European Journal of Human Genetics (2006) 14, 85–93

www.nature.com/ejhg

ARTICLE

Haplotype block structure study of the CFTR gene. Most variants are associated with the M470 allele in several European populations

Fiorenza Pompei[1], Bianca Maria Ciminelli[1], Cristina Bombieri[2], Cinzia Ciccacci[1], Monika Koudova[3], Silvia Giorgi[1], Francesca Belpinati[2], Angela Begnini[2], Milos Cerny[4], Marie Des Georges[5], Mireille Claustres[5], Claude Ferec[6], Milan Macek Jr[3], Guido Modiano[1] and Pier Franco Pignatti*[,2]

[1]*Department of Biology, University of Roma-Tor Vergata, Italy;* [2] *Section of Biology and Genetics, Department of Mother and Child and of Biology-Genetics, University of Verona, Strada le Grazie 8, 37134 Verona, Italy;* [3]*Department of Molecular Genetics, CF-Centre, Charles University, Prague, Czech Republic;* [4]*Department of Neonatology, Clinic of Obstetrics and Gynecology, Charles University Prague, Czech Republic;* [5]*Institute of Biology, University of Montpellier, France;* [6]*Centre de Biogénétique, CDTS, Brest, France*

An average of about 1700 CFTR (cystic fibrosis transmembrane conductance regulator) alleles from normal individuals from different European populations were extensively screened for DNA sequence variation. A total of 80 variants were observed: 61 coding SNSs (results already published), 13 noncoding SNSs, three STRs, two short deletions, and one nucleotide insertion. Eight DNA variants were classified as non-CF causing due to their high frequency of occurrence. Through this survey the CFTR has become the most exhaustively studied gene for its coding sequence variability and, though to a lesser extent, for its noncoding sequence variability as well. Interestingly, most variation was associated with the M470 allele, while the V470 allele showed an 'extended haplotype homozygosity' (EHH). These findings make us suggest a role for selection acting either on the M470V itself or through an hitchhiking mechanism involving a second site. The possible ancient origin of the V allele in an 'out of Africa' time frame is discussed.

European Journal of Human Genetics (2006) **14,** 85–93. doi:10.1038/sj.ejhg.5201498; published online 26 October 2005

Keywords: CFTR; random variability; allele-restricted haplotype variability

Introduction

The CFTR (cystic fibrosis transmembrane conductance regulator) gene has been extensively characterized for its pathologic variability. However, the study of its overall random variability, wherein disease causing mutations should be framed, has started only recently.[1]

***Correspondence: Professor PF Pignatti, Section of Biology and Genetics, Department of Mother and Child and of Biology-Genetics, University of Verona, Strada le Grazie 8, 37134 Verona, Italy. Tel: +39 045 584602; Fax: +39 045 8027180; E-mail: pierfranco.pignatti@univr.it**
Received 3 May 2005; revised 19 August 2005; accepted 23 August 2005; published online 26 October 2005

For many of the several hundreds CFTR variants reported, it is not known whether they are or not CF-causing and this may produce difficulties for genetic counselling. Besides the obvious criteria to identify with certainty the CF-causing mutations (eg frameshift and termination mutations), a purely statistical approach to identify not fully penetrant CF-causing mutations has been proposed by Bombieri *et al*.[2] It is based on the consideration that every CFTR variant with a frequency certainly higher than the cumulative frequency of the not unambiguously identified CF-causing alleles cannot be a fully penetrant CF-causing allele. It was applied to a random

sample of 191 Europeans (=382 genes), a population where the cumulative frequency of the not unambiguously identified CF-causing alleles is 0.004 (the difference between 0.02, the total frequency of the CF-causing alleles, and 0.016, the total frequency of the well identified CF-causing alleles: WHO Report[3]). In that study, 10 certainly non-CF-causing alleles were classified.[2]

Present paper reports eight further certainly not fully penetrant CF-causing alleles identified through this purely statistical approach. However, the most interesting finding concerned the very different patterns of variability found on the CFTR genes carrying the M470 or the V470 allele.

Materials and methods

The sample

A large part of the present sample was the same as previously described.[1] It consists of 1337 healthy, unrelated individuals (selected on the basis of the birth place of the four grandparents) from six geographical areas: Northern Italy, Verona ($n=300$); Central Italy, Rome ($n=300$); Southern France, Montpellier ($n=300$); Northern France, Brest ($n=278$); Czech Republic, Prague ($n=118$); and Spain, Barcelona ($n=41$). All individuals gave their informed consent. Since not all individuals have been studied for all the 27 exons of the CFTR gene, an average sample size has been computed. It amounts to about 1700 haploid genomes. For a detailed list of the sample size studied for each exon see Modiano *et al*.[1]

Mutation analysis

Genomic DNA was extracted from blood samples, amplified *in vitro* by PCR and analysed by DGGE[2] or DHPLC.[4] Every mutant discovered by these methods was sequenced with the ABI PRISM 377 or 310 Sequence Analyser. Some variants have been studied, on a fraction of the total sample, with a restriction enzyme specific method: the following cSNSs numbered as in Table A1, nos. 1, 12, 20, 24–26, 28, 29, 37, 45, 48, 56, 59, and 60; and the following intronic variants 3041–71g/c, 1001 + 11c/t, and 2752–15c/g (methods available on request cristina.bombieri@medgen.univr.it).

