 kb smaller than originally reported (2). This discrepancy likely reflects the prior assessment based on measuring the sizes of large restriction fragments. Similarly, the genomic sequence data allows more precise measurement of *CFTR* intron sizes, thereby refining prior estimates (52). The discovery of two previously unknown genes (*HSA_C7orf7* and *HSA_C7orf8*) flanking *CFTR* is also interesting. Of note, large, CF-causing genomic deletions have not been encountered to date; perhaps this reflects an important role(s) for one or both of these genes, with a heterozygous deletion resulting in a morbid phenotype.

Our studies also provide the ability to perform detailed sequence comparisons between the corresponding human and mouse regions, leading to a number of valuable insights. First, the genomic regions containing *CFTR* and *Cftr* are highly conserved throughout. Although conservation of the coding regions and the ≈10 kb of sequence upstream of exon 1 is not unexpected, the large segments of conserved sequences within the introns is intriguing. For example, PIP analysis reveals several areas of allegedly noncoding DNA with particularly high percent identity values between the human and mouse sequences (e.g., see regions at 94, 340, and 393 kb of the PIP at http://genome.nhgri.nih.gov/chr7/cftr). Similarly, the segments harboring *WNT2* and *Wnt2* contain several striking areas of conserved sequences within the introns flanking exon 3 (Fig. 3*A*).

The presence of discrete regions of highly conserved sequence within noncoding human and mouse DNA may suggest that these segments serve important functional roles (60), such as regulating gene (e.g., *CFTR*/*Cftr*) expression. It is notable that, despite major advances in understanding numerous aspects of CF biology, the mechanisms underlying the control of *CFTR*/*Cftr* expression remain poorly defined. Thus, the above conserved regions represent candidates for possible regulatory elements and deserve more careful study. Interestingly, DNase I hypersensitivity sites have been detected upstream of *CFTR*/*Cftr* (27) as well as within intron 1 (31); there is notably high human–mouse sequence conservation at the expected locations of these sites (see http://genome.nhgri.nih.gov/chr7/cftr), especially for the one in intron 1 (near position 210 kb in Fig. 3*D*). In addition, preliminary data indicates that the conserved regions in *WNT2*/*Wnt2* affect the expression level of a reporter construct when transfected into mammalian cells (L. Mei and R. C. Hardison, personal communication). Whether these or the other conserved sequences function to regulate *WNT2*/*Wnt2*, *CFTR*/*Cftr*, or another gene(s) in the area remains to be established.

The human–mouse genomic sequence comparison reported here highlights potential limitations of this approach for finding regulatory elements. For example, recognition sites for transcription factors may be as small as 6–8 nucleotides and, therefore, not detectable by PIP analysis. Another possibility is that, because the rate of fixation of mutations at this particular genomic locus is relatively slow, there may not have been sufficient time to accumulate sequence differences between humans and mice within nonfunctional segments; thus, some highly conserved segments may not be functionally important. Toward that end, it would be valuable to also generate the sequence of the *CFTR* region in more distantly related animal species (e.g., rat, cow, chicken) for comparative purposes. Alternatively, the mouse may be too divergent from human in terms of *Cftr* expression, relying on different regulatory sequences. Nonetheless, generating the mouse *Cftr* sequence will likely enhance studies of CF animal models. A number of mouse strains have been generated with defined *Cftr* mutations (12–15); although these animals show some characteristic CF symptoms, pancreatic insufficiency does not occur, and lung disease is minimal (17). These findings may be explained by differences in *Cftr* expression in mouse tissues. The distribution of submucosal glands in the lung is different in mouse compared with human, and *Cftr* expression in the fetal lung is notably lower in the mouse (17). In addition, the mouse pancreas does not express *Cftr* at high levels, again different from the human (33). Thus, the differences seen between humans and mice harboring *CFTR*/*Cftr* mutations may partly reflect differences in their tissue-specific expression of the gene. The sequence data reported here may help to clarify the mechanisms underlying such differences as well as to enhance other studies investigating the regulation of *CFTR*/*Cftr* expression, such as those involving the use of *CFTR*/*Cftr*-containing yeast artificial chromosome constructs (61, 62) (P. J. Mogayzel, Jr. and M. A. Ashlock, personnel communication).

