The Cystic Fibrosis Gene: A Molecular Genetic Perspective

Lap-Chee Tsui1 and Ruslan Dorfman2

1The University of Hong Kong, Hong Kong, Special Administrative Region, China
2Geneyouin Inc., Maple, Ontario, Canada
Correspondence: tsuilc@hku.hk; Ruslan.dorfman@geneyouin.ca

The positional cloning of the gene responsible for cystic fibrosis (CF) was the important first step in understanding the basic defect and pathophysiology of the disease. This study aims to provide a historical account of key developments as well as factors that contributed to the cystic fibrosis transmembrane conductance regulator (CFTR) gene identification work. A redefined gene structure based on the full sequence of the gene derived from the Human Genome Project is presented, along with brief reviews of the transcription regulatory sequences for the CFTR gene, the role of mRNA splicing in gene regulation and CF disease, and, various related sequences in the human genome and other species. Because CF mutations and genotype-phenotype correlations are covered by our colleagues (Ferec C, Cutting GR. 2012. Assessing the disease-liability of mutations in CFTR. Cold Spring Harb Perspect Med doi: 10.1101/cshperspect.a009480), we only attempt to provide an introduction of the CF mutation database here for reference purposes.

The cloning of the gene responsible for cystic fibrosis (CF) is a classic example of disease gene identification based on genetic linkage analysis. The first genetic analysis for CF could be traced to 1946, when Andersen and Hodges proposed that the disease was caused by a recessive mutation (Andersen and Hodges 1946), but linkage analyses were only briefly applied by Morton and Allen for cystic fibrosis of the pancreas in 1956 (Allen et al. 1956; Morton and Steinberg 1956). In fact, genetic linkage for disease gene identification was not broadly attempted before the discovery of polymorphic DNA, termed “restriction fragment length polymorphisms” markers in the early 1980s, and the suggestion of their use as markers for genetic mapping (Botstein et al. 1980). Hence, early CF linkage and association studies were conducted with polymorphic biochemical markers, such as immunoglobulins and protein markers (Tsui et al. 1985b; Schmiegelow et al. 1986). Ironically, however, the first CF linkage was detected with a serum enzyme marker, paraoxonase (PON) (Eiberg et al. 1985).

IDENTIFICATION OF CFTR GENE

There have been previous reviews on the identification of the cystic fibrosis gene (Tsui and Buchwald 1991; Tsui 1995). To avoid repeating
all that has been said, we have decided to provide only a historical account of key developments as well as factors that contributed to the CF gene identification.

1. CF was first found linked to PON, but this linkage did not directly result in identification of the CF gene because the chromosome location of PON was not known. Nevertheless, the finding confirmed CF to be a relatively homogeneous genetic disease and suggested that it would be possible to identify the CF gene by genetic linkage analysis with a large number of nuclear (two-generation) families.

2. The first DNA marker found linked to CF was an anonymous marker known as CRI-917 (Tsui et al. 1985a), which was subsequently designated D7S15.

3. Before the chromosome location of D7S15 was eventually published (Knowlton et al. 1985), rumor spread that the CF gene was localized to chromosome 7 (Newmark 1985). This information quickly led to the detection of linkage between CF and several other DNA markers on the long arm of chromosome 7 (Wainwright et al. 1985; White et al. 1985). In other words, the localization of CF to the long arm of chromosome 7 was almost instantly confirmed. However, further gene mapping was limited by the lack of human sequence data (Botstein and Risch 2003; Cardon and Abecasis 2003; Flint et al. 2005) because, essentially for each candidate region, a physical map had to be created, overlapping DNA segments isolated and analyzed for the presence of coding sequences (known as “chromosome walking”), and each gene assessed for its candidacy as the sought-after gene. This general approach was subsequently dubbed “positional cloning” (Collins 1990, 1992). The technique of gene hopping or jumping techniques (Collins et al. 1987) was also used in an attempt to reduce the number of DNA fragments required to be isolated while moving along the chromosome (Kerem et al. 1989a,b; Rommens et al. 1989a,b).

4. The identification of the CF gene was published in a series of three papers in 1989 (Kerem et al. 1989a; Riordan et al. 1989; Rommens et al. 1989a). The gene was named cystic fibrosis transmembrane conductance regulator (CFTR for short). The supporting evidence for the gene was the discovery of the most common mutation in CF, named ΔF508 (now renamed p.F508del). Because this 3-bp deletion was found on a relatively rare extended chromosome haplotype, it was possible to apply allelic association (commonly known as “linkage disequilibrium”) in the fine mapping of the CF gene, and haplotype mapping (ancestral recombination) was proposed as a means to map disease genes by Cox and Chakravarti (Cox et al. 1989).

