Hum Genet (2000) 106:172–178
Digital Object Identifier (DOI) 10.1007/s004399900228

ORIGINAL INVESTIGATION

Cristina Bombieri · Silvia Giorgi · Soukeyna Carles · Rafael de Cid · Francesca Belpinati · Caterina Tandoi Nathalie Pallares-Ruiz · Conxi Lazaro · Bianca Maria Ciminelli · Marie-Catherine Romey · Teresa Casals Fiorenza Pompei · Giorgio Gandini · Mireille Claustres · Xavier Estivill · Pier Franco Pignatti · Guido Modiano

A new approach for identifying non-pathogenic mutations. An analysis of the cystic fibrosis transmembrane regulator gene in normal individuals

Received: 24 November 1999 / Accepted: 6 December 1999 / Published online: 11 January 2000

Abstract Given *q* as the global frequency of the alleles causing a disease, any allele with a frequency higher than *q* minus the cumulative frequency of the previously known disease-causing mutations (threshold) cannot be the cause of that disease. This principle was applied to the analysis of cystic fibrosis transmembrane conductance regulator (CFTR) mutations in order to decide whether they are the cause of cystic fibrosis. A total of 191 DNA samples from random individuals from Italy, France, and Spain were investigated by DGGE (denaturing gradient gel electrophoresis) analysis of all the coding and proximal non-coding regions of the gene. The mutations detected by DGGE were identified by sequencing. The sample size was sufficient to select essentially all mutations with a frequency of at least 0.01. A total of 46 mutations was detected, 20 of which were missense mutations. Four new mutations were identified: 1341+28 C/T, 2082 C/T, L1096R, and I1131V. Thirteen mutations (125 G/C, 875+40 A/G, TTGAn, IVS8–6 5T, IVS8–6 9T, 1525–61 A/G, M470V, 2694 T/G, 3061–65 C/A, 4002 A/G, 4521 G/A, IVS8 TG10, IVS8 TG12) were classified as non-CF-causing alleles on the basis of their frequency. The remaining mutations have a cumulative frequency far exceeding *q*; therefore, most of them cannot be CF-causing mutations. This is the first random survey capable of detecting all the polymorphisms of the coding sequence of a gene.

C. Bombieri[1] (✉) · F. Belpinati · P. F. Pignatti
Department of Mother and Child, Biology and Genetics, University of Verona, Italy

S. Giorgi · C. Tandoi · B. M. Ciminelli · F. Pompei · G. Modiano
Department of Biology "E.Calef", University "Tor Vergata", Rome, Italy

S. Carles · N. Pallares-Ruiz · M.-C. Romey · M. Claustres
Institute of Biology, University of Montpellier, France

R. de Cid · C. Lazaro · T. Casals · X. Estivill
Medical and Molecular Genetics Center, Hospital Duran i Reynals, Barcelona, Spain

G. Gandini
Blood Center, Verona Hospital, Italy

Present address:
Sezione di Biologia e Genetica, MIBG, Università di Verona, Strada Le Grazie 8, I-37134 Verona, Italy, e-mail: cristy@borgoroma.univr.it, Tel.: +39 045 8098673, Fax: +39 045 8098180

Introduction

In an age when molecular genetic analysis has identified several hundreds of different mutations for many disease-causing genes, a common problem is to decide which mutations are disease-causing and which are not. The role of a not-obviously-pathogenic (e.g., frameshift or stop codon) gene mutation in causing a disease is usually assessed through its frequency in clinically characterized patients and confirmed through evolutionary conservation comparisons or functional studies in vitro or in experimental animal models. However, given the enormous number of known gene mutations for most genetic diseases, this is not feasible for all mutations. Therefore, alternative approaches to establish the possible involvement of a gene in disease etiopathogenesis would be useful.

We propose here a conceptually straightforward method to identify non-pathogenic mutations. We have applied this method to the analysis of the cystic fibrosis transmembrane conductance regulator (CFTR) gene in which more than 800 different mutations have previously been described (Cystic Fibrosis Genetic Analysis Consortium, CFGAC website). One must remember, however, that CFTR gene mutations could be related to different diseases. Indeed, they are not only related to the classical form of cystic fibrosis (CF) but also to atypical forms of the disease (Estivill 1996), congenital bilateral absence of the vas deferens (Chillon et al. 1995), disseminated bronchiectasis (Pignatti et al. 1995; Girodon et al. 1997; Bombieri et al. 1998), chronic pancreatitis (Cohn et al. 1998; Sharer et al. 1998), and others. Since the only well-characterized spectrum of CFTR gene mutations in many different populations is related to CF, this will be the only disease considered here. This project concerns the classi-

