European Journal of Human Genetics (2006) 14, 567–576

www.nature.com/ejhg

ARTICLE

Gross genomic rearrangements involving deletions in the *CFTR* gene: characterization of six new events from a large cohort of hitherto unidentified cystic fibrosis chromosomes and meta-analysis of the underlying mechanisms

Claude Férec*[,1,2,3,4], Teresa Casals[5], Nadia Chuzhanova[6], Milan Macek Jr[7], Thierry Bienvenu[8], Andrea Holubova[7], Caitriona King[9], Trudi McDevitt[9], Carlo Castellani[10], Philip M Farrell[11], Molly Sheridan[12], Sarah-Jane Pantaleo[13], Ourida Loumi[14], Taieb Messaoud[15], Harry Cuppens[16], Francesca Torricelli[17], Garry R Cutting[12], Robert Williamson[18], Maria Jesus Alonso Ramos[19], Pier Franco Pignatti[20], Odile Raguénès[1,4], David N Cooper[21], Marie-Pierre Audrézet[1,4] and Jian-Min Chen[1,2,3]

[1]*INSERM, U613 (Génétique Moléculaire et Génétique Epidémiologique), Brest, France;* [2]*Université de Bretagne Occidentale, Faculté de Médecine de Brest et des Sciences de la Santé, Brest, France;* [3]*Etablissement Français du Sang – Bretagne, Brest, France;* [4]*CHRU Brest, Hôpital Morvan, Laboratoire de Génétique Moléculaire et d'Histocompatibilité, Brest, France;* [5]*Medical and Molecular Genetics Center-IRO, Hospital Duran i Reynals, Barcelona, Spain;* [6]*Biostatistics and Bioinformatics Unit, Cardiff University, Cardiff, UK;* [7]*Institute of Biology and Medical Genetics-Cystic Fibrosis Center, Charles University, Prague, Czech Republic;* [8]*Laboratoire de Biochimie et Génétique Moléculaire, Hôpital Cochin, Paris, France;* [9]*National Centre for Medical Genetics and Department of Paediatrics, University College Dublin, Our Lady's Hospital for Sick Children, Crumlin, Dublin, Ireland;* [10]*Cystic Fibrosis Centre, Azienda Ospedaliera di Verona, Verona, Italy;* [11]*Department of Pediatrics, University of Wisconsin Medical School, Madison, WI, USA;* [12]*Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA;* [13]*Genetic Health Services Victoria, Royal Children's Hospital, Parkville, Australia;* [14]*Faculté des Sciences Biologiques, Bab-Ezzouar Institut de Biologie, Université des Sciences et de la Technologie, Alger, Algeria;* [15]*Laboratoire de Biochimie, Clinique-Hopital d'Enfants, Tunis, Tunisia;* [16]*Center for Human Genetics, Catholic University of Leuven, Leuven, Belgium;* [17]*Unita' Operativa Citogenetica e Genetica, Azienda Ospedaliera Careggi, Florence, Italy;* [18]*Murdoch Children's Research Institute, Parkville, Australia;* [19]*Laboratorio de Genética-B2-IBGM, Instituto de Biologia y Genética Molecular, Valladolid, Spain;* [20]*Section of Biology and Genetics, University of Verona, Verona, Italy;* [21]*Institute of Medical Genetics, Cardiff University, Cardiff, UK*

Gross genomic rearrangements involving deletions in the *CFTR* gene have recently been found to account for ~20% of unidentified cystic fibrosis (CF) chromosomes in both French and Italian patients. Using QMPSF and walking quantitative DHPLC, six novel mutations (three simple deletions, two complex deletions with short insertions of 3–6 bp, and a complex deletion with a 182 bp inverted downstream sequence) were characterized by screening 274 unidentified CF chromosomes from 10 different countries. These lesions increase the total number of fully characterized large *CFTR* genomic rearrangements involving deletions to 21. Systematic analysis of the 42 associated breakpoints indicated that all 21 events were caused by nonhomologous recombination. Whole gene complexity analysis revealed a significant

*Correspondence: Professor C Férec, INSERM, U613, Université de Bretagne Occidentale, Etablissement Français du Sang – Bretagne, 46 rue Félix Le Dantec, Brest, F-29220 France. Tel: +33 2 98445064; Fax: +33 2 98430555; E-mail: claude.ferec@univ-brest.fr
Received 24 October 2005; revised 6 January 2006; accepted 10 January 2006; published online 22 February 2006

correlation between regions of low sequence complexity and the locations of the deletion breakpoints. Known recombination-promoting motifs were noted in the vicinity of the breakpoints. A total of 11 simple deletions were potentially explicable in terms of the classical model of replication slippage. However, the complex deletions appear to have arisen via multiple mechanisms; three of the five complex deletions with short insertions and both examples of large inverted insertions (299 and 182 bp, respectively) can be explained by either a model of serial replication slippage in *cis* (SRS*cis*) or SRS in *trans* (SRS*trans*). Finally, the nature and distribution of large genomic rearrangements in the *CFTR* gene were compared and contrasted with those of two other genes, *DMD* and *MSH2*, with a view to gaining a broader understanding of DNA sequence context in mediating the diverse underlying mutational mechanisms.
European Journal of Human Genetics (2006) **14,** 567–576. doi:10.1038/sj.ejhg.5201590; published online 22 February 2006

Keywords: breakpoint; *CFTR*; deletion; gross genomic rearrangements; mutation; mutational mechanisms

Introduction

Cystic fibrosis (CF; MIM #219700) is a very common life-shortening autosomal recessive disorder in Caucasians. Mutations in the cystic fibrosis transmembrane conductance regulator gene (*CFTR/ABCC7*; MIM #602421) are responsible for a broad spectrum of clinical phenotypes ranging from severe CF to male sterility due to congenital bilateral aplasia of the vas deferens (CBAVD; MIM #277180), bronchiectasis (MIM #211400), and idiopathic chronic pancreatitis (MIM #167800).