Maximum likelihood (ML)

Estimates of haplotype frequencies, of linkage disequilibria and of their statistical significance were calculated by ARLEQUIN, ver. 2.000.[5]

Degree of heterozygosity (H)

The degree of heterozygosity (2pq for diallelic sites) has been calculated for each variable site both within the M and the V CFTR genes utilizing the allele frequencies for each variable site within the M (or the V) CFTR genes (see Table 1).

Web resources

Information about CFTR gene sequence and mutations are available at the Cystic Fibrosis Genetic Analysis Consortium Web Site: http://genet.sickkids.on.ca/cftr

Results

A total of 4443 coding and 2367 noncoding bp (2184 bp intronic plus 183 bp of the UTR regions) had been studied by DGGE (denaturing gradient gel electrophoresis) or DHPLC (denaturing high performance liquid chromatography). Table A1 is an update of that already published in Modiano *et al*[1] and reports the absolute and relative frequencies of the 61 cSNSs (single-nucleotide substitutions in a coding sequence) found in a Czech sample, larger than that already published, together with the updated European frequencies.

A detailed analysis of the cSNS variability has been presented elsewhere.[1] Among the 61 cSNSs (45 nonsynonymous, and 16 synonymous) observed in the entire length of the gene, three (ref. nos. 16, 32, 60) were frankly polymorphic ($q>0.05$) and eight only slightly polymorphic ($0.005<q<0.05$); all the other cSNSs showed very low frequencies (34 of them were singletons).

Table A2 reports the frequencies of the 19 non-cSNS variant sites detected in this study: 16 intronic (12 SNSs, three STRs, and one nucleotide insertion) and three exonic (one SNS in the 5′UTR and two trinucleotide deletions in the coding sequence).

The density of polymorphic SNSs in the coding and in the noncoding regions turned out to be compatible (12/4443 = 1/370 and 4/2367 = 1/592 bp, respectively; $P\approx0.4$); on the contrary, the density of rare SNSs appeared to be three-fold higher in the coding region (49/4443 = 1/91 bp in the exons and 9/2367 = 1/263 bp in the introns; $P\approx0.002$).

It has been possible to classify as not fully penetrant CF-causing alleles, on the basis of their frequencies, six cSNSs (ref. nos. 6, 20, 26, 29, 37, and 59, black arrows in Table A1), besides the four already classified in the previous investigation[2] (ref. nos. 16, 32, 54, and 60, white arrows in Table A1), and two noncoding variants (ref. letter E and O in Table A2).

The availability of a large number of mutants collected on a random sample of individuals made it possible to perform a comparison between the indirectly estimated relative mutation rates of the 12 possible type of substitutions (Figure 1). The expected numbers of cSNSs have been computed assuming that μ is the same for all of them and that all the mutational events had the same probability to be detected; therefore, since the four nonsynonymous cSNP (ref. nos. 6, 16, 26 and 29) may have been not neutral they have been excluded, and the total number of cSNSs was 57 instead of 61. The T↔A rate was much lower than

Table 1 Frequencies of the CFTR variants within the M or the V alleles

Ref. No.[a]	exon or intron	VARIANT SITES	ABSOLUTE AND RELATIVE FREQUENCIES in the M genes (MM subjects)	ABSOLUTE AND RELATIVE FREQUENCIES in the V genes (VV subjects)	distance from the M470V site[b] (Kb)
A	5' UTR	125 g/c	8/144 (0.056)	3/356 (0.008)	-80
1	2	R31C	5/226 (0.004)	1/576 (0.002)	-56
6	2	R75Q	1/226 (0.004)	**15/576 (0.026)**	-51
7	3	G85E	0/226 (0)	1/576 (0.002)	-51
B	i 3	406-6 t/c	0/226 (0)	6/576 (0.010)	-29
8	4	R117H	2/226 (0.009)	0/576 (0)	-29
10	4	I148T	**3/224 (0.013)**	0/576 (0)	-29
C	i 4	621+3 a/g	1/224 (0.004)	0/576 (0)	-29
12	5	R170H	1/158 (0.006)	0/402 (0)	-26
D	i 6a	875+40 a/g	**6/36 (0.167)**[c]	0/118 (0)[c]	-25
	i 6b	$(ttga)_6$	**13/36 (0.361)**	1/118 (0.008)	-23
E	i 6b	1001+11 c/t	**5/60 (0.083)**	0/166 (0)	-23
F	i 8	1341+28 c/t	1/152 (0.007)	0/464 (0)	-18
	i 8	$(tg)_{10}$	**39/76 (0.513)**	5/218 (0.023)	-11
	i 8	$(tg)_{11}$	21/76 (0.276)	**205/218 (0.940)**	-11
	i 8	$(tg)_{12}$	**16/76 (0.211)**	8/218 (0.037)	-11
	i 8	t_5	4/76 (0.053)	2/218 (0.009)	-11
	i 8	t_7	48/76 (0.632)	**214/218 (0.982)**	-11
	i 8	t_9	**24/76 (0.316)**	2/218 (0.009)	-11
16	10	M470V	←		
H	ex 10	F508del	**3/226 (0.013)**	0/572 (0)	0
19	10	F508C	0/226 (0)	1/572 (0.002)	0
20	10	1716g/a	**15/226 (0.066)**	0/572 (0)	0
21	11	G542X	1/158 (0.006)	0/400 (0)	+28
24	12	V562I	1/226 (0.004)	0/576 (0)	+30
25	12	V562L	1/226 (0.004)	0/576 (0)	+30
26	12	G576A	**3/226 (0.013)**	0/576 (0)	+30
28	13	2082c/t	1/104 (0.010)	0/226 (0)	+32
29	13	R668C	**3/224 (0.013)**	0/562 (0)	+32
32	14a	2694t/g	**45/70 (0.643)**	9/208 (0.043)	+35
I	i 14a	2752-15 c/g	0/226 (0)	5/576 (0.009)	+44
37	15	3030g/a	1/158 (0.006)	7/402 (0.017)	+44
O	i 15	3041-71 g/c	**5/226 (0.022)**	0/576 (0)	+47
39	17a	L997F	1/226 (0.004)	4/576 (0.007)	+51
40	17a	A1009T	0/226 (0)	1/572 (0.002)	+51
42	17b	F1052V	1/226 (0.004)	0/572 (0)	+52
43	17b	G1069R	1/226 (0.004)	0/572 (0)	+52
44	17b	Q1071H	1/226 (0.004)	0/572 (0)	+52
45	17b	3417a/t	0/226 (0)	4/572 (0.007)	+52
46	17b	L1096R	1/226 (0.004)	0/572 (0)	+52
52	19	3813a/g	0/118 (0)	1/484 (0.002)	+68
53	19	S1235R	**3/100 (0.030)**	0/294 (0)	+68
54	20	4002a/g	**5/56 (0.089)**	1/168 (0.006)	+83
56	21	4029a/g	0/194 (0)	3/506 (0.006)	+93
57	21	N1303K	1/92 (0.011)	0/272 (0)	+93
59	24	4404c/t	3/226 (0.013)	14/576 (0.024)	+107
60	24	4521g/a	**21/56 (0.375)**	2/172 (0.012)	+107