5. The strong allelic association led to the suggestion that the p.F508del mutation arose from one single event (Cutting et al. 1989; Kerem et al. 1989a). Furthermore, the occurrence of this common mutation in a region with extended linkage disequilibrium was used to form part of the argument for the initial identification of the CFTR gene (Kerem et al. 1989a).

6. It is of interest to note, however, that haplotype analysis was used to assess the genetic heterogeneity of the disease before the cloning of the gene (Tsui and Buchwald 1988). For example, it was used to show that pancreatic-sufficient CF patients had more variable and different haplotypes than pancreatic-insufficient patients (Kerem et al. 1989b).

7. It is also of interest to note that a great deal of effort was used to argue for the 3-bp p.F508del deletion as a bona fide mutation because it was not immediately apparent. Its status was somewhat questionable until the results of subsequent functional analysis of the p.F508del-CFTR protein (Drumm et al. 1991). The status of CFTR as the CF-causing gene was confirmed earlier, however, by the discovery of additional disease-causing mutations (Cutting et al. 1990).

8. Although CF is a relatively homogeneous, monogenic disease with distinct clinical
features, it is possible that a very small number of clinically diagnosed CF patients do not have mutations in the CFTR gene. In other words, some patients may be “phenocopies” of CF; despite their classical CF appearance, the disease in these patients may be caused by mutations in other genes. In support of this possibility, it was reported that mouse models with ENaC channel overexpression showed a phenotype reminiscent of CF (Donaldson and Boucher 2007). In fact, some patients with CF-like disease but mutations in only one copy of their CFTR genes were found to carry gain-of-function mutations in ENaC-encoding subunits (Sheridan et al. 2005; Azad et al. 2009; Fajac et al. 2009).

CDNA CLONING, MUTATION ANALYSIS, AND CFTR GENE STRUCTURE

The initial characterization of the CFTR gene was based on the isolation and alignment of genomic clones with the cDNA clones derived from the T84 human colonic adenocarcinoma cell line and those containing the p.Phe508del mutation isolated from a CF sweat gland cDNA library. Some of corresponding genomic DNA segments were missed because either the exons therein were small and thus refractory to isolation by hybridization or certain exons were skipped in the cDNA clones used for the construction of consensus sequence. Therefore, the CFTR gene was thought to contain 24 exons (Riordan et al. 1989). Subsequently, it became apparent that the gene contains 27 exons (Table 1); exons 6b, 14b, and 17b were missed in the original publication.

It is now well established that the full-length CFTR mRNA contains 6128 nucleotides. Alignment of this sequence with the genomic DNA sequence derived from the Human Genome Project has provided the precise exon–intron structure and intronic sequences of the CFTR gene (Table 1).

Based on the recent sequencing and functional data, the transcription module of CFTR was established to be >216 kb in size (see Fig. 1). The CFTR gene itself spans only 189.36 kb; however, the immediate promoter can be extended as far as 20.9 kb upstream, where the CTCF-dependent insulator binding element is located—the expanded promoter element is located—the expanded promoter element is located—in the regulatory expression unit includes the regulatory binding element required for proper gene expression (Blackledge et al. 2007). An insulator element 6.8 kb downstream from CFTR includes a DHS site with a CTCF binding site followed by additional DHS sites, at +6.8 kb, +7.0 kb, and +15.6 kb, which appear to contribute to DNA looping, limiting the boundary of the CFTR transcriptional unit (Blackledge et al. 2009). Thus, the total size of the CFTR transcription unit between these two insulating elements is ~216.7 kb.

SEGMENTAL DUPLICATION OF THE EXON 9 REGION

One of complications in the CFTR gene structure analysis stemmed from the segmental duplication of the ~30-kb region including exon 9 (Rozmahel et al. 1997). These multiple segmental duplications containing exon 9 and its flanking intron sequences (Fig. 2) have been found in multiple locations across the human genome (Liu et al. 2004). Although how these sequences became amplified with two flanking LINE1 elements is not exactly clear, multiple copies of exon 9-related sequences become problematic when the exon 9 sequence is being examined for mutation analysis, especially when PCR amplification is used (El-Seedy et al. 2009). The pyrimidine track immediately upstream of exon 9 also appears to influence mRNA splicing, causing exon 9 skipping (see below) (Chu et al. 1991, 1993).