The *CFTR* gene, which was characterized some 15 years ago,[1] represents one of the most extensively studied human disease genes. To date, >1300 different *CFTR* gene lesions have been deposited in the Cystic Fibrosis Mutation Database (www.genet.sickkids.on.ca/cftr/). The vast majority of these mutant alleles are either single base-pair substitutions or microinsertions/deletions, with a 3-bp deletion that results in the loss of a phenylalanine at amino-acid position 508 (F508del) accounting for about two third of all *CFTR* mutations worldwide.[2,3] A further 10–20 mutations are present at a frequency of >0.1% whereas the remaining lesions are confined to a relatively small proportion of patients and may even be found in single individuals.[3]

The above notwithstanding, a significant proportion of CF alleles remain to be identified in most of the studied populations.[4] This is not merely due to ethnogeographic differences in the distribution of CF alleles and the different mutation detection methodologies employed. Rather, it is clear that some gross rearrangements of the *CFTR* gene are refractory to analysis by conventional PCR-based methods. Having performed the first systematic screen for such mutations in the *CFTR* gene using quantitative multiplex PCR of short fluorescent fragments (QMPSF), we found that some 16% of previously unidentified CF chromosomes (after extensive and complete screening of the gene by both denaturing gradient gel electrophoresis[5,6] and denaturing high-performance liquid chromatography (DHPLC))[7] carried a gross deletion-containing rearrangement of the *CFTR* gene.[8] These findings have now received broad support from additional studies.[9–12]

Gross genomic rearrangements of the *CFTR* gene comprise ~1.5% of known *CFTR* gene lesions (Human Gene Mutation Database; http://www.hgmd.org).[13] These often complex mutations exhibit extensive allelic heterogeneity and arise through the action of diverse mutational mechanisms. To obtain further insights into these findings, we have extended our search for large genomic rearrangements to CF chromosomes with hitherto unidentified *CFTR* gene lesions obtained from 10 different countries including Australia, Algeria, Belgium, Czech Republic, France, Ireland, Italy, Spain, Tunisia and the USA.

Materials and methods

Recruitment of unidentified CF chromosomes

A total of 274 chromosomes were recruited from 10 countries through 15 different laboratories: Australia (four), Algeria (29), Belgium (three), Czech Republic (50), France (12), Ireland (57), Italy (four), Spain (28), Tunisia (51), USA (36). All chromosomes were derived from CF patients but had not been found to carry any known *CFTR* mutations after screening the coding regions by DGGE and/or DHPLC.

QMPSF analysis and molecular characterization of the genomic rearrangements

Mutation detection and characterization were performed as previously described.[8]

Mutation nomenclature

All newly identified mutations were named in accordance with the standard nomenclature guidelines proposed by the Human Genome Variation Society (http://www.

hgvs.org/; ie cDNA-based numbering with the A of the ATG translational initiation codon as +1). In addition, for the purpose of easily locating the breakpoints, conventional nomenclature using IVS+ or – was also provided. The annotated genomic sequence of the *CFTR* gene deposited in the Cystic Fibrosis Mutation Database (http://www.genet.sickkids.on.ca/cftr/) was used as the reference sequence.

Collation of previously characterized gross rearrangements involving gross deletions of the *CFTR* gene

All fully characterized gross *CFTR* deletions reported in the literature were collated for analysis.

Computer-assisted sequence analysis

DNA sequence ±500 bp to each deletion breakpoint was searched for both low complexity/simple repeats and interspersed repeats by the *RepeatMasker* program available at http://www.repeatmasker.org. Sequence similarity between the ±500 bp flanking the 5′ breakpoint and the ±500 bp flanking the 3′ breakpoint of each deletion was compared wherever possible using the *BLAST 2 sequences* tool available at http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi. Both programs were used with default parameters.

The occurrence of 142 specific motifs of length ≥5 bp, known to play a role in the breakage and rejoining of DNA molecules (partially listed in Abeysinghe *et al*[14]), 17 deletion/insertion 'super hot spots' associated with microdeletions/microinsertions,[15] and the indel hot spot[16] were sought in the vicinity (±25 bp) of all breakpoint junctions. Finally, *complexity analysis*[17] was used to assess the regularity of the genomic *CFTR* gene sequence in relation to the positions of the deletion breakpoints.

Results and discussion

Characterization of six novel gross *CFTR* genomic rearrangements involving deletions

Using previously established techniques,[8] we characterized six novel large *CFTR* genomic rearrangements involving deletions (Figure 1) from the 274 CF chromosomes with hitherto unidentified *CFTR* gene lesions: IVS1-5842_IVS4+401del33104 in three apparently unrelated Irish patients; IVS16-449_IVS18+644del5288 in one French patient; IVS19-24_IVS20+601del781 in three Spanish patients; IVS16-908_c.3085del1005insGACAG in one French patient; c.4344_Stop+486del585insTTG in one Spanish patient; and IVS1-5811_IVS2+2186del8108ins182 in one Czech patient. In addition, we identified a complete deletion of the *CFTR* gene in an Italian patient but have been unable to characterize its breakpoints. Furthermore, some previously known large deletions were also found among these chromosomes (data not shown).

New insights into the mutational mechanisms underlying large *CFTR* genomic rearrangements involving deletions obtained by meta-analysis

The six novel deletions reported here have lent further support to the notion that large *CFTR* genomic rearrangements manifest extensive allelic heterogeneity.[8] More importantly, the addition of these new lesions has increased the total number of fully characterized large *CFTR* genomic rearrangements involving deletions from 15 to 21 (Figure 2, Table 1). The availability of a total of 42 independent breakpoints made it possible to perform a meta-analysis of the large genomic rearrangements that have occurred at the *CFTR* locus.