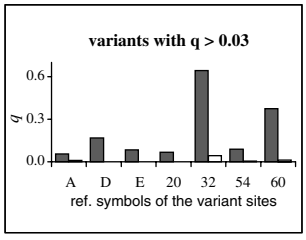
"slow evolution" markers

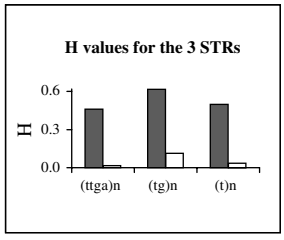
H associated with the….	
M	V
2.214	0.362

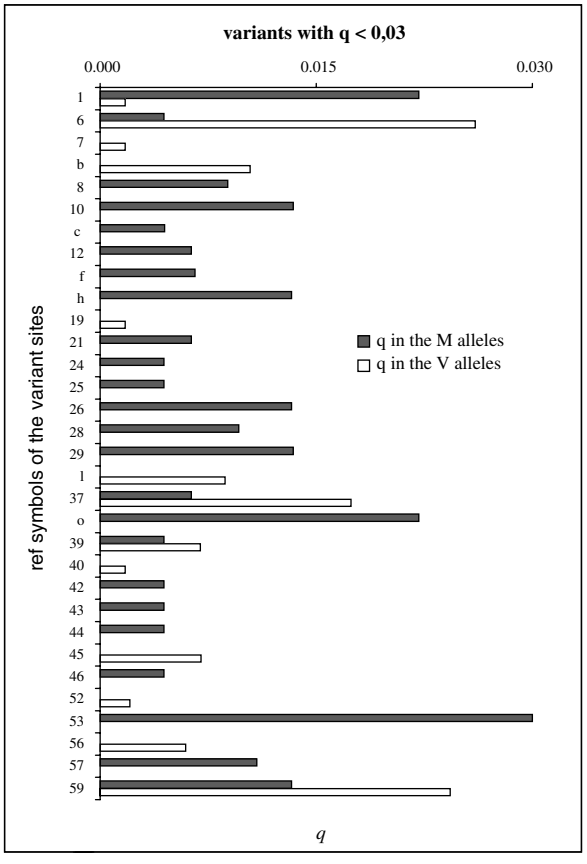
"fast evolution" markers (i.e. STRs)

Polymorphic sites	H in M genes	H in V genes
$(ttga)_n$	0.461	0.017
$(tg)_n$	0.616	0.114
$(t)_n$	0.499	0.036

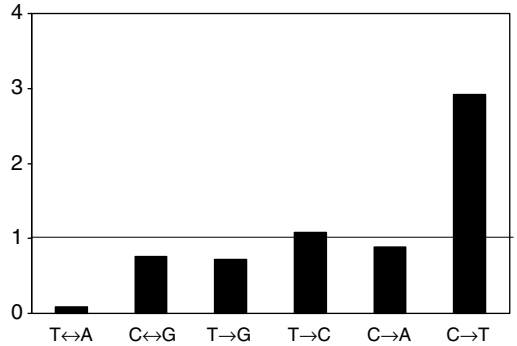
H is the sum of the degrees of heterozygosity of all the markers





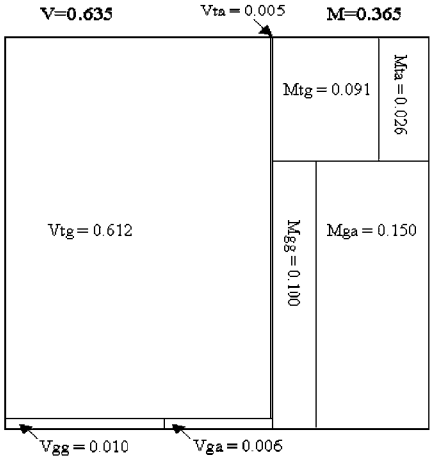


Figures referring to sites where q_M and q_V are very different are boldfaced.
[a]Reference symbol: numbers refer to cSNSs (Table A1) and letters to other variants (Table A2).
[b]Derived from the GDB sequence accession no. AC000061, AC000111.
[c]The Czech Republic has been excluded due to its very different frequency.