REGULATION OF CFTR TRANSCRIPTION

To understand the pathophysiology of CF, one of the approaches is to discover the CFTR expression pattern in different tissue surveys (Gregory et al. 1990; Trezise and Buchwald 1991; Ward et al. 1991). CFTR is found to be expressed in the epithelial cells of a variety of tissues and organs, whose functions are significantly affected in CF patients: lung and trachea, pancreas, liver, intestines, and sweat glands. Low levels of CFTR transcripts can be found in kidney, uterus, ovary,
Table 1. Summary of CFTR coding regions according to historical and current nomenclature

<table>
<thead>
<tr>
<th>Exon 1</th>
<th>Historical exon name</th>
<th>Start (117,120,017)</th>
<th>End (117,120,201)</th>
<th>Length</th>
<th>Exon</th>
<th>Start (117,120,202)</th>
<th>End (117,144,306)</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1</td>
<td>Exon 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 2</td>
<td>Exon 2</td>
<td>117,144,307</td>
<td>117,144,417</td>
<td>111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 3</td>
<td>Exon 3</td>
<td>117,149,088</td>
<td>117,149,196</td>
<td>109</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 4</td>
<td>Exon 4</td>
<td>117,171,168</td>
<td>117,171,419</td>
<td>216</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 5</td>
<td>Exon 5</td>
<td>117,174,420</td>
<td>117,174,420</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 6</td>
<td>Exon 6</td>
<td>117,175,466</td>
<td>117,175,466</td>
<td>164</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 7</td>
<td>Exon 7</td>
<td>117,176,602</td>
<td>117,176,602</td>
<td>126</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 8</td>
<td>Exon 8</td>
<td>117,180,400</td>
<td>117,180,400</td>
<td>247</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 9</td>
<td>Exon 9</td>
<td>117,182,070</td>
<td>117,182,162</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 10</td>
<td>Exon 10</td>
<td>117,188,877</td>
<td>117,188,877</td>
<td>183</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 11</td>
<td>Exon 11</td>
<td>117,199,518</td>
<td>117,199,709</td>
<td>192</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 12</td>
<td>Exon 12</td>
<td>117,227,793</td>
<td>117,227,887</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 14</td>
<td>Exon 14</td>
<td>117,231,988</td>
<td>117,232,711</td>
<td>724</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 15</td>
<td>Exon 15</td>
<td>117,234,984</td>
<td>117,235,112</td>
<td>129</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 16</td>
<td>Exon 16</td>
<td>117,242,880</td>
<td>117,242,917</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 17</td>
<td>Exon 17</td>
<td>117,243,586</td>
<td>117,243,836</td>
<td>251</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 18</td>
<td>Exon 18</td>
<td>117,246,728</td>
<td>117,246,807</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 19</td>
<td>Exon 19</td>
<td>117,250,573</td>
<td>117,250,723</td>
<td>151</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 20</td>
<td>Exon 20</td>
<td>117,251,635</td>
<td>117,251,862</td>
<td>228</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 21</td>
<td>Exon 21</td>
<td>117,254,667</td>
<td>117,254,767</td>
<td>101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 22</td>
<td>Exon 22</td>
<td>117,267,576</td>
<td>117,267,824</td>
<td>249</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 23</td>
<td>Exon 23</td>
<td>117,282,492</td>
<td>117,282,647</td>
<td>156</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 24</td>
<td>Exon 24</td>
<td>117,292,896</td>
<td>117,292,985</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 25</td>
<td>Exon 25</td>
<td>117,304,742</td>
<td>117,304,914</td>
<td>173</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 26</td>
<td>Exon 26</td>
<td>117,305,513</td>
<td>117,305,618</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exon 27</td>
<td>Exon 27</td>
<td>117,306,962</td>
<td>117,308,715</td>
<td>1754</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data are based on the Ensembl release 61–Feb 2011 (http://www.ensembl.org/Homo_sapiens/Transcript/Exons?db=core&g=ENSG0000001626;r=7:117119358-117308719;t=ENST00000003084).
thyroid, and even higher levels in salivary gland and bladder, but the epithelial cell function is not seriously compromised in tissues and organs of CF patients. It is possible that there is sufficient compensation of the missing function by other ion transporters. It is of interest to note that the low levels of CFTR expression in these tissues are driven off an alternative promoter (McCarthy and Harris 2005). These studies have also led to the identification of several tissue-specific splicing isoforms and splicing elements (see below).

The promoter region of the CFTR gene lacks a TATA box and is quite GC rich, although generally it is not methylated. As such, the CFTR promoter has not been well defined. The immediate promoter region contains an important cis-acting element within 632 nucleotides upstream of the translation start site (Lewandowska et al. 2010). The definition of the key promoter regions is also suggested by several mutations that appear to disrupt transcription initiation; for example, the c.-234T>A (-102T>A) mutation was first reported by Claustres and coworkers (Romey et al. 1999). Mutations in more distant regions have been shown to affect the efficiency of transcription: The c.-1750A>G polymorphism reduces the transcription efficiency by >50% (Lewandowska et al. 2010).