None of the 21 characterized large* CFTR *genomic rearrangements appear to have been generated by homologous recombination Large genomic rearrangements may be classified as being due either to homologous or nonhomologous recombination, based upon the presence or absence, respectively, of significant nucleotide sequence similarity between the parental sites of recombination. In this regard, the minimal efficient processing segment (MEPS), which describes the minimum length of sequence identity between two homologous sequences required for efficient homologous recombination to occur,[18] has been estimated to be between 337 and 456 bp in humans.[19] Consistent with this estimate, full-length (together with their poly(A) tails) *Alu* sequences (which comprise >10% of the human genome sequence[20] and have often been found to mediate gross deletions causing human genetic disease through homologous recombination)[21] have a length of >300 bp.[22] We have shown that *Alu*-mediated homologous recombination is unlikely to be able to account for our previously reported five gross *CFTR* deletions.[8] We have now systematically searched ±500 bp flanking each deletion breakpoint of an additional 16 mutational events and found that only three nonidentical breakpoints resided within *Alu* repeats. In other words, in none of the 21 characterized large *CFTR* genomic rearrangements were homologous *Alu* repeats present at both the 5′ and 3′ breakpoints.

Other interspersed repeats such as LINE-1 and SINE/MIR were also found to occur in the vicinity of certain breakpoints but again no examples of homologous repeats being present at both the 5′ and 3′ breakpoints of a given mutational event were noted. Furthermore, since none of the 21 characterized large deletions exhibited any significant sequence similarity between their 5′ and 3′ breakpoints, homologous recombination may be effectively excluded as the underlying mutational mechanism in these cases.

Known recombination-promoting motifs are often present in the vicinity of the* CFTR *deletion breakpoints Nonhomologous recombination can be promoted by common sequence features or motifs. We have thus investigated the

Figure 1 The six newly identified gross *CFTR* genomic rearrangements involving deletions. Wild-type sequences spanning the 5′ and 3′ breakpoints of each rearrangement (deleted sequences are barred), irrespective of whether they are simple or complex deletions, are provided. Coding sequences are in upper case. Short direct repeats, which are present only at the 5′ and 3′ breakpoints of the three simple deletions, are underlined. The sequences of the short insertions found in the first two of the three complex deletions are GACAG and TTG, respectively, as indicated by the relevant mutational nomenclature. The 182 nucleotides inserted as part of the remaining complex deletion correspond to a downstream sequence tract (from IVS3 + 6780 to IVS3 + 6961) in inverted orientation (boxed). Note also that the TTG insertion in c.4344_Stop + 486del585insTTG corresponds to a downstream trinucleotide, caa, in inverted orientation (boxed). These two complex deletions that contain inverted downstream sequences are potentially explicable by the model of intrachromosomal serial replication slippage in *trans* (SRS*trans*).[33] The different pairs of short inverted repeats (shaded or shaded in bold) thought to mediate SRS*trans* in the two cases are indicated. *CFTR* genomic sequence deposited in the Cystic Fibrosis Mutation Database (http://www.genet.sickkids.on.ca/cftr/) was used as the reference sequence.

occurrence of 142 specific motifs of length 5 bp known to play a role in the breakage and rejoining of DNA molecules and partially listed in Abeysinghe *et al*,[14] 17 deletion/insertion 'super hot spots' associated with microdeletions/microinsertions[15] and an indel hot spot[16] in the vicinity (±25 bp) of the 42 breakpoint junctions. The most frequently encountered motifs are listed in Table 2.

Alternating purine/pyrimidine and polypurine tracts have been found to be significantly over-represented in the vicinity of gross deletions.[14] These sequences, as well as polypyrimidine tracts, are known to form non-B DNA structures and to induce gross deletions and other forms of genomic instability.[23] Intriguingly, polypurine tracts were found to be the most frequently encountered motifs in the vicinity of the *CFTR* deletion breakpoints. The second most frequent motif was reminiscent of immunoglobulin heavy chain class switch repeats, sequences that might, therefore, facilitate double strand breaks.[24] Other motifs, such as WGGAG and its complement CTCCW (associated with replication fork arrest[25]) and the deletion hot spot consensus TGRRKM[26] were also found in the vicinity of some breakpoints (Table 2).

The deletion breakpoints exhibit a significant correlation with regions of low complexity Three measures of complexity, with respect to direct repeats, inverted repeats and symmetric elements[17] were used to assess the regularity of the studied genomic *CFTR* gene sequence (comprising all introns and exons of the *CFTR* gene, 19801 bp 5′ to the first exon, and 25051 bp 3′ to the last exon). Position numbering is given from the beginning of this extended sequence. To assess the local sequence regularity, the complexity profile of the sequence was calculated for each of the three measures of complexity by scanning the *CFTR*