	T↔A	C↔G	T→G + A→C	T→C + A→G	C→A + G→T	C→T + G→A
obs/exp	1/11.1	6/7.9	8/11.1	12/11.1	7/7.9	23/7.9
P	0.002	0.502	0.351	0.790	0.752	$7x10^{-8}$

Figure 1 Indirect relative mutation rate estimates of the 12 types of cSNSs. The expected number for each of 12 cSNS, say X→Y, has been obtained by multiplying the proportion of X among the 4443 coding CFTR bp (T = 1236; C = 873; A = 1362 and G = 972) by 57 (the number of cSNSs) and dividing this figure by 3 (each nucleotide can mutate to the other 3). Complementary cSNSs have been combined because the two DNA strands exhibited compatible mutational behaviours (expressed by the ratio obs/exp).



Two sites haplotypes

sites (ref.no.s)	haplotype	exp. freq	obs. freq	D	D'
16-32	Mg	0.097	0.249	0.152	0.905
	Mt	0.268	0.116		
	Vg	0.168	0.016		
	Vt	0.467	0.619		
16-60	Ma	0.068	0.176	0.107	0.902
	Mg	0.297	0.190		
	Va	0.119	0.012		
	Vg	0.516	0.623		
32-60	ga	0.050	0.156	0.106	0.774
	gg	0.215	0.109		
	ta	0.138	0.031		
	tg	0.598	0.704		

Three sites haplotypes

haplotype (16-32-60)	exp. freq	obs. freq	D
Mga	0.018	0.150	0.132
Mgg	0.079	0.100	0.021
Mta	0.050	0.026	-0.024
Mtg	0.218	0.091	-0.127
Vga	0.032	0.006	-0.025
Vgg	0.137	0.010	-0.127
Vta	0.087	0.005	-0.082
Vtg	0.379	0.612	0.234

site	cSNS	allele	freq
16	M470V	M	0.365
32	2694t/g	g	0.265
60	4521g/a	a	0.187

Figure 2 Haploid assortments for the three highly polymorphic cSNSs of the CFTR gene. Lack of haplotype variability associated to the V470 allele compared with M470. V and M indicate the M470V alleles. Three letter words indicate haplotypes; V or M in the first position: M470V (site 16); t or g in the second position: 2694 t/g (site 32); a or g in the third position: 4521 a/g (site 60). The areas indicate haplotype frequencies. *D* and *D'* = absolute and relative linkage disequilibrium values, respectively. Reference numbers are as in Table A1. Accession numbers for these three highly polymorphic cSNSs in the dbSNP public database (http://www.ncbi.nlm.nih.gov/SNP/) are: rs213950 for M470V, rs1042077 for 2694t/g, and rs2800136 for 4521g/a.

expected (obs. = 1; exp. = 11.1; $P \approx 0.002$; ≈ 0.02 with the Bonferroni correction). The combined rate of the complementary C→T and G→A SNSs is about threefold higher than expected (obs. = 23; exp. = 7.9; $P \approx 10^{-7}$), confirming already known notions.[6–8] It is commonly accepted that the strong excess of these two SNSs is due to a particularly high probability of the C nucleotide (both in the sense and in the nonsense DNA strand), when it is followed by G, to mutate to T (see, for example, Cooper and Krawczak[9] and Cooper *et al*[10]). This is strongly supported by present data. In fact, since 873 is the number of C in the 4443 coding bp of the CFTR gene, 873 is the number of CpN dinucleotides of this gene. Among them only 57 (6.5%) are CpG, whereas six out of the eight (75%) C→T mutations of the present study were in a CpG dinucleotide ($P \approx 0$). Similarly, the total number of NpG dinucleotides is 972 and only 57 are CpG (5.9%), whereas five out of the 15 (33.3%) G→A mutations of the present study were in a CpG dinucleotide ($P \approx 10^{-4}$).

An almost complete linkage disequilibrium (LD) between the M470V and the two other highly polymorphic cSNS sites of the gene (ref. nos. 32 and 60) has been observed. These LDs ($D' = 0.91$ and 0.90, respectively) are shown in Figure 2; they are not due to blocks of absence of recombination.[11–13] The two sites 32 and 60, in fact, are not in strong disequilibrium between themselves in the CFTR genes carrying the M470 allele, thus suggesting that they are very ancient. This situation is strongly reminiscent of that of Rh, the genetic system where the LD phenomenon was first discovered.[14,15] In fact, this system too consists of one polymorphic locus, D/d, and of two additional loci, C/c and E/e, which are highly polymorphic within the chromosomes carrying the D allele and barely polymorphic within the chromosomes with the d allele.