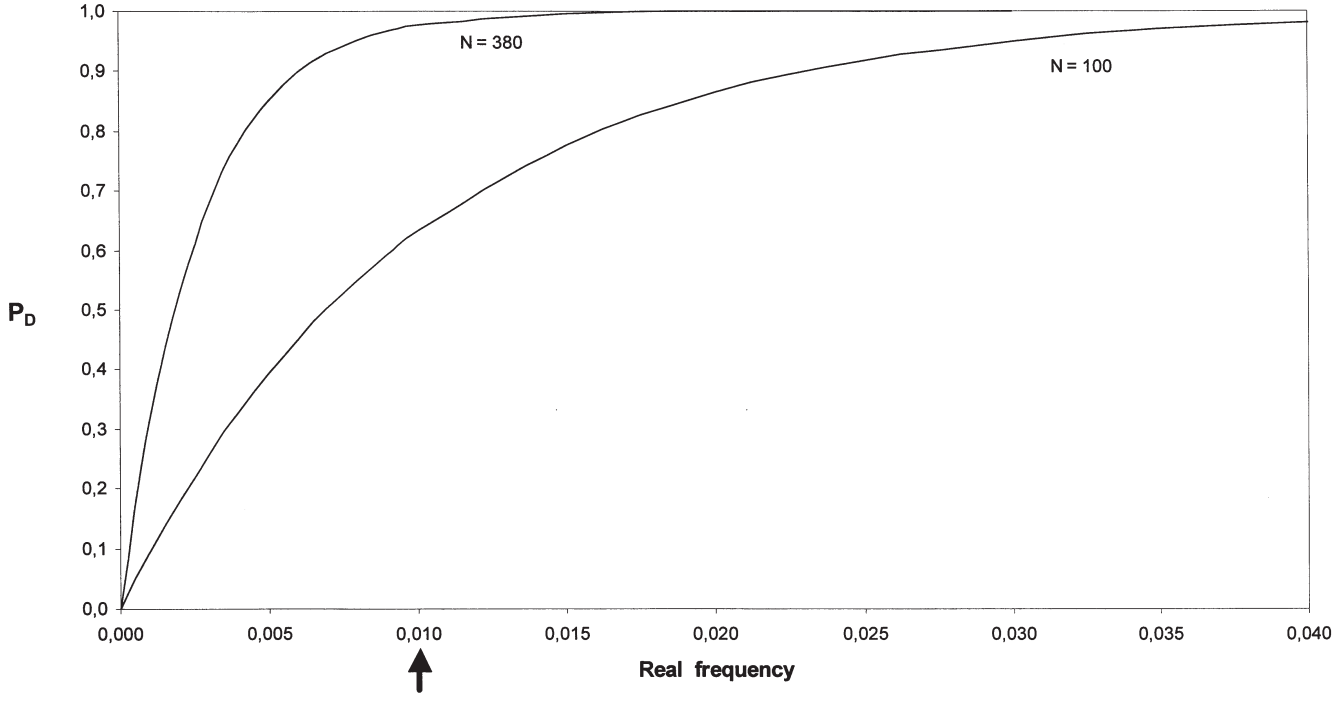

Fig. 1 Probability $P_{D\ (detection)}$ that a mutation happens to be included in the sample, as a function of its frequency. *N* is the number of genes examined. One hundred and 380 indicate the approximate size of each of the four subsamples and that of the total sample of the present work, respectively. The *arrow* indicates the detectability limit, i.e., the value above which all mutations are expected to be detected with the *N*=380 sample size used in this study

fication of non-CF-causing mutations with a thorough analysis of all the coding and flanking non-coding sequences of the CFTR gene in random individuals from various European countries.

Materials and methods

Statistical analysis

The rationale of the present approach for identifying pathogenic genes is based on the consideration that any mutation more frequent than the difference between the global frequency of the disease-causing mutations and the cumulative frequency of the previously known disease-causing mutations cannot be the cause of the disease ("common", hence non-pathogenic, mutations). This difference is the "threshold" above which a mutation is defined as a common (C) mutation. The statistical efficiency of the approach corresponds to the proportion of C mutations that are expected to be discovered and recognized as common. For each C mutation, this efficiency depends on two distinct probabilities: (1) the probability of being included in the sample under study, and (2) the probability that it is recognized as being common.

The probability that a mutation is included in the sample was calculated as follows: $P=(1-P_0)$, where P_0 is the probability of not being included. P_0 was calculated according to Poisson's distribution: $P_0=(\mu^0/0!)\ e^{-\mu}$, where μ is the expected absolute frequency of the mutation in the sample (Fig. 1). As far as point (2) above is concerned, a mutation can be classified with certainty as being common if its frequency has been established to be higher than the threshold. This condition is fulfilled if f_{obs}–2.5SE is greater than the threshold. In the case of CF in southern Europe, the threshold is 0.004, since the global frequency of CF mutations in Europe is about 0.02, that of ΔF508 is about 0.01 in Italy (Bonizzato et al. 1994; Rendine et al. 1997), in Southern France (Claustres et al. 1993), and in Spain (Estivill et al. 1997), and the combined frequency of the other most common CF mutations is about 0.006. Therefore the probability that a C mutation is classified as being common (hence non-CF-causing) was calculated as follows: f_{obs} must be at least equal to 0.004+2.5SE; this value $f_{minimal}$ is 0.025, 0.015, 0.012, and 0.009 when *N* is 380, 1000, 1500, and 3000, respectively. This is the same as saying, for example, that, in a sample of 3000 genes, only the mutations with an observed frequency of at least 0.009 can be classified with certainty as C mutations, i.e., 0.009–2.5 (0.009×0.991)/3000. For each mutation with a true frequency f_{true} higher than 0.004, its probability of exhibiting a frequency higher than $f_{minimal}$ was obtained as follows: the algebric difference $D = f_{minimal} - f_{true}$ is standardized by dividing it by the SE of f_{true}; the one-tail probability of standardized *D* is derived from the normal distribution table. When *D* is negative, the value to be plotted is 1 – the value found in the one-tail probability, instead of the value itself.