The predominant CFTR transcription start site is as described in the original study by Rommens et al. (1989a), meaning that the majority of CFTR transcripts are driven from the key promoter, described in the previous section. However, although the canonical transcripts are found in cells with high CFTR expression, alternative transcription start sites are apparently used in cell lines with low expression levels (McCarthy and Harris 2005). CFTR transcripts from even more distant transcription start sites between 2868 and 2794 can be found in CFPAC and T84 cell lines (McCarthy and Harris 2005).

The immediate promoter region has also been characterized by consensus binding sites for several transcription factors: CTCF, AP-1, SP1, GRE, CRE, C/EBP, and Y-box proteins (McCarthy and Harris 2005). DNase I hypersensitive site ( DHS) mapping has been used to map various putative enhancer sequences.
within the CFTR intragenic regions (Table 2). Presumably, multiple transcription factors can bind to chromatin at these sequences, opening the DNA and extending the physical interactions with the promoter, thereby affecting transcription. HNF1α binding sites, indicative of putative enhancer elements, can be found in multiple locations inside introns 10, 17a, and 20 (Mouchel et al. 2004; McCarthy et al. 2009); it has been shown that RNAi-mediated inhibition of the HNF1α could lead to reduction of CFTR expression. Additional enhancer elements have been located in introns 1 and 11, and HNF1α and p300 are involved in the regulation of CFTR expression (Ott et al. 2009a). However, understanding of the transcriptional control of CFTR gene expression is far from comprehensive.

**SPLICING**

As described in the previous section, the majority of CFTR transcripts start from around the same site. There are, however, transcripts with alternative splicing, resulting in skipping of different exons. Some alternatively spliced species can produce truncated CFTR protein with partial function; for example, an alternatively spliced CFTR transcript missing exon 5 found in heart may produce an active channel albeit with significantly reduced function (Xie et al. 1995).

Many of the alternatively spliced transcripts are expected to introduce translation reading frameshifts, thereby introducing premature translation termination that can induce nonsense-mediated RNA decay, removing the aberrant mRNA species. For some mutations,
however, the resulting mRNA species with exon skipping can persist and produce a dysfunctional protein (Hull et al. 1994). In the case of the c.2491G>T (p.Glu831X, also known as 2623G>T, E831X) variant, the presence of an alternative spliced mutant transcript may explain a mild disease presentation and pancreatic sufficiency in patients carrying this nonsense mutation (Hinzpeter et al. 2010).

Based on studies with minigene constructs, Hinzpeter et al. (2010) have proposed that the c.2491G>T mutation not only introduces a premature stop codon, but also creates a binding site (CAGTAG) for the splicing factor that recognizes the consensus sequence NAGNAG. The study shows that the mutation can lead to the production of three different mRNAs: CFTR_831-873del, CFTR_E831X, and CFTR_E831del, and, unexpectedly, the latter transcript could bypass the premature termination codon and produce a functional p.E831del protein.

### ROLE OF SPlicing REGULATION IN CF DISEASE

Mutations occurring at the intron–exon boundaries with highly conserved sequences, that is,
splicing junctions at $-1$, $-2$, $-3$, and $+1$, $+2$, $+3$ positions are expected to affect splicing of the immediately adjacent exons. The consequence of such mutations, for example, c.1117-1G $>$ A and c.1209+1G $>$ A, are generally considered to be severe, whereas mutations occurring at more distant positions, for example, $+5$, $+6$, or $-5$ and $-6$, are mild, typically associated with mild CF disease. Although the sequences between $+6$ and $+30$ are to some degree conserved in evolution, suggestive of a possible role of them in splicing (Aznarez et al. 2008), there is not much information regarding the severity of mutations found beyond the $+6$ positions, except for c.1117-18G $>$ T, c.1209+18A $>$ C, and c.2909-15T $>$ G, which are frequently found in infertile males.

Some of the mutations are found to affect splicing efficiency, presumably because they alter the binding sites of splicing factors (Aznarez et al. 2003). The most noted example in the CFTR gene is the polypyrimidine track located in intron 8 in front of exon 9; the length of the T-track varies from five to nine Ts (Chu et al. 1993). Although the 5T variant is associated with low efficiency of splicing, the effect by itself is insufficient to cause any known disease. However, if the 5T allele is present in combination with another mild mutation, p.Arg117His (R117H) (Gervais et al. 1993), the combined effect of R117H-5T (reduced protein function and lower abundance of correctly spliced transcript) is apparently sufficient to cause CF, albeit the pancreatic sufficient form (Chu et al. 1993; Massie et al. 2001; Zuccato et al. 2004). Other examples of mutations affecting splicing efficiency include several missense (D648V and T665S) and nonsense (E664X) mutations, presumably due to the disruption of the ESE elements within exon 13 (Aznarez et al. 2003).