The large number of CFTR genes studied allowed us to subdivide the total sample into two subsamples consisting of genes carrying the M470 or the V470 allele, respectively. Table 1 compares the degree of variability of the CFTR genes in these two subsamples. It clearly appears that the CFTR genes carrying the M allele are much more variable than those carrying the V allele for most of the markers suitable for such comparison (ie those for which a variant

Table 2 The intensity of the M-restricted variability depends on the distance from the M470V site

Distance from the M470V site (kb)	*Number of sites variable only within the M genes*	*Number of all other variable sites*
<50	16	5
≥ 50	6	12

$P \approx 0.007$.

allele was found in at least one MM or VV homozygote, respectively) both for the 'slow' (mutation rate in the order of 10^{-8} to 10^{-7}; $n=39$) and for the 'fast' (mutation rate 10^{-4} to 10^{-2}; $n=3$; ie STRs) evolution markers.[16] Thus, the estimate of the overall variability of the CFTR gene is the weighted mean of two very different patterns of variability: that of the CFTR genes carrying the M and that of the CFTR genes carrying the V allele, plus, obviously, that due to the M470V site itself. These findings show the existence of an 'extended haplotype homozygosity' region (EHH),[17] namely of an almost 'allele-restricted' monomorphic region concerning only the CFTR genes with the V allele. Such strong preferential concentration of variability within the M CFTR genes turned out to be correlated with the distance from the 470 site being stronger in the DNA sequence around ± 50 kb from it (Table 2).

Discussion

Sabeti *et al*[17] suggested that an EHH implies a recent positive selection, and verified this hypothesis in two genes (G6PD and TNSF5) known for having been recently subjected to positive selection. Thus, the present finding of an EHH region encompassing the M470V site strongly suggests that the CFTR gene recently underwent selection. This suggestion is in accordance with previous findings of extended homogeneous haplotypes associated with specific CF mutations.[18,19]

Some features of the CFTR gene suggest a possible scenario for the selection process:

(1) M470 is the ancestral allele. It is in fact the allele found in all the other species studied so far;[20–22]
(2) M470 is almost fixed among the sub-Saharian Africans: the combined V frequency in the three sub-Saharian African populations we have studied (Mossì, Burkina Faso, $n=146$ individuals; Ewondo, Ghana, $n=10$; Pygmies, CAR, $n=10$; unpublished data) was 0.02 ± 0.01. This high prevalence of M470 presumably applies to all sub-Saharian populations.
(3) the V470 allele outside of Africa is very frequent, it is even more common than the M allele (eg for the Europeans[1] and for the Asians[23]);
(4) the bulk of CFTR gene variability is restricted to the haplotypes carrying the M470 allele (Table 1).

The time elapsed since the radiation of *H. sapiens* from Africa to the rest of the Old World (only two/four thousand generations[24]) has been far too short to account in terms of genetic drift only[25] for such a tremendous increase of the V allele frequency. Therefore, a selective process seems more likely. As far as the time of onset of the selection process, the great extension of the region encompassing the M470V site with an almost complete LD suggests that it is recent (see also Slatkin and Bertorelle[26]). We wish to suggest the involvement of a selectively advantageous mutation X that would have caused, in relatively few generations, the increase of the V allele frequency outside of Africa.

V allele frequency could have increased by one of the following three possible mechanisms, the first one relates to the V mutation itself, the other two pertain to the hitchhiking phenomenon:[27]

(1) X is V. The V allele is very common outside of Africa because only there it has been advantageous. There are indications that the V allele might produce a less functional protein. The V470 allele has, in fact, been reported to have a 1.7 times lower intrinsic chloride channel activity,[28] as confirmed by different studies.[23,29,30] It might have conferred a selective advantage in particular environments as suggested for CF heterozygotes and tubercolosis,[31] or lung infections,[32] or diarrhea caused by enterotoxic bacteria;[33,34]
(2) X is not V, and was already present in Africa, in CFTR genes with the V allele, before the migration of *H. sapiens*, but it did not confer any selective advantage. Its frequency increased dramatically in Europe following human exposure to different environmental conditions that made it advantageous;
(3) X is not V, and was born in Europe, in one CFTR gene with the V allele, before the migration of *H.sapiens* towards Asia.

The first two possibilities require the additional hypothesis that the African V was carried by only one haplotype, while the third possibility is independent from the number of different African haplotypes carrying the V allele.

A choice among these three possibilities would require, at least, the ascertainment of variability, if any, of the African haplotypes carrying V allele.

Acknowledgements

This work was funded by the The Italian Ministry of University and Research; the Italian Ministry of Health, 'National Project for Standardization and Quality Assurance of Genetic Tests' (D.lg 505/92); The Italian Cystic Fibrosis Research Foundation; MZCR IGA 1A/8236-3, 00000064203 and EC-CF Chip and CRMGEN to MM.