Samples

Four population samples, with a total of 191 individuals, were studied. One sample consisted of 50 individuals, each of whom had all four grandparents born in Veneto (North-East Italy). Another sample consisted of 50 individuals, each with four grandparents born in Latium or Umbria or Abruzzi (Central Italy). The third sample consisted of 50 individuals whose parents and grandparents lived in Southern France. The fourth sample consisted of 41 individuals who were spouses of CF carriers and born in Spain. In order to avoid strictly local effects, each subsample was made up of individuals living in a large town. Blood specimens were collected after informed consent.

Mutation analysis

The whole coding (with the exception of the first 10 nucleotides of the 13th exon) and partial proximal non-coding sequences were analysed (see Table 1). Genomic DNA was prepared from periph-

Table 1 A list of the analyzed non-coding sequences

Region	Length of the sequence analysed		
		5’	3’
UTRs			
5’	147[a]		
3’	36		
Introns			
1		57	21
2		26	79
3		55	52[a]
4		22[a]	29
5		40	25
6a		65[a]	70[a]
6b		70	13
7		15	51
8		68[a]	51[a]
9		61	30
10		40	29
11		20	31
12		109	–
13		45	33
14a		47	31
14b		19	28[a]
15		28	81[a]
16		97	29
17a		28	29
17b		47	58
18		38	32
19		41	18
20		53[a]	64
21		31	47
22		30	22
23		34	45
Total	183	1186	998

[a]At least one mutation was found in these sequences

eral blood samples by standard methods. Polymerase chain reaction (PCR) DNA amplification and denaturating gradient gel electrophoresis (DGGE) were performed as described (Fanen et al. 1992; Costes et al. 1993) or after local modifications (M. Claustres et al., unpublished). Every mutant discovered by this method was sequenced either manually or with the ABI Prism 377 or 310 sequence analyser. The IVS8–6 polymorphism (a stretch of 5, 7, or 9 T) was studied as described (Chillon et al. 1995). The IVS8 $(TG)_m$ polymorphism (a stretch of 10, 11, or 12 TG) was also studied as described (Strasberg et al. 1997). For the Spanish sample only, 12 CFTR gene exons (1, 2, 4, 6b, 7, 10, 13b, 16, 17a, 19, 22, 24) were analyzed by single-strand conformation analysis (SSCA) as previously described (Chillon et al. 1994). The 4521 G/A mutation was studied in all subjects by a restriction modification assay (Gasparini et al. 1992), because DGGE does not distinguish between the two types of homozygotes. According to Macek et al. (1997), the sensitivity of the DGGE method with respect to detecting mutations in the CFTR gene is about 100%. In our hands, this method exhibited a sensitivity of 98.6%, as we have easily detected 73 out of 74 mutations not originally discovered by DGGE.

Results

Statistical approach

The purpose of the present approach is to identify as many C mutations as possible. The probability of each of these mutations being classified as C is the result of: (1) the probability of being detected, (2) the probability of conclusively demonstrating that its frequency is higher than 0.004. From a statistical point of view, both points (1) and (2) are a function of the frequency q of the mutation and of the sample size of random genes examined.

Figure 1 shows the probability of detecting a mutation as a function of frequency and sample size. With a sample of 380 genes (as that of the present study), roughly 15% of the mutations with a frequency around 0.005 are expected not to be included in the sample. Figure 2 shows the probability of a mutation being classified as a C mutation as a function of the frequency for various sample sizes. With a sample of 3000 genes, 50% of the mutations with a frequency around 0.009 are expected not to be demonstrated as C mutations. C mutations can be subdivided into three classes differing from one another with respect to (1) the rate of statistical detectability, i.e., the probability that the mutation happens to be included in the sample, and (2) statistical demonstrability, i.e., the probability of showing that the mutation is a C mutation by finding that its q_{obs}–2.5SE is greater than 0.004. The classification of C mutations is shown in Table 2. The most relevant feature is the dramatic difference between the sample size required to ensure a satisfactory rate of statistical detectability and that required to attain a satisfactory rate of statistical demonstrability, which is one order of magnitude larger. The obvious implication is that a full detectability of all the C mutations can be reasonably achieved, but not a full demonstrability. In particular, mutations with a frequency $0.004<q<0.01$, i.e., the “impossible” to demonstrate (ID-C) mutations, may not be identified with certainty as C mutations.