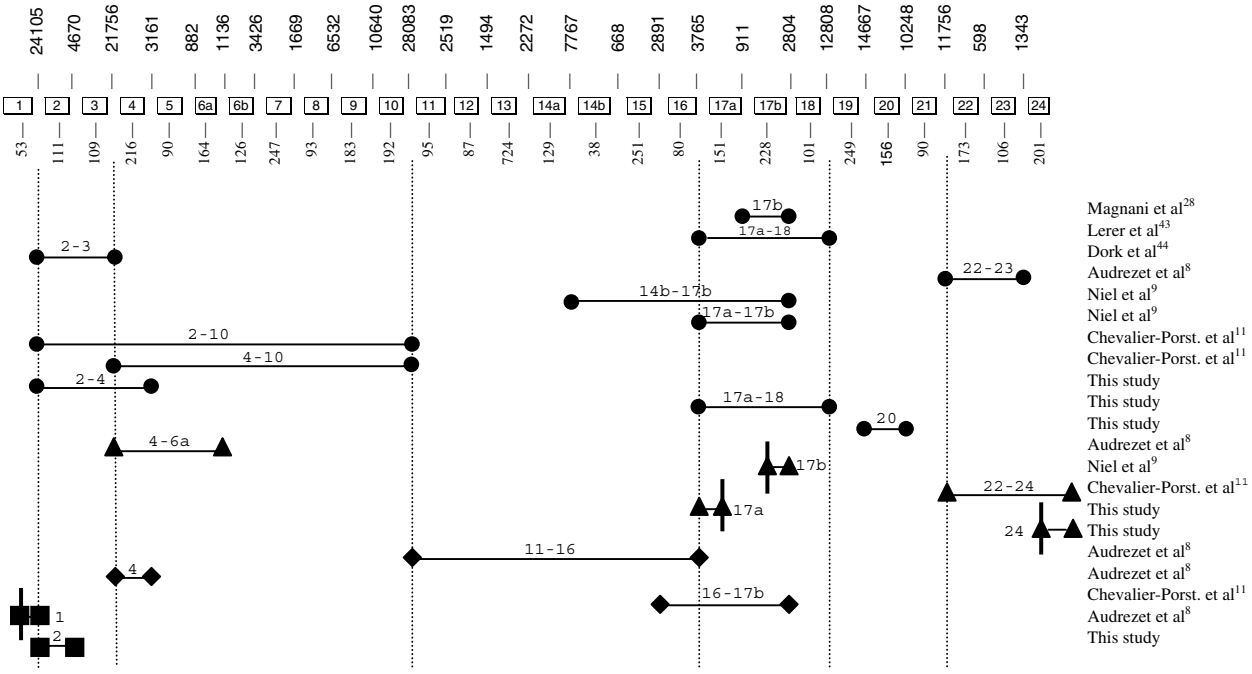

Figure 2 Schematic diagram of fully characterized gross *CFTR* genomic rearrangements involving deletions. *Upper panel* Genomic structure of the *CFTR* gene. Spanning ~189 kb on chromosome 7q31.3,[41] the gene comprises 27 exons[42] and encodes a ~6.5 kb transcript.[1] Numbers above and below denote the sizes (bp) of the introns and exons respectively. *Lower panel* Fully characterized large genomic rearrangements involving deletions of the *CFTR* gene. ●, simple deletions with short direct repeats at 5′ and 3′ breakpoints. ▲, complex deletions with short insertions of 3–6 bp. ◆, complex deletions with small insertions of 32–41 bp. ■, Complex deletions with large insertions of >100 bp. The vertical bars indicate that the breakpoints have occurred within coding sequences. Note that specific PCR reactions for genotyping 12 of the 21 characterized large *CFTR* genomic rearrangements were established (see Table 1).

Table 1 PCR conditions employed for the rapid screening of 12 large *CFTR* genomic rearrangements

Mutation[a]	*Primer sequence (5′ to 3′)*[b]	*Cycling program*[c]	*Size of amplified mutant allele (bp)*
1	F: GCAGGCACCCAGAGTAGT R: GGGATTAACCCCTATGACAGTCT	94°C × 30 s, 57°C × 30 s, 72°C × 2 min	2150
2	F: TGACGTTTTGACAGTTGCACAAGTT R: CCCTGGGCAATGTAGTTTTAG	94°C × 30 s, 57°C × 30 s, 72°C × 5 min	4522
2–4	F: TTAGCAGCAAATAATTAGTTC R: AATACATGCAGAAAAGCAGAGTAG	94°C × 30 s, 57°C × 30 s, 72°C × 30s	419
4	F: GCACAGTAGGATGTTATTTATAT R: CAAAACATTATCAGAGCAATATGA	94°C × 30 s, 57°C × 30 s, 72°C × 30s	232
4–6a	F: GGAGAAGTCCTGGTACTCATGTTG R: TTTTCCATTGCTTCTTCC	94°C × 30 s, 57°C × 30 s, 72°C × 5 min	5072
11–16	F: GCAGAGTACCTGAAACAGGA R: CAGTATTATTATTTTGCAGGGGTAG	94°C × 30 s, 57°C × 30 s, 72°C × 1 min	912
17a	F: ATGTAGAAATATAGAAATTTATTAA R: GAAGGTAACAGCAATGAAGAAGA	94°C × 30 s, 57°C × 30 s, 72°C × 2 min	1236
17a–18	F: TTCTGAATGCGTCTACTGTGATC R: ACCCAGGAAAGGCTACTTGTGCTA	94°C × 30 s, 57°C × 30 s, 72°C × 5 min	5008
20	F: TAAGTCGTATCCACTTTGG R: GGAGTTATGTGTGGACATAGATTC	94°C × 30 s, 57°C × 30 s, 72°C × 2 min	2292
22–23	F: GTGCCATATTCTGTGGGAGC R: AGAGTTGAGCCTCTTCCACAGT	94°C × 30 s, 57°C × 30 s, 72°C × 3 min	3301
22–24	F: CAGCCTGGGCAATAGAGCAAGATT R: TGGCCTTCTATCATTGTGGGAACC	94°C × 30 s, 57°C × 30 s, 72°C × 30s	338
24	F: AAGCATTTGCTGATTGCACAG R: TGAAAACATACCACCATCAACC	94°C × 30 s, 57°C × 30 s, 72°C × 2 min	2028

[a]Nomenclature in accordance with Figure 2.
[b]F, forward; R, reverse.
[c]40 cycles of the indicated program after a denaturation step at 94°C for 3 min using conventional *Taq* DNA polymerase.

sequence using a window of size $W = 100$ bp (Figure 3). Complexity profiles comprise regions of relatively low, high and medium complexity. Regions of low complexity are rich in direct and inverted repeats or symmetric elements and they have the potential to contribute to DNA breakage through formation of slipped structures, cruciforms or triplexes. Regions of medium complexity may or may not form secondary structures whereas fragments corresponding to regions of high complexity are patternless and irregular. Inspection of the generated complexity profiles indicates that the *CFTR* deletion breakpoint junctions tend to occur in regions of relatively low complexity.