References

1 Modiano G, Bombieri C, Ciminelli BM *et al*: A large-scale study of the random variability of a coding sequence: a study on the CFTR gene. *Eur J Hum Genet* 2005; **13**: 184–192.
2 Bombieri C, Giorgi S, Carles S *et al*: A new approach for identifying non-pathogenic mutations. An analysis of the cystic fibrosis transmembrane regulator gene in normal individuals. *Hum Genet* 2000; **106**: 172–178.
3 WHO Report: The molecular genetic epidemiology of cystic fibrosis, 2004, www.who.int/genomics/publications/en/.
4 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.
5 Schneider S, Roessli D, Excoffier L: *Arlequin ver. 2000: a software for population genetics data analysis*. Switzerland: Genetics and Biometry Laboratory, University of Geneva.
6 Modiano G, Battistuzzi G, Motulsky AG: Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc Natl Acad Sci* 1981; **78**: 1110–1114.
7 Stephens JC, Schneider JA, Tanguay DA *et al*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293**: 489–493.
8 Strachan T, Read AP: *Human Molecular Genetics*, 3rd edn. London and New York: Garland Science, 2004.
9 Cooper DN, Krawczak M: *Human gene mutation*. Oxford: BIOS Scientific Publishers, 1993.
10 Cooper DN, Krawczak M, Antonarakis SE: The nature and mechanisms of human gene mutation; in Scriver C, Beaudet AL, Sly WS, Valle D (eds): *Metabolic and molecular bases of inherited disease*. New York: McGraw Hill, 1995, pp 259–291.
11 Reich DE, Cargill M, Bolk S *et al*: Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
12 Goldstein DB: Islands of linkage disequilibrium. *Nat Genet* 2001; **29**: 109–111.
13 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
14 Race RR: The Rh genotypes and Fisher's theory. *Blood* 1948; **3** (suppl 2): 27–42.
15 Race RR, Sanger R: *Blood groups in man*. Oxford: Blackwell Scientific Publication, 1958.
16 Jobling MA, Hurles ME, Tyler-Smith C: *Human evolutionary genetics: origins, peoples and disease*. New York: Garland Science, Taylor & Francis Group, 2004.
17 Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
18 Dork T, Neumann T, Wulbrand U *et al*: Intra- and extragenic marker haplotypes of CFTR mutations in cystic fibrosis families. *Hum Genet* 1992; **88**: 417–425.
19 Sereth H, Shoshani T, Bashan N, Kerem BS: Extended haplotype analysis of cystic fibrosis mutations and its implications for the selective advantage hypothesis. *Hum Genet* 1993; **92**: 289–295.
20 Tucker SJ, Tannahill D, Higgins CF: Identification and developmental expression of the *Xenopus laevis* cystic fibrosis transmembrane conductance regulator gene. *Hum Mol Genet* 1992; **1**: 77–82.
21 Vuillaumier S, Kaltenboeck B, Lecointre G, Lehn P, Denamur E: Phylogenetic analysis of cistyc fibrosis transmembrane conductance regulator gene in mammalian species argues for the development of a rabbit model for cystic fibrosis. *Mol Biol Evol* 1997; **14**: 372–380.
22 Wine JJ, Glavac D, Hurlock G *et al*: Genomic DNA sequence of Rhesus (*M. mulatta*) cystic fibrosis (CFTR) gene. *Mamm Genome* 1998; **9**: 301–305.
23 Lee JH, Choi JH, Namkung W *et al*: A haplotype-based molecular analysis of CFTR mutations associated with respiratory and pancreatic diseases. *Hum Mol Genet* 2003; **12**: 2321–2332.
24 Klein RG: *The human career. Human biological and cultural origins*. Chicago: The University Chicago Press, 1989.
25 Neuhauser C: Mathematical models in population genetics; in Balding DJ *et al* (eds):: *Handbook of statistical genetics*. New York John Wiley & Sons, 2001, pp 153–177.
26 Slatkin M, Bertorelle G: The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* 2001; **158**: 865–874.
27 Wagener DK, Cavalli-Sforza LL: Ethnic variation in genetic disease: possible roles of hitchhiking and epistasis. *Am J Hum Genet* 1975; **27**: 348–364.
28 Cuppens H, Lin W, Jaspers M *et al*: Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes. The polymorphic (Tg)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation. *J Clin Invest* 1998; **101**: 487–496.
29 Noone PG, Pue CA, Zhou Z *et al*: Lung disease associated with the IVS8 5T allele of the CFTR gene. *Am J Respir Crit Care Med* 2000; **162**: 1919–1924.
30 Wei L, Vankeerberghen A, Jaspers M, Cassiman J, Nilius B, Cuppens H: Suppressive interactions between mutations located in the two nucleotide binding domains of CFTR. *FEBS Lett* 2000; **473**: 149–153.
31 Meindl RS: Hypothesis: a selective advantage for cystic fibrosis heterozygotes. *Am J Phys Anthropol* 1987; **74**: 39–45.
32 Pier GB, Grout M, Zaidi TS *et al*: Role of mutant CFTR in hypersusceptibility of cystic fibrosis patients to lung infections. *Science* 1996; **271**: 64–67.
33 Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ: Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* 1994; **266**: 107–109.
34 Quinton PM: Human genetics. What is good about cystic fibrosis? *Curr Biol* 1994; **4**: 742–743.

Appendix

Table A1 is an update of that already published in Modiano *et al*[1] and reports the absolute and relative frequencies of the 61 cSNSs (single-nucleotide substitutions in a coding sequence) found in a Czech sample, larger than that already published, together with the updated European frequencies. Table A2 reports the frequencies of the 19 non-cSNS variant sites detected in this study: 16 intronic (12 SNSs, three STRs, and one nucleotide insertion) and three exonic (one SNS in the 5′UTR and two trinucleotide deletions in the coding sequence).