On the basis of the above considerations, this is a two-step project. The first phase aims are (1) to detect and identify every “easy” to demonstrate (ED-C) mutation as C; (2) to detect every “difficult” to demonstrate (DD-C) mutation present in the populations under study. The second phase aim is to identify which mutations, if any, among those detected in the first phase and still not classified with respect to pathogenicity, can be recognized with certainty as being C mutations. The first phase demands the full analysis of the CFTR gene (all of its 27 exons and flanking intronic sequences) in a relatively small sample (a few hundred); the second phase requires a specific search limited to the mutations detected in the first phase and of still uncertain clinical meaning, but on a larger sample (a few thousand).

Mutations found in the sample

The complete list of the mutations found in this work and their frequency is given in Table 3; 42 sites with two alleles and two sites with three alleles have been detected corresponding to a total 44 sites and 90 alleles. These findings include four novel mutations described here for the first time: 1341+28 C/T, 2082 C/T, L1096R, and I1131V.

For each mutation, the frequencies observed in the four subsamples were not significantly different from each other after correcting for multiple comparisons. They

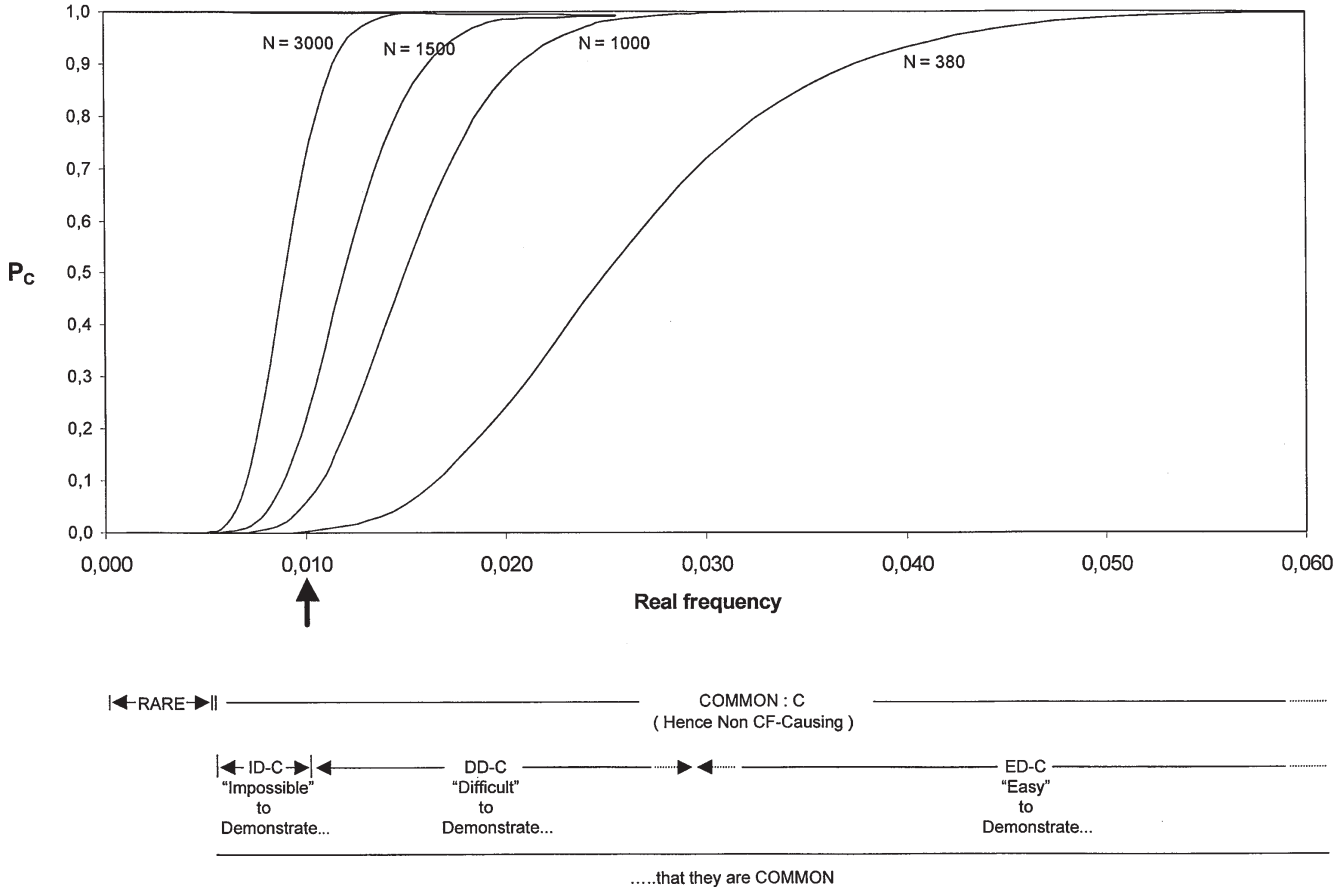

Fig. 2 Probability $P_{C\ (classification)}$ of classifying C mutations as a function of their frequency, for different sample sizes. The *arrow* indicates the frequency above which it becomes possible to demonstrate that a mutation is a C mutation. *N* is the number of genes, for an explanation of *ID-C*, *DD-C*, *ED-C* see Table 2