It is of interest to note that synonymous variants (or polymorphisms) are much less common in the CFTR gene than in other disease-related genes. The reason for this is unknown, but it may be a property of genomic regions with strong allelic association (or linkage disequilibrium). However, some of these rare synonymous variants, for example, c.2679G $>$ T (p.Gly893Gly), c.2898G $>$ A, c.3204C $>$ T, c.3285A $>$ T, and c.3897A $>$ G (rs1800131), could have an effect on splicing because they occur in the consensus sequences for ESEs (exon splicing enhancers) or ESRs (exon splicing repressors). The exon 12 variant c.2679G $>$ T (p.Gly893Gly), for example, is associated with a mild form of CF; the mutation affects a highly conserved sequence and causes skipping of exon 12 (Pagani et al. 2005; Raponi et al. 2007). In fact, the sequence alteration in exon 12 has been found coupled with a new 5' splice site in exon 15, resulting in a transcript lacking 76-amino-acid codons (Faa et al. 2010). This and other observations bring about a suggestion that some of the synonymous changes can affect CFTR mRNA folding and, in turn, protein synthesis (Bartoszewski et al. 2010). Supporting this view, synonymous mutations have been shown to affect posttranslational protein modification and protein stability and were found to be overrepresented among the “top hits” identified in several Genome Wide Association Studies of common diseases (Chen et al. 2010; Li et al. 2010; Plotkin and Kudla 2010), again showing a crucial, but frequently underestimated, role of synonymous variations in disease pathophysiology.

**TRANS-SPLICING**

Trans-splicing has been described for the CFTR gene, making the initial characterization of the gene structure difficult (Rommens et al. 1989a). It is unclear how often it happens for other genes, but trans-splicing has been suggested as a possible strategy for the correction of CFTR mutations (Wardle and Harris 1995; Mansfield et al. 2000; Buratti and Baralle 2001; Liu et al. 2002). For example, it has been feasible to reconstruct a full-length CFTR transcript with recombinant DNA containing exons 10–24, provided in trans, in transfected cells (Mansfield et al. 2000, 2003; Liu et al. 2002). It remains to be determined, however, whether it is possible to engineer gene replacement therapy by introducing a large CFTR cDNA construct into patient cells. Nevertheless, this approach may be useful for gene repair for rare missense and frameshift mutations, which are
unlikely to benefit from drugs with molecular chaperone properties (see Thomas 2013).

**STUDY OF CFTR IN MICE AND OTHER SPECIES**

Because of the difficulty in studying the *CFTR* gene in human tissues, extensive studies have been devoted to the isolation and characterization of genes homologous to *CFTR* in other species. *CFTR* (ABCC7) is a member of the C subfamily of ABC transporters, which includes sulfonylurea receptors (SURs) and multidrug resistance–associated proteins (MRPs). The ABCC subfamily proteins have two transmembrane membrane-spanning domains (MSDs), each consisting of six transmembrane (TM) segments, although some of the MRP channels have an additional MSD0 (Deeley et al. 2006; Toyoda et al. 2008). All ABC proteins also have two nucleotide-binding domains (NBDs), but only CFTR has the novel R domain (CFTR protein structure and function) (for review, see Hunt et al. 2013; Hwang and Kirk 2013). Better understanding of CFTR function may derive from studies of orthologous genes in other species. One line of such investigation has been attempts to search for naturally occurring *CFTR* gene mutations in various animal species, including sheep, guinea pigs, ferrets, dogs, and even chickens (Tebbutt 1995). Unfortunately, although some sequence variations have been identified, none of them appear to correlate with any CF-like disease.

Because mouse is the best-studied experimental animal for human diseases, several *CFTR*-deficient murine lines were created with the use of recombinant DNA technologies (Snouwaert et al. 1992). These include simple knockout (KO) and specific gene replacements on various genetic strains of mice, providing a spectrum of clinical phenotypes (Davidson and Rolfe 2001; Guilbault et al. 2007; Wilke et al. 2011). Despite undetectable CFTR transcripts in all the complete KO strains—*Cftr<sup>tm1Unc</sup>, *Cftr<sup>tm1Cam</sup>, *Cftr<sup>tm1Hsc</sup>, *Cftr<sup>tm3Bay</sup>, *Cftr<sup>tm5Ush</sup>—the survival varied from <5% to 40%. Furthermore, survival in the hypomorphic strains *Cftr<sup>tm1Hgu</sup>* and *Cftr<sup>tm1Bay</sup>* was largely unaffected, reminiscent of the residual CFTR function for some CF mutations and disease severity in humans, particularly with respect to pancreatic status and extent of CF lung disease (Wilke et al. 2011). The *Cftr<sup>tm1Ksh</sup>, *Cftr<sup>tm2Cam</sup>, and *Cftr<sup>tm1Eur</sup>* mice were created to study the p.Phe508del mutation, whereas other lines, such as *Cftr<sup>tm1G551D</sup>* for other missense mutations found in CF patients. Despite intestinal obstruction and reduced weight, the symptoms observed in CFTR-deficient mice appear to be very different from those of the human disease, particularly in terms of pulmonary involvement.