The *runs test* was used to assess the significance or otherwise of these findings. A dataset of 10 'quasi-breakpoints' was chosen randomly and the corresponding complexities were combined with the complexities of the known breakpoints. All entries were then arranged in ascending order of their complexities and represented as a sequence of 1's and 2's with each entry from both the

Table 2 Numbers of breakpoints containing DNA sequence motifs or their complements known to be associated with site-specific recombination, mutation, cleavage and gene rearrangement[a]

Motif description	*Motif sequence*	*Number of breakpoints*
Polypurine tract	RRRRR	28
Polypyrimidine tract	YYYYY	24
Alternating purine-pyrimidine tract	RYRYR	24
Immunoglobulin heavy chain class switch repeats	GAGCT, GGGCT, GGGGT, TGGGG, TGAGC	26
DNA polymerase α arrest site	WGGAG	9
TCF-1/LEF-1 (HMG-box)	WWCAAAG	15
U2 splice branch consensus	YNYTRAY	10
Micro-deletion/insertion/indel 'super hotspot'	GGAGAA	8
Deletion hotspot consensus	TGRRKM	15

[a]IUPAC code symbols: R = A/G, Y = C/T, K = G/T, M = A/C, W = A/T, N = A/C/G/T. Each motif was counted once in the case of multiple occurrences in the vicinity of the same breakpoint junction.

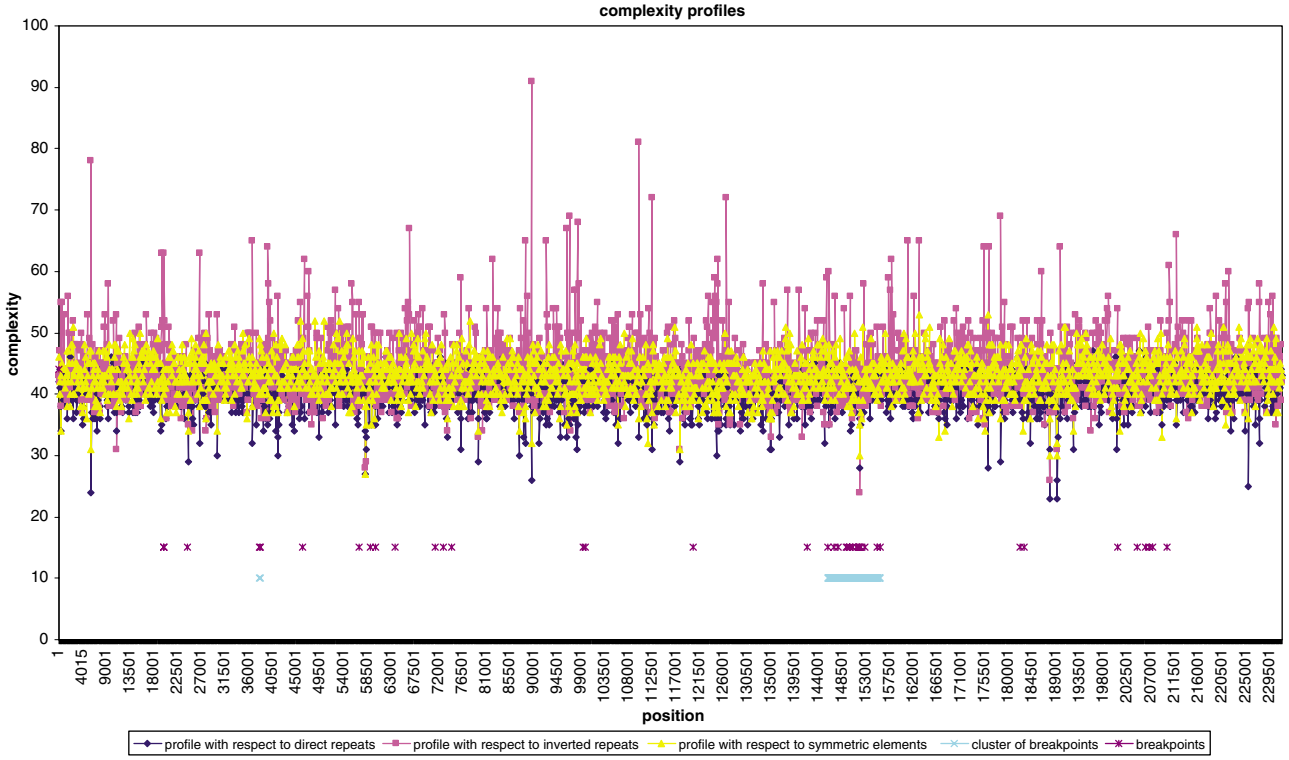

Figure 3 Complexity analysis of the full-length *CFTR* genomic sequence and the occurrence of breakpoints. Complexity profiles were computed with respect to direct repeats, inverted repeats and symmetric elements. The locations of the breakpoints are denoted by 'X'. Two clusters of deletion breakpoints are indicated by solid lines.

randomly chosen and breakpoint datasets being marked by 1 and 2, respectively. Were these two data sets to be significantly different from each other, then one would expect that the complexities of one of the two data sets would be smaller than the other, and therefore that the sequences from one data set would tend to cluster at the beginning (or end) of the sequence of 1's and 2's. The probability of finding this by chance alone can be estimated using *Z*-scores. This process was applied iteratively 10 000 times. The average *Z*-scores were found to be −2.3946, −1.9365 and −1.703 for complexity measures with respect to direct repeat, inverted repeat and symmetric elements with corresponding probabilities, $P = 0.0084$, 0.0262 and 0.0446. This allows us to conclude that the breakpoint junctions have occurred disproportionately in regions that display a relatively low level of sequence complexity, especially with respect to direct repeats.

The order statistics, *r-scans*,[27] were used to assess the extent of clustering of the breakpoint junctions along the extended *CFTR* gene sequence. We assessed whether these breakpoints were evenly dispersed with a Poisson-like distribution throughout the *CFTR* gene sequence, or alternatively whether they were clustered within specific regions. Two clusters, comprising 3 and 16 breakpoints respectively, were noted (see Figure 3). The corresponding probabilities of finding these clusters by chance alone were calculated to be ≤ 0.05 and ≤ 0.01, respectively.