Table 2 A subdivision of C mutations based on frequency and sample size

Frequency of the C mutations	Detectability (%)		Demonstrability (%)[a]	
	380 genes	3000 genes	380 genes	3000 genes
$0.004<q<0.01$	80–98	100	0	0–70[b]
$0.01<q<0.03$	98–100	100	0–70	70–100[c]
$q>0.03$	100	100	>70	100[d]

[a]That an observed mutation is a C mutation
[b]ID-C: "impossible" to demonstrate, in the sense that a sample size of 3000 genes may be insufficient to demonstrate with certainty that one such mutation is indeed a C mutation (by showing that its observed frequency q_{obs} fulfils the condition q_{obs}–2.5SE>0.004)
[c]DD-C: "difficult" to demonstrate, in the sense that the sample size required to show that q_{obs}–2.5SE>0.004 is in the order of 3000 genes
[d]ED-C: "easy" to demonstrate, because a relatively small sample size (380 genes) is expected to be sufficient to show with certainty that such mutations are C mutations

were therefore pooled, and a combined frequency was obtained. Since both the statistical efficiencies (for mutations with a $q>0.01$) and experimental efficiencies were substantially 100%, one can state that all mutations present with a frequency of at least 0.01 in the area under study have been detected and are included in the list shown in Table 3. The 46 different mutations listed in Table 3 affect sequences of the following types: 31 in the coding sequence (20 missense, 10 samesense, and one trinucleotide deletion: ΔF508), 12 in the intronic sequences, and one in the 5'UTR. Out of the 20 missense mutations, three (G85E, ΔF508, and N1303K) are certainly CF-causing, and several (R31C, K68E, R75Q, I148T, V562L, G576A-R668C, L997F, F1052V, S1235R) have been described in congenital bilateral absence of the vas deferens, in disseminated bronchiectasis, in pancreatitis, or in atypical CF cases mutations as reported in the CFGAC website ().

Many (13 out of 20) of the missense mutations change highly conserved (5/5 species analyzed) amino acid residues (R75Q, G85E, I148T, I506V, R668C, G622D, L997F, I1027T, F1052V, L1096R, I1131V, R1162L, N1303K); others affect amino acid residues conserved in 4/5 species (K68 E, R170H, M470V, V562L, S1235R), or in 3/5 species (R31C and G576A; Tucker et al. 1992). This distribution is compatible with that observed by comparing these species for the whole coding sequence.

The sum of the absolute frequencies of the 46 mutations was 513 on a total sample of 382 genes (165/100, 113/100, 137/100, and 98/82 in North-East Italy, Central

Table 3 Frequency of CFTR gene mutations found in random individuals (mutations in *bold*: novel mutations found in this work). The TG repeat (i 8) was analyzed in three samples only: North-East Italy, Central Italy, and Southern France. On a total of 100 genes for each sample, TG_{10} was detected 27, 24, and 36 times, respectively, and TG_{12} 11, 8, and 7 times, respectively. The most common allele at this site is TG_{11}. Moreover, 1525–61 A/G (i 9) and 3601–65 C/A (i 18) were detected by SSCA performed in the Spanish sample only (14/82 and 12/80, respectively); these mutations were not identifiable by DGGE as used in the present work