The availability of CFTR-deficient mice has, however, allowed the study of possible modifying genes for CF, because the same CFTR mutation(s) can result in different disease presentations among different CF patients. Variability in phenotypic manifestations was also observed in mice, and it appeared to be largely dependent on the genetic background of the knockout mice (Rozmahel et al. 1996). Indeed, these differences were used to map the *CFM1* and several other genetic modifier loci associated with varying “disease severity” in CFTR-deficient mice (Rozmahel et al. 1996; Haston et al. 2002a,b; Haston and Tsui 2003).

The CFTR-deficient pig (Rogers et al. 2008 a,b) appears to be a very promising animal model for CF because the mutant animals essentially recapitulate all of the CF-specific gastrointestinal abnormalities and the critical aspects of CF lung disease (Meyerholz et al. 2010). For example, meconium ileus is fully penetrant in the mutant animals: All of the newborn CFTR-deficient piglets suffer from intestinal obstruction and defects in gallbladder development. However, there are differences in disease severity between the completely null CF pigs (Rogers et al. 2008a,b) and the p.Phe508del pigs (Ostedgaard et al. 2011).

Ferrets with *CFTR* gene knockout may provide yet another promising animal model for CE. The CFTR-deficient animals have been found to show many of the characteristics of human CF disease, including defective airway chloride transport and submucosal gland fluid secretion;
meconium ileus with variable penetrance (75% MI); pancreatic, liver, and vas deferens diseases; and a predisposition to lung infection in early life (Sun et al. 2010).

CROSS-SPECIES STUDIES OF CFTR SEQUENCES

Study of sequences conserved across multiple species may identify potential regulatory sequences important for CFTR gene expression and splicing. Such information may be obtained from the ENCODE project (ENCylopedia Of DNA Elements), which has been developed to identify conserved regions and putative gene expression regulatory regions; CFTR is one of the several candidate genes included in the study (ENCODE Project Consortium 2004; Birney et al. 2007). The ChIP-seq method that involves DNA digestion and chromatin immunoprecipitation followed by sequencing was used to identify the transcription factor and histone binding sites. RNA sequencing was used for more accurate gene and exon mapping. DNA methylation and other DNA modifications were assessed in multiple cell types. The conserved regions and functional DNA modification and protein–DNA interaction sites identified by ENCODE (http://encodeproject.org) are compared with the CFTR variants recorded in the Cystic Fibrosis Mutation Database in the PhenCode browser (http://globin.bx.psu.edu/phencode/) (Giardine et al. 2007).

However, not many CFTR mutations have been found associated with these putative regulatory regions thus far (Ott et al. 2009b). There are several possible explanations for the apparent paucity of mutations detected in these regions. First, the biological significance of these sequences has not been investigated. Second, mutation screening studies in CF patients have been primarily concentrated in the coding regions, including at most the immediate flanking intronic regions. Third, mutations in regulatory regions may only affect the expression levels of CFTR, resulting in very mild disease, not recognized as CF.

Identification of crucial regulatory sequences that can direct proper tissue specificity may be critically useful in the construction of CFTR expression vectors for CF gene therapy because, although CFTR deficiency could be rescued by overexpression of wild-type cDNA (Drumm et al. 1990), expression of the gene in atopic tissues could lead to adverse side effects. One study showed that atopic expression of CFTR in the heart could cause arrhythmia and sudden death of transgenic animals (Ye et al. 2010). Therefore, most of the CF gene therapy studies have only used heterologous promoters that are well characterized. Among these, the KRT18 promoter appears to be promising in targeting CFTR expression in the lung epithelial cells (Koehler et al. 2001).

CF MUTATION DATABASE AND MUTATION NOMENCLATURE

The Cystic Fibrosis Genetic Analysis Consortium was formed to facilitate CFTR mutation analysis, at a time before the gene structure was fully characterized. A set of general agreements and guidelines was agreed upon at the North American CF Conference in Florida on October 13, 1989. Briefly, members of the Consortium were encouraged to report data to the Consortium before submission for publication. Because it would usually take at least three months for any publication to come to print, early communication within the group would allow other members to use the information and screen their respective populations in a coordinated and concerted effort. The Consortium also helped disseminate information and ideas regarding novel methods and technical tips that would improve genetic testing. Members agreed that the group that reported the information would hold the priority to publish the primary data. Other members of the Consortium might use the information for their own research before the publication of the original report but might not publish any secondary observations without the consent of the original group.