Further delineation of the nature and underlying mutational mechanisms of the 21 characterized large* CFTR *genomic rearrangements Although the above analyses have served to reveal some of the intrinsic sequence features that may have promoted nonhomologous recombination at the *CFTR* locus, they did not provide any information as to how these rearrangements originated. Here, we have addressed this issue by dividing the 21 mutational events into two categories: that is, simple and complex deletions.

Simple deletions Some 52% (11) of the 21 fully characterized *CFTR* mutations are simple deletions. Since short direct repeats are invariably present at both their 5′ and 3′ breakpoints (see Figure 1 eg), they can in principle be explained by the classical model of replication slippage invoking a single cycle of primer–template dissociation and reassociation.[8,9,28,29]

Complex deletions The remaining 10 *CFTR* mutations are complex deletions in which additional sequences have become inserted at the newly formed chromosomal junctions. In sharp contrast to the 11 simple deletions, none of these mutational events exhibited short direct repeats at their 5′ and 3′ deletion breakpoints. Consequently, the classical model of replication slippage cannot account for these complex deletions.

The 10 complex deletions can however be further divided into three categories based upon the length of the inserted sequences. The first category comprises five lesions – IVS3 + 7983_IVS6a + 362del18654insACCTCG,[8] 3413del355insTGTTAA,[9] IVS21-3890_Stop + 3143del9454 insTAACT,[11] and the newly identified IVS16-908_c.3085de l1005insGACAG and c.4344_Stop + 486del585insTTG – all of which contain short insertions of 3–6 nucleotides in, or within the immediate vicinity of, the aberrant chromosomal junctions. This type of short insertion may result from either the untemplated addition or capture of preformed DNA oligonucleotides.[30,31] However, it is possible that some of these short insertions could also have resulted from nascent template-dependent DNA synthesis. In this regard, the six-nucleotide insertion in IVS3 + 7983_ IVS6a + 362del18654insACCTCG and some short insertions of a similar size have been proposed to result from a model of serial replication slippage in *cis* (SRS*cis*)[32] whereas some short insertions could be further explained by a model of SRS in *trans* (SRS*trans*).[33]

Here, we have examined ±50 bp flanking each breakpoint of the remaining four complex *CFTR* deletions with short insertions and found that two of them, IVS21-3890_Stop + 3143del9454insTAACT[11] (Figure 4) and the newly identified c.4344_Stop + 486del585insTTG (Figure 1), are consistent with the SRS*trans* model; both cases involve sequence replacement by an downstream sequence in the near vicinity of the 3′ breakpoint. It is possible that the other two complex deletions, 3413del355insTGTTAA[9] and the newly identified IVS16-908_c.3085del1005insGACAG, could also be explicable by either SRS*cis* or SRS*trans*, if we were to extend the sequence under analysis beyond the ±50 bp flanking each breakpoint. This was not performed, however, because we would necessarily have had a much reduced degree of confidence in the results obtained given the extremely short length of these insertions.

The second category of lesion comprises three complex deletions, for which the length of the inserted sequences ranges from 32 to 41 bp. Unlike the above-mentioned short insertions, these relatively long insertions are unlikely to arise via untemplated DNA incorporation (for a detailed discussion of this topic, see Chen *et al*[29]). Indeed, all three insertions appear to be templated and, interestingly, each probably involves a different mechanism. The 35-bp

gcaatttcctgtttgtttgt••••••aagctcactatgccaaagggcaagttaagctt

Figure 4 IVS21-3890_Stop + 3143del9454insTAACT.[11] The TAACT insertion corresponds to a downstream sequence tract, agtta (boxed), in inverted orientation. The generation of this complex mutation is explicable by the model of SRS*trans*.[33] The different pairs of short inverted repeats (shaded or shaded in bold) thought to mediate SRS*trans* are indicated.

insertion in (IVS10 + 10T > C; IVS10 + 12_IVS16 + 403del47.5kbins35bp) represents a duplication of the 35 nucleotides immediately downstream of the 3′ breakpoint.[8] Given that these 35 nucleotides comprise a pair of inverted repeats, we considered this complex mutation to be formed by a stem-loop structure which induced staggered cleavage followed by subsequent repair and replication.[8] The IVS3−5938_IVS4 + 2011del8165bpins41bp[8] was highly unusual in that it involved the insertion of a 41 bp sequence with partial homology to a retrotranspositionally-competent LINE-1 element. The insertion of this ultra-short LINE-1 element (dubbed a 'hyphen element'[8]) may constitute a novel type of mutation associated with human genetic disease. However, the origin of the 32-bp insertion in IVS15−636_IVS17b−1611del6965ins32[11] is unclear; 24 nucleotides (positions 3–26) are identical to the ±12 bp flanking the 3′ breakpoint of the mutation (ie gggtccaactgc/agtctactctgc).

The third category of mutation comprises two complex deletions (ie c.4_IVS1 + 69del119bpins299bp[8] and the newly identified IVS1−5811_IVS2 + 2186del8108ins182 (Figure 1)), both of which contain quite large insertions. Coincidentally, both insertions represent a downstream sequence in inverted orientation and both are explicable in terms of the intrachromosomal SRS*trans* model.[33]

Comparison of mutational mechanisms underlying characterized large genomic rearrangements in the *CFTR* gene and other genes

Recently, the detection rate of disease-causing large genomic rearrangements has increased significantly thanks to the availability of quantitative multiplex PCR-based techniques. However, most of these large genomic rearrangements have not been fully characterized, as is exemplified by three recent studies involving the *STK11*[34] (MIM #602216), *RB1*[35] (MIM #180200), and *DMD*[36] (MIM #300377) genes. This notwithstanding, studies over the last two decades have led to the characterization of a significant number of large genomic rearrangements causing human genetic disease. In this regard, the Gross Rearrangement Breakpoint Database (GraBD; http://archive.uwcm.ac.uk/uwcm/mg/grabd) currently contains 397 breakpoints from 90 different genes, of which 104 were derived from large deletions and 116 from large deletions with short insertions (note that ~2/3 of these are of somatic origin).[37] A survey of GraBD, together with a perusal of the recent literature suggests that, in the context of fully characterized large inherited gene rearrangements, *CFTR* is probably the best-studied gene with the possible exceptions of *DMD* and *MSH2* (MIM #120435). We have thus sought to compare the nature and distribution of large genomic rearrangements in these three genes, with a view to improving our understanding of the diverse mutational mechanisms that operate upon them.