Mutation	Position	North-East Italy 100 genes	Central Italy 100 genes	Southern France 100 genes	Spain 82 genes	Total	
						382 genes	%
125 G/C	5'UTR	1	2	7	3	13	3.4
R31C	2	1	1	1	0	3	0.8
K68E	3	1	0	0	0	1	0.3
R75Q	3	1	1	2	0	4	1.0
G85E[e]	3	0	1	0	0	1	0.3
406–6 T/C	i 3	0	0	1	0	1	0.3
I148T	4	1	0	0	0	1	0.3
621+3 A/G	i 4	0	1	0	0	1	0.3
R170H	5	1	0	0	0	1	0.3
875+40 A/G	i 6a	11	5	5	2	23	6.0
$(TTGA)_6$[f]	i 6a	17	11	7	13	48	12.6
1341+28 C/T	i 8	1	0	0	0	1	0.3
IVS8–6[g] T_5	i 8	8	2	4	3/78	17[a]	4.5
IVS8–6[g] T_9	i 8	10	7	10	11/78	38[a]	10.0
M470V[h]	10	42	30	39	27	138	36.1
I506V	10	1	0	0	0	1	0.3
ΔF508[e]	10	1	0	2	0	3	0.8
1716 G/A	10	2	1	0	5	8	2.1
V562L	12	0	0	1	0	1	0.3
G576A	12	1	0/80	1	0	2[b]	0.6
G622D	13	0	0/80	1	0	1[b]	0.3
R668C	13	1	0/80	1	0	2[b]	0.6
2082 C/T	13	1	0/80	0	0	1[b]	0.3
2377 C/T	13	0	0/80	0	1	1[b]	0.3
2694 T/G	i 14a	33	23	33	14/80	103[c]	27.1
2752–15 C/G	i 14b	0	3	0	0	3	0.8
3041–71 G/C	i 15	0	1	2	0	3	0.8
L997F	17a	0	2	0	0	2	0.5
I1027T	17a	1	0	0	0	1	0.3
F1052V	17b	1	0	0	0	1	0.3
L1096R	17b	0	0	1	0	1	0.3
3417 A/T	17b	1	0	1	0	2	0.5
I1131V	18	0	1	0	0	1	0.3
R1162L	19	0	1	0	0	1	0.3
3690 A/G	19	0	0	0	1/80	1[c]	0.3
S1235R	19	1	0	0	0	1	0.3
4002 A/G	20	2	3	3	3/80	11[c]	2.9
4005+28insA	i 20	0	1	0	0	1	0.3
4029 A/G	21	1	0	0	0	1	0.3
N1303K[e]	21	1	0	0	0	1	0.3
4404 C/T	24	1	0	1	0	2	0.5
4521 G/A	24	21	16	14/80	15/76	66[d]	18.5
Total		165	113	137	98	513	

The totals are: [a]378; [b]362; [c]380; [d]356 genes
[e]Certainly a CF-causing mutations
[f]The most common allele at this site is $(TTGA)_7$
[g]The most common allele at this site is T_7
[h]The frequency shown is that of the M allele

Italy, Southern France, and Spain samples, respectively, with an overall highly significant excess in North-East Italy as compared with the other three samples). Therefore, in many cases, two or more mutations are localized in the same gene. In addition to very common mutations, which are often associated with other mutations in the same gene, two sporadic mutations were found together in the same individual (G576A and R668C once; 3041–71 G/C and 4002 A/G twice) in three cases. On the basis of personal observations and of published data (Bombieri et al. 1998), we believe that these mutations are located in the same genes. In three further cases where more than one sporadic mutation was observed in the same individual (1341+28 C/T and F1052V and S1235R; ΔF508 and I1027T; 1716G/A and N1303K), data from the literature were not available, and segregation analysis was not possible; thus, their phase could not be established.

Discussion

Medical genetics

Thirteen of the 46 variant alleles detected in the present study show a frequency high enough to allow the conclusion that their frequency in the population is certainly higher than 0.004; therefore, they have been classified as C mutations (Table 4). These are ED-C mutations according to the nomenclature adopted in Table 2, because their observed frequency (≥0.03) is much higher than the threshold of 0.004. The remaining 30 (after having excluded the certainly CF-causing mutations G85E, ΔF508 and N1303K) have been found a few times only: once (20 mutations), twice (five mutations), three times (three mutations), four times (one mutation), and eight times (one mutation). Two reasons for these mutations having been

Table 4 A list of the 13 ED-C mutations detected in this survey

Mutation	q±SE	q–2.5SE
125 G/C	0.0340±0.0093	0.011
875+40 A/G	0.0602±0.0122	0.030
876–5 $(GATT)_6$	0.1257±0.0170	0.083
IVS8 T_5	0.0450±0.0107	0.018
IVS8 T_9	0.1005±0.0155	0.062
IVS8 $(TG)_{10}$	0.2900±0.0262	0.223
IVS8 $(TG)_{12}$	0.0867±0.0162	0.046
1525–61 A/G[a]	0.1750±0.0420	0.070
M470V	0.3613±0.0246	0.300
2694 T/G	0.2711±0.0228	0.214
3601–65 C/A[a]	0.1500±0.0399	0.050
4002 A/G	0.0289±0.0086	0.007
4521 G/A	0.1854±0.0206	0.134

[a]Searched by SSCA in the sample of 40 Spanish individuals only: frequency and standard error are those of that sample

Table 5 Distribution in the four subsamples of mutations found a few times but not classified

Total number of times the mutation has been found	Subsample			
	NE Italy	Central Italy	Southern France	Spain
Twice				
G576A	1	–	1	–
R668C	1	–	1	–
L997F	–	2	–	–
3417 A/T	1	–	1	–
4404 C/T	1	–	1	–
Three times				
R31C	1	1	1	–
2752–15 C/G	–	3	–	–
3041–71 G/C	–	1	2	–
Four times				
R75Q	1	1	2	–
Eight times				
1716 G/A[a]	2	1	–	5