Table A1 Fequencies of the 61 CFTR cSNSs

		cSNS[a] *variants*		*Czech Republic*		*Overall data*			
Ref. no.	*Exon*	*SNS*	*SAS*[b]	*genes n = 236*[c]	*Abs. freq.*	*Total sample size*	*q × 10³*	*SE × 10³*	
1	2	223C>T	R31C	1	12	2668	4.50	1.30	
2		224G>T	R31L	0	1	2168	0.46	0.46	
3		257C>T	S42F	0	1	2168	0.46	0.46	
4	3	334A>G	K68E	0	1	2668	0.37	0.37	
5		352C>T	R74W	0	1	2668	0.37	0.37	
6		356G>A	R75Q	2	40	2668	14.99	2.35	⬅
7		386G>A	G85E	0	2	2668	0.75	0.53	
8	4	482G>A	R117H	0	3	2466	1.22	0.70	
9		528T>G	I132M	0	1	2466	0.41	0.41	
10		575T>C	I148T	1	6	2466	2.43	0.99	
11	5	640C>T	R170C	0	1	1672	0.60	0.60	
12		641G>A	R170H	1	5	2166	2.31	1.03	
13	8	1281G>A	L383	0	1	1680	0.60	0.60	
14	9	1402G>A	G424S	0	1	1176	0.85	0.85	
15		1459G>T	D443Y	0	1	1176	0.85	0.85	
16	10	1540A>G	M470V[d]	165/428	736	1912	384.94	11.13	⇦
17		1598C>A	S489X	0	1	2610	0.38	0.38	
18		1648A>G	I506V	0	1	2610	0.38	0.38	
19		1655T>G	F508C	0/428	2	2466	0.81	0.57	
20		1716G>A	E528	11	62	2114	29.33	3.67	⬅
21	11	1756G>T	G542X	0/428	2	2340	0.85	0.60	
22		1764T>G	G544	0	1	2148	0.47	0.47	
23		1784G>A	G551D	0/428	1	2340	0.43	0.43	
24	12	1816G>A	V562I	0	1	2668	0.37	0.37	
25		1816G>C	V562L	0	3	2668	1.12	0.65	
26		1859G>C	G576A	0	27	2668	10.12	1.94	⬅
27	13	1997G>A	G622D	0	1	2238	0.45	0.45	
28		2082C>T	F650	0	1	2648	0.38	0.38	
29		2134C>T	R668C	0	32	2682	11.93	2.10	⬅
30		2377C>T	L748	0	1	624	1.60	1.60	
31	14a	2670G>A	W846X	0	1	1174	0.85	0.85	
32		2694T>G	T854	67	357	1174	304.09	13.43	
33		2695G>A	V855I	0	1	1174	0.85	0.85	
34	15	2816G>C	S895T	0	2	1160	1.72	1.22	
35		2831A>C	N900T	0	1	1160	0.86	0.86	
36		2988G>C	M952I	0	1	1160	0.86	0.86	
37		3030G>A	T966		25	1814	13.78	2.74	⬅
38		3032T>C	L967S	0	1	1160	0.86	0.86	
39	17a	3123G>C	L997F	0	17	2666	6.38	1.54	
40		3157G>A	A1009T	0	2	2666	0.75	0.53	
41		3212T>C	I1027T	0	1	2666	0.38	0.38	
42	17b	3286T>G	F1052V	0	3	2364	1.27	0.73	
43		3337G>A	G1069R	0	1	2364	0.42	0.42	
44		3345G>T	Q1071H	0	1	2364	0.42	0.42	
45		3417A>T	T1095	0	8	2670	3.00	1.06	
46		3419T>G	L1096R	0	1	2364	0.42	0.42	
47		3477C>A	T1115	0	1	2364	0.42	0.42	
48	18	3523A>G	I1131V	0	2	2072	0.97	0.68	
49		3586G>C	D1152H	0	1	1676	0.60	0.60	

Table A1 (*Continued*)

Ref. no.	Exon	cSNS[a] variants SNS	SAS[b]	Czech Republic genes $n=236$[c]	Abs. freq.	Overall data Total sample size	$q \times 10^3$	$SE \times 10^3$	
50	19	3617G>T	R1162L	0	2	2426	0.82	0.58	
51		3690A>G	Q1186	0	1	2426	0.41	0.41	
52		3813A>G	L1227	0	1	2426	0.41	0.41	
53		3837T>G	S1235R	1	15	2426	6.18	1.59	
54	20	4002A>G	P1290	3	37	1176	31.46	5.09	⇦
55	21	4009G>A	V1293I	0	1	1480	0.68	0.68	
56		4029A>G	T1299	0	8	2454	3.26	1.15	
57		4041C>G	N1303K	0	1	1480	0.68	0.68	
58		4085T>C	V1318A	0	1	1480	0.68	0.68	
59	24	4404C>T	Y1424	3	33	2636	12.52	2.17	⬅
60		4521G>A	Q1463	52	245	1144	214.16	12.13	⇦
61		4563T>C	D1477	0	1	1144	0.87	0.87	

Present data include those already reported in Modiano *et al*[1] and further data concerning the Czech Republic.
The ref. nos. are as in Modiano *et al*.[1]
[a]Single-nucleotide substitution in a coding sequence.
[b]Single- amino-acid substitution; for the synonymous the amino acid is indicated.
[c]This figure includes the 72 genes already reported in Modiano *et al*.[1]
[d]The frequency of the M allele is reported.
⬅: Variants classified as non-CF-causing in the present study.
⇦: Variants classified as non-CF-causing in the previous investigation (Bombieri *et al*[2]).