CFTR vs MSH2 As with gross genomic rearrangements in the *CFTR* gene, those in the *MSH2* gene also show extensive allelic heterogeneity. However, the majority of *MSH2* gene deletions encompass exon 1 (Charbonnier *et al*[38] and references therein). The characterization of 17 large *MSH2* genomic rearrangements, all involving an exon 1 deletion, revealed that up to 15 cases may have resulted from *Alu*-mediated homologous recombination.[38] This high frequency appears to be due to a remarkably high density of *Alu* repeats in the 5′ region of the *MSH2* gene; indeed, this region contains three to four times more *Alu* sequences than the average manifested by the spatially matched regions of 24336 human genes.[38] By contrast, none of the 21 *CFTR* large genomic rearrangements can be explained by *Alu*-mediated homologous recombination. Interestingly, analysis using *RepeatMasker* revealed that *Alu* sequences (10 883 bp) account for only 5.8% of the *CFTR* gene sequence (189 014 bp from position −1000 upstream of the translational initiation codon to +1000 downstream of the translational stop codon), significantly lower than the fraction (10.6%) of *Alu* sequences present in the human draft genome sequence.[20]

CFTR vs DMD An unusual feature of DMD is that deletions of one or more exons in the *DMD* gene are found in ~65% of cases (Lalic *et al*[36] and references therein). Despite extensive heterogeneity in terms of both deletion size and location, two hot spots have been identified. A study of 20 *DMD* deletion junctions involving the major hot spot exons (40–50) has revealed that although all the deletions were presumed to have resulted from nonhomologous recombination, no sequence elements including minisatellite core sequences, *Chi* elements, translin-binding sites, *Pur* elements, matrix attachment regions, and motifs conferring sequence-dependent DNA curvature and duplex stability known to be involved in illegitimate recombination, were found to be significantly associated with the deletion breakpoints.[39] Further, of the 20 deletion events, six (30%) were simple deletions occurring in the absence of short direct repeats which can thus only be accounted for by a model of nonhomologous end joining.[39] Interestingly, in another study,[40] four of the 14 deletion events contained duplicational junctions ranging from 9 to 24 bp; three of these (junctions 9, 10, and 12) can be explained by SRS mutational models (data not shown).

Conclusions

In summary, through an international collaborative effort, we have characterized six novel large *CFTR* genomic rearrangements involving deletions. These lesions, when evaluated together with those previously reported, have increased our knowledge of the diverse nature and mechanisms of these mutational events at the *CFTR* locus. We have also, for the first time, performed a whole-gene

complexity analysis and observed a significant correlation between the locations of the deletion breakpoints and regions of low sequence complexity. This type of analysis would appear to be worth repeating in other systems in order to explore its possible generality.

Acknowledgements

JM Chen is a visiting Professor of Genetics supported by the Ministère de la Jeunesse, de l'Éducation Nationale et de la Recherche, France. This work was supported by the INSERM (Institut National de la Santé et de la Recherche Médicale) and the VLM (Vaincre La Mucoviscidose), France; the Grant FIS/FEDER PI020099 (to T Casals); and the grant VZ FNM 00064203 (to M Macek).