[a]Given its frequency and distribution, this mutant will probably turn out to be a C mutant

encountered in the present survey are possible. (1) The mutation may be one of a very large set of extremely rare mutations, namely one of a set of mutations each having a very low probability of being included by pure chance in a random sample of 400 genes (for example, such a probability would be 4×10^{-3} for a mutation with a frequency of 10^{-5}). Obviously, such rare mutations are not to be expected to be found again in future surveys. If n is the number of such mutations, one expects to find, in an additional survey of 400 random genes, once again about n such extremely rare mutations, which should however be different from the present ones. (2) The mutation may be relatively common in the population in which it was found, but rarer in the other populations of the present survey. In a much enlarged sample, a mutation of this kind should be found again in several specimens of the population in which it was originally found and not in the remaining populations. Indeed, in some cases, one may reach the conclusion that a mutation is a C mutation even though its overall observed frequency does not fulfil the condition q_{obs}–2.5SE is greater than 0.004: in order to be classified as a C mutation, it is sufficient that, even in one subsample, its q_{obs}–2.5SE value is higher than 0.004. This is the case for two mutations detectable only with a method used for the Spanish sample (Table 4). The distribution in the four samples of mutations observed a few times is given in Table 5. Even though, considering a multiple comparison correction, no significant sample heterogeneity was found, it appears likely that one or more of the observed inter-ethnic differences is the expression of an actual relatively high frequency of a mutation in the area in which it was found.

Population genetics

The relatively high frequency of the 13 alleles of Table 4 makes them possible candidates as anthropogenetic markers, namely markers of potential usefulness for characterizing populations. Surprisingly enough, this aspect of the CFTR gene seems to have been overlooked. This is probably because of the overwhelming clinical interest in the CFTR gene.

The common variants of the CFTR gene appear to be evenly distributed in Southern Europe since, considering a multiple comparison correction, no obvious differences in their frequencies have been found between the four random samples of the present survey. The next step will consist of estimating the frequency of these alleles for other European and non-European populations in order to identify which, if any, will be a potentially useful anthropogenetic marker.

Structural gene evolution

The present survey is, to the best of our knowledge, the first deliberate and systematic random search for polymorphic mutations in a structural gene with a method close to 100% efficiency. This approach, therefore, fulfils the conditions required to obtain a reliable estimate of the common variability of the structural gene under study, with a characterization of the functional distribution, e.g., introns vs exons, missense-single-nucleotide-substitution (SNS) vs samesense-SNS mutations, etc., and the spatial distribution of the mutations. The limit of this survey is that it concerns a single, although large, gene; the results cannot therefore be extrapolated to all structural genes.

Four types of variable sites, such as SNS (31 exonic and 8 intronic), deletions (one), insertions (one), and simple tandem repeats (three), were found in the present study.

The density of polymorphic sites, which is essentially the proportion of sites susceptible to evolving among the total number of sites (bp) of that gene, can be estimated by counting the number n of polymorphic sites existing in

the stretch under study of *N* bp and dividing *n* by *N*. Usually, the number of sites identified as polymorphic sites is merely a minimum estimate of the total number *n* of polymorphic sites of the *N* stretch, because the number of polymorphic sites of the stretch that escaped detection remains unknown. The present investigation is an exception to this rule, since the sample size and the technique adopted should have enabled us to detect every existing mutation with a polymorphic frequency at least once (with the practically negligible exception of those with a frequency *q* ranging between 0.005 and 0.01). Therefore, the highest possible number of polymorphic sites in the DNA under study simply corresponds to the number of mutations observed (assuming that all the observed mutations are polymorphic). Thus, the present data identify both the minimum (n_{min}=13, i.e., the number of sites already classified as polymorphic; Table 4) and the maximum number of polymorphisms (n_{max}: if all the 46 variations of Table 3 are polymorphic). It is therefore possible to define the range of uncertainty of the density of polymorphic sites *n/N* precisely. The true *n* will become known by searching for the 33 remaining mutations not yet classified as C mutations in a random sample of 2500 more alleles (second phase of the research project).

The overall density is a coarse parameter to describe *n/N*. Therefore, this density of polymorphic sites will be considered separately for each SNS type. The maximum number of missense, samesense, and intronic SNSs possible in the studied sequence is 2862, 9815, and 6552, respectively. The number of SNSs found was 6, 6, and 3, respectively. The ratio $n_{obs}/n_{possible}$ of samesense mutations was three times larger than that of missense mutations.

These findings, although preliminary, suggest that the rate of common samesense variation is not much higher than that of the missense variation as implied by the theory that polymorphisms are, as a rule, neutral polymorphisms.