In addition, a standardized mutation nomenclature system was proposed (Beaudet and Tsui 1993). This was the first time a systematic nomenclature was used to describe human disease mutations. These guidelines were adopted
by the genetic community and applied in the naming of mutations in other genes, and laid the basis for subsequent versions of mutation nomenclature developed by the Human Genome Variation Society.4

After the launching of the Consortium in 1989, its membership quickly grew to include 130 groups of CF laboratories from more than 30 countries. The Consortium exchanged mutation reports and facilitated the discussions and research on genotype–phenotype correlations, confirmed clinical data, and enabled development of genetic screening tools. With the development of Internet technologies, because the data held by the Consortium were essentially open to the entire scientific community, a Web-based Cystic Fibrosis Mutation Database (CFMDB) was launched in 1995. Even to date, it is one of the most cited genetic databases and is visited by more than 500 unique users per day. The Database contains information for more than 1890 disease-causing mutations and variants of different types5 (Tables 3 and 4: mutation distribution summary). Although the changes in HGV nomenclature rules have necessitated a revision (April 2010) of CFMDB, all of the previous (legacy) names for CF mutations are retained as a display option along with new maps and up-to-date numbering systems according to the current knowledge of the gene structure. All new mutation submissions, however, would only appear in the revised HGV format.

Table 3. Summary statistics for classification of CFTR mutations recorded in CFMDB (January 2011)

<table>
<thead>
<tr>
<th>Mutation type</th>
<th>Count</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missense</td>
<td>765</td>
<td>40.41</td>
</tr>
<tr>
<td>Frameshift</td>
<td>303</td>
<td>16.01</td>
</tr>
<tr>
<td>Splicing</td>
<td>225</td>
<td>11.89</td>
</tr>
<tr>
<td>Nonsense</td>
<td>160</td>
<td>8.45</td>
</tr>
<tr>
<td>In frame in/del</td>
<td>37</td>
<td>1.95</td>
</tr>
<tr>
<td>Large in/del</td>
<td>49</td>
<td>2.59</td>
</tr>
<tr>
<td>Promoter</td>
<td>15</td>
<td>0.79</td>
</tr>
<tr>
<td>Sequence variation</td>
<td>269</td>
<td>14.21</td>
</tr>
<tr>
<td>Unknown</td>
<td>70</td>
<td>3.7</td>
</tr>
<tr>
<td>Total</td>
<td>1893</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Mutation distribution summary

CONCLUDING REMARKS

It has often been said that, had the full human genome been mapped and sequenced, the task of finding the cystic fibrosis gene would have taken only a fraction of the time and cost. The identification of the CFTR gene has, however, provided an entry point to understand the basic defect underlying this multisystemic disease, whereby better detection and treatment can be devised for the sake of all of the affected individuals.

In addition to providing the first example of positional cloning, the CF gene cloning work also provided the concept of haplotype mapping, which was illustrated on the cover of the September 8, 1989 issue of Science (see Fig. 3). Although it was not properly introduced in the report of Kerem et al. (1989), the solution for fine-mapping of human traits by haplotype analysis was initially attempted by Cox and Chakravarti (Cox et al. 1989) and later properly introduced by Lander and Kruglyak (Kruglyak 1997, 2005, 2008).

The spirit of sharing and cooperation was exemplified in CF genetic research as early as the CF gene mapping stage; the primary DNA marker data set used in the CF gene mapping study was shared with the human gene mapping community, so that different algorithms on multipoint mapping could be tested (Spence et al. 1989).

After identification of the CFTR gene, crucial information for the detection of additional mutations was shared among members of the CF Genetic Analysis Consortium before the gene structure was entirely elucidated and published (Zielenski et al. 1991). New mutation data were also shared before formal publication as describe in the preceding section.

Most remarkable, however, was the honorable competition and cooperative spirit that could be found even during the race to the cloning of the CF gene. In 1987, the group led by Robert Williamson had used a rather elegant technique and identified a DNA fragment very

Cite this article as Cold Spring Harbor Perspectives in Medicine 2013;3:a009472
Table 4. Distribution of mutations in CFTR protein domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Exon number</th>
<th>Missense</th>
<th>Frameshift/stop</th>
<th>In frame in/del</th>
<th>Splicing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSD1</td>
<td>Exon 4</td>
<td>18</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>MSD2</td>
<td>Exon 4-5</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>MSD3</td>
<td>Exon 6a-6b</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>MSD4</td>
<td>Exon 6b</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>MSD5</td>
<td>Exon 7-8</td>
<td>13</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>MSD6</td>
<td>Exon 8</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>NBD1</td>
<td>Exon 10-13</td>
<td>147</td>
<td>37</td>
<td>3</td>
<td>1</td>
<td>188</td>
</tr>
<tr>
<td>R</td>
<td>Exon 13-14a</td>
<td>58</td>
<td>87</td>
<td>1</td>
<td>1</td>
<td>147</td>
</tr>
<tr>
<td>MSD7</td>
<td>Exon 15</td>
<td>13</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>MSD8</td>
<td>Exon 16-17a</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>MSD9</td>
<td>Exon 17b</td>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>MSD10</td>
<td>Exon 17b</td>
<td>12</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>MSD11</td>
<td>Exon 19</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>MSD12</td>
<td>Exon 20</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>NBD2</td>
<td>Exon 20-24</td>
<td>74</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>139</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>422</td>
<td>271</td>
<td>13</td>
<td>4</td>
<td>710</td>
</tr>
</tbody>
</table>