Table A2 Frequencies of the 19 CFTR gene variants other than cSNSs

Ref. no.	Variant	Exon or intron	Czech Republic abs freq	Czech Republic q	Central Italy[a] abs freq	Central Italy[a] q	NE Italy[a] abs freq	NE Italy[a] q	Southern France[a] abs freq	Southern France[a] q	Northern France abs freq	Northern France q	Overall data[a] abs freq	Overall data[a] q se	
A	125 g/c	5′ UTR	21	0.089	2	0.020	1	0.002	7	0.070	22	0.040	[b]		⇦
			236		**98**		**604**		**100**		**552**				
B	406-6 t/c	i 3	1	0.004	2	0.005	3	0.005	2	0.003	1	0.002	9	0.0038	
			236		**392**		**604**		**600**		**556**		**2388**	*0.0013*	
C	621+3 a/g	i 4	0	0.004	1	0.003	0		0		0		1	0.0004	
			236		**392**		**604**		**600**		**556**		**2388**	*0.0004*	
D	875+40 a/g	i 6a	2	0.008	5	0.047	11	0.110	5	0.050			21[c]	0.0686	⇦
			236		**106**		**100**		**100**		**0**		**306**[c]	*0.0144*	
	(ttga)n	i 6b	NT		11	0.104	17	0.170	7	0.071	NT		35	0.1151	⇦
					106		**100**		**98**				**304**	*0.0183*	
E	1001+11 c/t	i 6b	NT		4	0.036	10	0.100	6	0.060	38	0.069	58	0.0671	⬅
					110		**100**		**100**		**554**		**864**	*0.0085*	
F	1341+28 c/t	i 8	0	0.004	0		1	0.002	0		0		1	0.0006	
			236		**224**		**604**		**100**		**556**		**1720**	*0.0006*	
	(tg)n 9	i 8	1	0.002	0		0		0		NT		1	0.0014	
														0.0014	
	10		129	0.301	27	0.255	27	0.270	36	0.360			219	0.2984	⇦
														0.0169	
	11		267	0.624	71	0.670	61	0.610	58	0.580			457	0.6226	
														0.0179	
	12		31	0.072	8	0.075	12	0.120	6	0.060			57	0.0777	⇦
			428		**106**		**100**		**100**				**734**	*0.0099*	
	(t)n 5	i 8	3	0.022	2	0.019	8	0.080	4	0.040	NT		17	0.0385	⇦
														0.0091	
	7		118	0.868	97	0.916	82	0.820	86	0.860			383	0.8665	
														0.0162	
	9		15	0.110	7	0.066	10	0.100	10	0.100			42	0.0950	
			136		**106**		**100**		**100**				**442**	*0.0139*	

Table A2 (*Continued*)

Ref. no.	Variant	Exon or intron	Czech Republic abs freq	Czech Republic q	Central Italy[a] abs freq	Central Italy[a] q	NE Italy[a] abs freq	NE Italy[a] q	Southern France[a] abs freq	Southern France[a] q	Northern France abs freq	Northern France q	Overall data[a] abs freq	Overall data[a] q se	
G	I507del	ex 10	0	0.004	0		0		0		1	0.001	1	0.0004	
			236		**196**		**604**		**600**		**896**		**2532**	*0.0004*	
H	F508del	ex 10	3	0.013	2	0.010	2	0.003	10	0.017	17	0.019	34	0.0134	
			236		**196**		**604**		**600**		**896**		**2532**	*0.0023*	
I	1717−1 g/a	i 10	0		0		1	0.002	0		0		1	0.0004	
			144		**520**		**604**		**200**		**896**		**2364**	*0.0004*	
L	2752−15 c/g	i 14a	1	0.004	6	0.010	2	0.003	3	0.005	0	0.002	12	0.0046	
			236		**620**		**604**		**600**		**546**		**2606**	*0.0013*	
M	2752−97 c/t	i 14a	NT		NT		0		0		2	0.004	2	0.0011	
							600		**600**		**546**		**1746**	*0.0008*	
N	3040+23 t/c	i 15	0		0		0		0		1	0.002	1	0.0010	
			144		**100**		**100**		**100**		**536**		**980**	*0.0010*	
O	3041−71 g/c	i 15	2	0.008	6	0.010	3	0.005	9	0.015			20	0.0098	⬅
			236		**598**		**604**		**600**				**2038**	*0.0022*	
P	3272−26 a/g	i 17a	0		0		0		0		1	0.002	1	0.0005	
			72		**294**		**600**		**600**		**552**		**2118**	*0.0005*	
Q	4005+28 insA	i 20	0		1	0.005	0		0		0		1	0.0008	
			236		**222**		**100**		**100**		**554**		**1212**	*0.0008*	
R	4374+13 a/g	i23	NT		0		2	0.003	0		0		2	0.0015	
					100		**604**		**100**		**548**		**1352**	*0.0011*	

Figures in bold indicate the total number of genes studied.
For the length of the intronic sequences examined see Bombieri *et al*.[2]
NT = not tested.
⬅: Variants classified as non CF-causing in the present study.
⟸: Variants classified as non CF-causing in the previous survey (Bombieri *et al*[2]).
[a]These figures include those already published in Bombieri *et al*.[2]
[b]A pooled frequency has not been computed since the five samples are largely heterogenous.
[c]The Czech Republic has been excluded due to its very different frequency.