In summary, we have generated just under 2 Mb of mammalian DNA sequence from a medically important genomic region. Our efforts have provided the definitive genetic blueprint for the extensively studied *CFTR* gene as well as several flanking genes. The additional infrastructure provided by knowledge of the complete *CFTR* and *Cftr* sequences should contribute to myriad studies of CF genetics and pathophysiology, including ongoing efforts to develop pharmacological and gene therapy-based treatments. In addition, our comparative analyses have facilitated the detection and characterization of previously unidentified genes residing on either side of *CFTR*/*Cftr* and yielded some tantalizing clues about evolutionarily conserved sequences in the region, including those that they may serve a role in the regulation of *CFTR*/*Cftr* expression. Our studies in conjunction with an increasing number of examples in which homologous regions of human and mouse DNA have been sequenced (54, 60, 63–67) are beginning to illustrate the power of comparative sequence analysis for understanding genome structure and function. Such efforts come at an exciting time in the Human Genome Project, as large amounts of human sequence are actively being generated (35) and an earlier-than-anticipated program to sequence the mouse genome has been launched (68).

We thank Drs. Jim Thomas, Bill Pavan, and Melissa Ashlock for critical review of this manuscript. We thank the numerous members of the Washington University Genome Sequencing Center and the National

Institutes of Health Intramural Sequencing Center for their dedicated work in generating the sequence data reported here. These studies were supported in part by funds provided to the Washington University Genome Sequencing Center by the National Institutes of Health for human genome sequencing, Grant LM05110 to W.M. from the National Library of Medicine, and funds to E.D.G. and the National Institutes of Health Intramural Sequencing Center by the Cystic Fibrosis Foundation for mouse genome sequencing.