*Please note that the total number of mutations is smaller because synonymous variants are not shown in the summary; also, different cDNA mutations that lead to identical missense changes and mutations outside of the functionally defined domains are now shown in the summary. Therefore, the total number of missense mutations is substantially smaller than indicated in Table 3.

Figure 3. Schematic diagram illustrating the concept of haplotype mapping. (A) The color bars represent chromosomes from different patients with cystic fibrosis; this diagram was used as the basis for the cover design of the September 8, 1989 issue of Science. (B) The DNA marker haplotypes associated with the CF chromosomes carrying ΔF508; the "1s" and "2s" inside the colored bars on the left represent schematic alleles of DNA markers around the CFTR gene; the original data appeared in Table 3 on p. 1076 of Kerem et al. (1989); the numbers on the right denote the counts of CF and normal (N) chromosomes with the corresponding DNA marker haplotypes shown on the left.
close to the CF gene (Scambler et al. 1986; Estivill et al. 1987). In fact, their data misled them and many others to believe that they had cloned the gene. Due credit must be given to William-son and his colleagues, however, because they were quick to admit to the scientific world that the candidate gene they had isolated was not the right gene (Wainwright et al. 1988). Their timely sharing of this valuable piece of information reversed the funding decision at the National Institutes of Health (of the United States), so that the CF gene cloning work in several laboratories could be continued without any interrup-tion of funding.

We would like to end this work with a quote from a previous review article (Tsui and Estivill 1991):

Finally, there has been tremendous competition in the field of disease gene cloning. . . . It is difficult to describe the feelings of performing repeated searches for clone after clone without an open reading frame, constructing and screening genomic and cDNA libraries one after another, watching the come and go of a candidate gene, and working under the fear that the gene has perhaps been identified by another group . . . although being the first to report appears to be the only consolation of this hard work, one should keep in mind that gene identification is in fact only the first step . . . in understanding a disease. . . .

REFERENCES

* Reference is also in this collection.


L.-C. Tsui and R. Dorfman


Cystic Fibrosis Gene


L.-C. Tsui and R. Dorfman


The Cystic Fibrosis Gene: A Molecular Genetic Perspective

Lap-Chee Tsui and Ruslan Dorfman

Cold Spring Harb Perspect Med 2013; doi: 10.1101/cshperspect.a009472

Subject Collection  Cystic Fibrosis

Antibiotic and Anti-Inflammatory Therapies for Cystic Fibrosis
James F. Chmiel, Michael W. Konstan and J. Stuart Elborn

Structure and Function of the Mucus Clearance System of the Lung
Brenda M. Button and Brian Button

New Pulmonary Therapies Directed at Targets Other than CFTR
Scott H. Donaldson and Luis Galietta

The Cystic Fibrosis Intestine
Robert C. De Lisle and Drucy Borowitz

The Cystic Fibrosis Transmembrane Regulator (ABCC7) Structure
John F. Hunt, Chi Wang and Robert C. Ford

Status of Fluid and Electrolyte Absorption in Cystic Fibrosis
M.M. Reddy and M. Jackson Stutts

The Influence of Genetics on Cystic Fibrosis Phenotypes
Michael R. Knowles and Mitchell Drumm

Perspectives on Mucus Properties and Formation—Lessons from the Biochemical World
Daniel Ambort, Malin E.V. Johansson, Jenny K. Gustafsson, et al.

The Cystic Fibrosis Airway Microbiome
Susan V. Lynch and Kenneth D. Bruce

Cystic Fibrosis Transmembrane Conductance Regulator
P. Andrew Chong, Pradeep Kota, Nikolay V. Dokholyan, et al.

The CFTR Ion Channel: Gating, Regulation, and Anion Permeation
Tzyh-Chang Hwang and Kevin L. Kirk

Assessing the Disease-Liability of Mutations in CFTR
Claude Ferec and Garry R. Cutting

Supramolecular Dynamics of Mucus
Pedro Verdugo

For additional articles in this collection, see http://perspectivesinmedicine.cshlp.org/cgi/collection/

Copyright © 2013 Cold Spring Harbor Laboratory Press; all rights reserved