Acknowledgements This work was funded by: the Italian Ministry of Health, CF Project, law 548/93; the Italian Ministry of University and Scientific and Technological Research; Consorzio Studi Universitari Verona; the Italian CNR Strategic Project Biotechnology; the Italian CNR Strategic Project Cultural Goods; Fondo de Investigaciones Sanitarias (FIS98/0977); and the European Union CEC/BIOMED2 (BMH4-CT97–2486). We thank the ECCACF for providing mutation controls and the PCR mix for multiplex DGGE of the CFTR gene. C. B. was supported by the CNR Biotechnology Project. F. B. was supported by CF Project, law 548/93. The experiments of this study comply with the current laws of Italy, France, and Spain.

References

Bombieri C, Benetazzo MG, Saccomani A, Belpinati F, Gilè LS, Luisetti M, Pignatti PF (1998) Complete mutational screening of the CFTR gene in 120 patients with pulmonary disease. Hum Genet 103:718–722

Bonizzato A, Bisceglia L, Marigo C, Nicolis E, Bombieri C, Castellani C, Borgo G, et al (1994) Analysis of the complete coding region of the CFTR gene in a cohort of CF patients from North-Easten Italy: identification of 90% of the mutations. Hum Genet 95:397–402

CFGAC website: http://www.genet.sickkids.on.ca/cftr

Chillon M, Casals T Gimenez J, Nunes V, Estivill X (1994) Analysis of the CFTR gene in the Spanish population: SSCP-screening for 60 known mutations and identification of four new mutations (Q30X, A120 T, 1812–1G/A, and 3667del4). Hum Mutat 3:2323–2330

Chillon M, Casals T, Mercier B, Bassas L, Lissens W, Silber S, Romey M-C, et al (1995) Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. N Engl J Med 332:1475–1480

Claustres M, Laussel M, Desgeorges M, Glansily M, Culard J-F, Razakatsara G, Demaille J (1993) Analysis of the 27 exons and flanking regions of the cystic fibrosis gene: 40 different mutations account for 91.2% of the mutant alleles in Southern France. Hum Mol Genet 2:1209–1213

Cohn JA, Friedman KJ, Noone PG, Knowles MR, Silverman LM, Jowell PS (1998) Relation between mutations of the cystic fibrosis gene and idiopathic pancreatitis. N Engl J Med 339: 653–658

Costes B, Fanen P, Goossens M, Ghanem N (1993) A rapid, efficient, and sensitive assay for simultaneous detection of multiple cystic fibrosis mutations. Hum Mutat 2:185–191

Estivill X (1996) Complexity in a monogenic disease. Nat Genet 12:348–350

Estivill X, Bancells C, Ramos C, the Biomed CF Mutation Analysis Consortium (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. Hum Mutat 10:135–154

Fanen P, Ghanem N, Vidaud M, Besmond C, Martin J, Costes B, Plassa F, Goossens M (1992) Molecular characterization of cystic fibrosis: 16 novel mutations identified by analysis of the whole cystic fibrosis conductance transmembrane regulator (CFTR) coding regions and splice site junctions. Genomics 12:770–776

Gasparini P, Bonizzato A, Dognini M, Pignatti PF (1992) Restriction site generating-polymerase chain reaction (RG-PCR) for the probeless detection of hidden genetic variation: application to the study of some common cystic fibrosis mutations. Mol Cell Probes 6:1–7

Girodon E, Cazeneuve C, Lebargy F, Chinet T, Costes B, Ghanem N, Martin J, Lemay S, Scheid P, Housset B, Bignon J, Goossens M (1997) CFTR gene mutations in adults with disseminated bronchiectasis. Eur J Hum Genet, 5:149–155

Macek M, Mercier B, Mackovà A, Weiner Miller P, Hamosh A, Férec C, Cutting GR (1997) Sensitivity of the denaturing gradient gel electrophoresis technique in detection of known mutations and novel Asian mutations in the CFTR gene. Hum Mutat 9:136–147

Pignatti PF, Bombieri C, Marigo C, Benetazzo MG, Luisetti M (1995) Increased incidence of cystic fibrosis gene mutations in adults with disseminated bronchiectasis. Hum Mol Genet 4:635–639

Rendine S, Calafell F, Capello N, Gagliardini R, Caramia G, Rigillo N, Silvetti M, et al (1997) Genetic history of cystic fibrosis mutations in Italy. I. Regional distribution. Ann Hum Genet 61:411–424

Sharer N, Schwarz M, Malone G, Howarth A, Painter J, Super M, Braganza J (1998) Mutations of the cystic fibrosis gene in patients with chronic pancreatitis. N Engl J Med 339:645–652

Strasberg PM, Friedman K.J, McGlynn-Steele L, Zielenski J, Ray PN (1997) Rapid characterization of both the variable length 5, 7, or 9 polythimidine (T) tract in intron 8 and the adjacent CA repeat unit of the CFTR gene: use in DNA diagnostics. Am J Hum Genet Suppl 61: no. 4

Tucker SJ, Tannahill D, Higgins CF (1992) Identification and developmental expression of the *Xenopus laevis* cystic fibrosis transmembrane conductance regulator gene. Hum Mol Genet 1:77–82