References

1 Riordan JR, Rommens JM, Kerem B *et al*: Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989; **245**: 1066–1073.
2 Kerem B, Rommens JM, Buchanan JA *et al*: Identification of the cystic fibrosis gene: genetic analysis. *Science* 1989; **245**: 1073–1080.
3 Bobadilla JL, Macek Jr M, Fine JP, Farrell PM: Cystic fibrosis: a worldwide analysis of *CFTR* mutations — correlation with incidence data and application to screening. *Hum Mutat* 2002; **19**: 575–606.
4 Girodon-Boulandet E, Cazeneuve C, Goossens M: Screening practices for mutations in the CFTR gene *ABCC7*. *Hum Mutat* 2000; **15**: 135–149.
5 Ferec C, Audrezet MP, Mercier B *et al*: Detection of over 98% cystic fibrosis mutations in a Celtic population. *Nat Genet* 1992; **1**: 188–191.
6 Audrezet MP, Mercier B, Guillermit H *et al*: Identification of 12 novel mutations in the *CFTR* gene. *Hum Mol Genet* 1993; **2**: 51–54.
7 Le Marechal C, Audrezet MP, Quere I, Raguenes O, Langonne S, Ferec C: Complete and rapid scanning of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene by denaturing high-performance liquid chromatography (D-HPLC): major implications for genetic counselling. *Hum Genet* 2001; **108**: 290–298.
8 Audrezet MP, Chen JM, Raguenes O *et al*: Genomic rearrangements in the *CFTR* gene: extensive allelic heterogeneity and diverse mutational mechanisms. *Hum Mutat* 2004; **23**: 343–357.
9 Niel F, Martin J, Dastot-Le Moal F *et al*: Rapid detection of *CFTR* gene rearrangements impacts on genetic counselling in cystic fibrosis. *J Med Genet* 2004; **41**: e118.
10 Bombieri C, Bonizzato A, Castellani C, Assael BM, Pignatti PF: Frequency of large *CFTR* gene rearrangements in Italian CF patients. *Eur J Hum Genet* 2005; **13**: 687–689.
11 Chevalier-Porst F, Souche G, Bozon D: Identification and characterization of three large deletions and a deletion/polymorphism in the *CFTR* gene. *Hum Mutat* 2005; **25**: 504.
12 Hantash FM, Redman JB, Starn K *et al*: Novel and recurrent rearrangements in the *CFTR* gene: clinical and laboratory implications for cystic fibrosis screening. *Hum Genet*, [E-pub ahead of print].
13 Stenson PD, Ball EV, Mort M *et al*: Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 2003; **21**: 577–581.
14 Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN: Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum Mutat* 2003; **22**: 229–244.
15 Ball EV, Stenson PD, Krawczak M, Cooper DN, Chuzhanova NA: Micro-deletions and micro-insertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 2005; **26**: 205–213.
16 Chuzhanova NA, Anassis EJ, Ball E, Krawczak M, Cooper DN: Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 2003; **21**: 28–44.
17 Gusev VD, Nemytikova LA, Chuzhanova NA: On the complexity measures of genetic sequences. *Bioinformatics* 1999; **15**: 994–999.
18 Shen P, Huang HV: Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* 1986; **112**: 441–457.
19 Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR: Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *Am J Hum Genet* 1998; **62**: 1023–1033.
20 Lander ES, Linton LM, Birren B *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
21 Deininger PL, Batzer MA: *Alu* repeats and human disease. *Mol Genet Metab* 1999; **67**: 183–193.
22 Chen JM, Stenson PD, Cooper DN, Ferec C: A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* 2005; **117**: 411–427.
23 Bacolla A, Jaworski A, Larson JE *et al*: Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci USA* 2004; **101**: 14162–14167.
24 van Gent DC, Hoeijmakers JHJ, Kanaar R: Chromosomal stability and the DNA double-stranded break connection. *Nature Rev Genet* 2001; **2**: 196–206.
25 Weaver DT, DePamphilis ML: The role of palindromic and non-palindromic sequences in arresting DNA synthesis *in vitro* and *in vivo*. *J Mol Biol* 1984; **180**: 961–986.
26 Cooper DN, Krawczak M: *Hum Gene Mutation*. Oxford: BIOS Scientific, 1993.
27 Karlin S, Macken C: Some statistical problems in assessment of inhomogeneities of DNA sequence data. *J Am Statist Assoc* 1991; **86**: 27–35.
28 Magnani C, Cremonesi L, Giunta A, Magnaghi P, Taramelli R, Ferrari M: Short direct repeats at the breakpoints of a novel large deletion in the *CFTR* gene suggest a likely slipped mispairing mechanism. *Hum Genet* 1996; **98**: 102–108.
29 Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN: Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Hum Mutat* 2005; **25**: 207–221.
30 Roth DB, Chang XB, Wilson JH: Comparison of filler DNA at immune, nonimmune, and oncogenic rearrangements suggests multiple mechanisms of formation. *Mol Cell Biol* 1989; **9**: 3049–3057.
31 Baysal BE, Willett-Brozick JE, Filho PA, Lawrence EC, Myers EN, Ferrell RE: An *Alu*-mediated partial *SDHC* deletion causes familial and sporadic paraganglioma. *J Med Genet* 2004; **41**: 703–709.
32 Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN: Complex gene rearrangements caused by serial replication slippage. *Hum Mutat* 2005; **26**: 125–134.
33 Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN: Intrachromosomal serial replication slippage in *trans* gives rise to diverse genomic rearrangements involving inversions. *Hum Mutat* 2005; **26**: 362–373.
34 Aretz S, Stienen D, Uhlhaas S *et al*: High proportion of large genomic *STK11* deletions in Peutz-Jeghers syndrome. *Hum Mutat* 2005; **26**: 513–519.
35 Albrecht P, Ansperger-Rescher B, Schuler A, Zeschnigk M, Gallie B, Lohmann DR: Spectrum of gross deletions and insertions in the *RB1* gene in patients with retinoblastoma and

association with phenotypic expression. *Hum Mutat* 2005; **26**: 437–445.
36 Lalic T, Vossen RH, Coffa J *et al*: Deletion and duplication screening in the *DMD* gene using MLPA. *Eur J Hum Genet* 2005; **13**: 1231–1234.
37 Abeysinghe SS, Stenson PD, Krawczak M, Cooper DN: Gross Rearrangement Breakpoint Database (GRaBD). *Hum Mutat* 2004; **23**: 219–221.
38 Charbonnier F, Baert-Desurmont S, Liang P *et al*: The 5′region of the *MSH2* gene involved in hereditary non-polyposis colorectal cancer contains a high density of recombinogenic sequences. *Hum Mutat* 2005; **26**: 255–261.
39 Sironi M, Pozzoli U, Cagliani R *et al*: Relevance of sequence and structure elements for deletion events in the dystrophin gene major hot-spot. *Hum Genet* 2003; **112**: 272–288.
40 Toffolatti L, Cardazzo B, Nobile C *et al*: Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene. *Genomics* 2002; **80**: 523–530.
41 Ellsworth RE, Jamison DC, Touchman JW *et al*: Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc Natl Acad Sci USA* 2000; **97**: 1172–1177.
42 Zielenski J, Rozmahel R, Bozon D *et al*: Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. *Genomics* 1991; **10**: 214–228.
43 Lerer I, Laufer-Cahana A, Rivlin JR, Augarten A, Abeliovich D: A large deletion mutation in the *CFTR* gene (3120+1Kbdel8.6Kb): a founder mutation in the Palestinian Arabs. Mutation in brief no. 231. Online. *Hum Mutat* 1999; **13**: 337.
44 Dork T, Macek Jr M, Mekus F *et al*: Characterization of a novel 21-kb deletion, CFTRdele2,3(21 kb), in the *CFTR* gene: a cystic fibrosis mutation of Slavic origin common in Central and East Europe. *Hum Genet* 2000; **106**: 259–268.