1. Welsh, M. J., Tsui, L.-C., Boat, T. F. & Beaudet, A. L. (1995) in *The Metabolic and Molecular Bases of Inherited Disease*, eds. Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D. (McGraw–Hill, New York), pp. 3799–3876.
2. Rommens, J. M., Iannuzzi, M. C., Kerem, B., Drumm, M. L., Melmer, G., Dean, M., Rozmahel, R., Cole, J. L., Kennedy, D., Hidaka, N., *et al.* (1989) *Science* **245,** 1059–1065.
3. Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.-L., *et al.* (1989) *Science* **245,** 1066–1073.
4. Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M. & Tsui, L.-C. (1989) *Science* **245,** 1073–1080.
5. Collins, F. S. (1992) *Nat. Genet.* **1,** 3–6.
6. Collins, F. S. (1995) *Nat. Genet.* **9,** 347–350.
7. Heng, H. H. Q., Shi, X.-M. & Tsui, L.-C. (1993) *Cytogenet. Cell Genet.* **62,** 108–109.
8. Bear, C. E., Li, C. H., Kartner, N., Bridges, R. J., Jensen, T. J., Ramjeesingh, M. & Riordan, J. R. (1992) *Cell* **68,** 809–818.
9. Welsh, M. J., Anderson, M. P., Rich, D. P., Berger, H. A., Denning, G. M., Ostedgaard, L. S., Sheppard, D. N., Cheng, S. H., Gregory, R. J. & Smith, A. E. (1992) *Neuron* **8,** 821–829.
10. Welsh, M. J. & Smith, A. E. (1993) *Cell* **73,** 1251–1254.
11. Riordan, J. R. (1999) *Am. J. Hum. Genet.* **64,** 1499–1504.
12. Dorin, J. R., Dickinson, P., Alton, E. W. F. W., Smith, S. N., Geddes, D. M., Stevenson, B. J., Kimber, W. L., Fleming, S., Clarke, A. R., Hooper, M. L., *et al.* (1992) *Nature (London)* **359,** 211–215.
13. Snouwaert, J. N., Brigman, K. K., Latour, A. M., Malouf, N. N., Boucher, R. C., Smithies, O. & Koller, B. H. (1992) *Science* **257,** 1083–1088.
14. Colledge, W. H., Ratcliff, R., Foster, D., Williamson, R. & Evans, M. J. (1992) *Lancet* **340,** 680.
15. Clarke, L. L., Grubb, B. R., Gabriel, S. E., Smithies, O., Koller, B. H. & Boucher, R. C. (1992) *Science* **257,** 1125–1128.
16. Dorin, J. R. (1995) *J. Inherited Metab. Dis.* **18,** 495–500.
17. Harris, A. (1997) *Hum. Mol. Genet.* **6,** 2191–2194.
18. Hagemann, T. (1996) *J. Pediatr. Health Care* **10,** 127–134.
19. Ramsey, B. W. (1996) *N. Engl. J. Med.* **335,** 179–188.
20. Boucher, R. C. (1999) *J. Clin. Invest.* **103,** 441–445.
21. Chou, J.-L., Rozmahel, R. & Tsui, L.-C. (1991) *J. Biol. Chem.* **266,** 24471–24476.
22. Yoshimura, K., Nakamura, H., Trapnell, B. C., Dalemans, W., Pavirani, A., Lecocq, J.-P. & Crystal, R. G. (1991) *J. Biol. Chem.* **266,** 9140–9144.
23. Koh, J., Sferra, T. J. & Collins, F. S. (1993) *J. Biol. Chem.* **268,** 15912–15921.
24. Denamur, E. & Chehab, F. F. (1994) *Hum. Mol. Genet.* **3,** 1089–1094.
25. McDonald, C. D., Hollingsworth, M. A. & Maher, L. J., III (1994) *Gene* **150,** 267–274.
26. Denamur, E. & Chehab, F. F. (1995) *DNA Cell Biol.* **14,** 811–815.
27. Smith, A. N., Wardle, C. J. C. & Harris, A. (1995) *Biochem. Biophys. Res. Commun.* **211,** 274–281.
28. McDonald, R. A., Matthews, R. P., Idzerda, R. L. & McKnight, G. S. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7560–7564.
29. Pittman, N., Shue, G., LeLeiko, N. S. & Walsh, M. J. (1995) *J. Biol. Chem.* **270,** 28848–28857.
30. Matthews, R. P. & McKnight, G. S. (1996) *J. Biol. Chem.* **271,** 31869–31877.
31. Smith, A. N., Barth, M. L., McDowell, T. L., Moulin, D. S., Nuthall, H. N., Hollingsworth, M. A. & Harris, A. (1996) *J. Biol. Chem.* **271,** 9947–9954.
32. Vuillaumier, S., Dixmeras, I., Messai, H., Lapoumeroulie, C., Lallemand, D., Gekas, J., Chehab, F. F., Perret, C., Elion, J. & Denamur, E. (1997) *Biochem. J.* **327,** 651–662.
33. Crawford, I., Maloney, P. C., Zeitlin, P. L., Guggino, W. B., Hyde, S. C., Turley, H., Gatter, K. C., Harris, A. & Higgins, C. F. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 9262–9266.
34. The Sanger Centre & the Washington University Genome Sequencing Center (1998) *Genome Res.* **8,** 1097–1108.
35. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. & Walters, L. (1998) *Science* **282,** 682–689.
36. Zenklusen, J. C., Weintraub, L. A. & Green, E. D. (1999) *Neoplasia* **1,** 16–22.
37. Ellsworth, R. E., Ionasescu, V., Searby, C., Sheffield, V. C., Braden, V. V., Kucaba, T. A., McPherson, J. D., Marra, M. A. & Green, E. D. (1999) *Genome Res.* **9,** 568–574.
38. Yorifuji, T., Lemna, W. K., Ballard, C. F., Rosenbloom, C. L., Rozmahel, R., Plavsic, N., Tsui, L.-C. & Beaudet, A. L. (1991) *Genomics* **10,** 547–550.
39. Tata, F., Stanier, P., Wicking, C., Halford, S., Kruyer, H., Lench, N. J., Scambler, P. J., Hansen, C., Braman, J. C., Williamson, R. & Wainwright, B. J. (1991) *Genomics* **10,** 301–307.
40. Marra, M. A., Kucaba, T. A., Dietrich, N. L., Green, E. D., Brownstein, B., Wilson, R. K., McDonald, K. M., Hillier, L. W., McPherson, J. D. & Waterston, R. H. (1997) *Genome Res.* **7,** 1072–1084.
41. Wilson, R. K. & Mardis, E. R. (1997) in *Genome Analysis: A Laboratory Manual*, eds. Birren, B., Green, E. D., Klapholz, S., Myers, R. M. & Roskams, J. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 1, pp. 301–395.
42. Wilson, R. K. & Mardis, E. R. (1997) in *Genome Analysis: A Laboratory Manual*, eds. Birren, B., Green, E. D., Klapholz, S., Myers, R. M. & Roskams, J. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 1, pp. 397–454.
43. The *C. elegans* Sequencing Consortium (1998) *Science* **282,** 2012–2018.
44. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8,** 175–185.
45. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
46. Gordon, D., Abajian, C. & Green, P. (1998) *Genome Res.* **8,** 195–202.
47. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. (1999) *Genome Res.* **9,** 1–4.
48. Huang, X. Q., Hardison, R. C. & Miller, W. (1990) *Comput. Appl. Biosci.* **6,** 373–381.
49. Schwartz, S., Miller, W., Yang, C. M. & Hardison, R. C. (1991) *Nucleic Acids Res.* **19,** 4663–4667.
50. Bouffard, G. G., Idol, J. R., Braden, V. V., Iyer, L. M., Cunningham, A. F., Weintraub, L. A., Touchman, J. W., Mohr-Tidwell, R. M., Peluso, D. C., Fulton, R. S., *et al.* (1997) *Genome Res.* **7,** 673–692.
51. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
52. Zielenski, J., Rozmahel, R., Bozon, D., Kerem, B., Grzelczak, Z., Riordan, J. R., Rommens, J. & Tsui, L.-C. (1991) *Genomics* **10,** 214–228.
53. Chu, C.-S., Trapnell, B. C., Curristin, S., Cutting, G. R. & Crystal, R. G. (1993) *Nat. Genet.* **3,** 151–156.
54. Endrizzi, M., Huang, S., Scharf, J. M., Kelter, A.-R., Wirth, B., Kunkel, L. M., Miller, W. & Dietrich, W. F. (1999) *Genomics*, **60,** 137–151.
55. Wainwright, B. J., Scambler, P. J., Stanier, P., Watson, E. K., Bell, G., Wicking, C., Estivill, X., Courtney, M., Boue, A., Pedersen, P. S., *et al.* (1988) *EMBO J.* **7,** 1743–1748.
56. Burge, C. & Karlin, S. (1997) *J. Mol. Biol.* **268,** 78–94.
57. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
58. Green, E. D. & Olson, M. V. (1990) *Science* **250,** 94–98.
59. Green, E. D. & Green, P. (1991) *PCR Methods Appl.* **1,** 77–90.
60. Hardison, R. C., Oeltjen, J. & Miller, W. (1997) *Genome Res.* **7,** 959–966.
61. Mogayzel Jr., P. J., Henning, K. A., Bittner, M. L., Novotny, E. A., Schwiebert, E. M., Guggino, W. B., Jiang, Y. & Rosenfeld, M. A. (1997) *Hum. Mol. Genet.* **6,** 59–68.
62. Moulin, D. S., Manson, A. L., Nuthall, H. N., Smith, D. J., Huxley, C. & Harris, A. (1999) *Mol. Med.* **5,** 211–223.
63. Oeltjen, J. C., Malley, T. M., Muzny, D. M., Miller, W., Gibbs, R. A. & Belmont, J. W. (1997) *Genome Res.* **7,** 315–329.
64. Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., Gorrell, J. H., Chinault, A. C., Belmont, J. W., Miller, W. & Gibbs, R. A. (1998) *Genome Res.* **8,** 29–40.
65. Jang, W., Hua, A., Spilson, S. V., Miller, W., Roe, B. A. & Meisler, M. H. (1999) *Genome Res.* **9,** 53–61.
66. Koop, B. F. & Hood, L. (1994) *Nat. Genet.* **7,** 48–53.
67. Brickner, A. G., Koop, B. F., Aronow, B. J. & Wiginton, D. A. (1999) *Mamm. Genome* **10,** 95–101.
68. Battey, J., Jordan, E., Cox, D. & Dove, W. (1999) *Nat. Genet.* **21,** 73–75.