European Journal of Human Genetics (2005) 13, 184–192

www.nature.com/ejhg

ARTICLE

# A large-scale study of the random variability of a coding sequence: a study on the CFTR gene

Guido Modiano*[,1], Cristina Bombieri[2], Bianca Maria Ciminelli[1], Francesca Belpinati[2], Silvia Giorgi[1], Marie des Georges[3], Virginie Scotet[4], Fiorenza Pompei[1], Cinzia Ciccacci[1], Caroline Guittard[3], Marie Pierre Audrézet[4], Angela Begnini[2], Michael Toepfer[5], Milan Macek Jr[5], Claude Ferec[4], Mireille Claustres[3] 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, Italy;* [3]*Institute of Biology, University of Montpellier, France;* [4]*Centre de Biogénétique, CDTS, Brest, France;* [5]*Department of Molecular Genetics, CF-Centre, Charles University, Prague, Czech Republic*

**Coding single nucleotide substitutions (cSNSs) have been studied on hundreds of genes using small samples ($n_g \approx 100-150$ genes). In the present investigation, a large random European population sample (average $n_g \approx 1500$) was studied for a single gene, the *CFTR* (Cystic Fibrosis Transmembrane conductance Regulator). The nonsynonymous (NS) substitutions exhibited, in accordance with previous reports, a mean probability of being polymorphic ($q > 0.005$), much lower than that of the synonymous (S) substitutions, but they showed a similar rate of subpolymorphic ($q < 0.005$) variability. This indicates that, in autosomal genes that may have harmful recessive alleles (nonduplicated genes with important functions), genetic drift overwhelms selection in the subpolymorphic range of variability, making disadvantageous alleles behave as neutral. These results imply that the majority of the subpolymorphic nonsynonymous alleles of these genes are selectively negative or even pathogenic.**
*European Journal of Human Genetics* (2005) **13,** 184–192. doi:10.1038/sj.ejhg.5201306
Published online 10 November 2004

**Keywords:** common *vs* rare alleles; synonymous and nonsynonymous variability; CFTR gene

## Introduction

The quantitative study of the DNA random diversity can only be approached through a suitable 'window' ($n_{bp} \times n_g$) consisting of $n_{bp}$ base pairs, studied on a number $n_g$ of genomes. The whole spectrum of possible 'windows' ranges between two opposite and complementary 'extreme windows': those where many genes (that is many $n_{bp}$) are studied on few $n_g$ ('1st type loophole windows') and those where only a single gene is studied on a very large $n_g$ sample ('2nd type loophole windows').

The type of molecular variations most suitable for these studies are the single nucleotide substitutions in a coding sequence (cSNSs); (hereafter, when no ambiguity is possible, we shall also refer to them simply as substitutions) because the number and the consequences of cSNSs that may occur in a sequence of known length are known, not to mention that they include a perfectly defined 'control class' (the synonymous cSNSs) mainly consisting of neutral mutations (although it is known that some of them affect the splicing: for the possible molecular mechanisms see, eg, Pagani *et al*[1,2]).

The majority of the previous investigations[3–9] adopted the '1st-type loophole' approach ($n_g \approx 100$); thus, in spite of

*Correspondence: Professor Guido Modiano, Department of Biology, University 'Tor Vergata' - Rome, Via della Ricerca Scientifica, 00133 Roma, Italy. Tel: +390672594341; Fax: +39062023500;
E-mail: modiano@uniroma2.it
Received 3 December 2003; revised 2 August 2004; accepted 5 August 2004

their extensiveness ($n_{bp} \approx 100$ Kb), which produced exhaustive information about the common variation, they could not, in principle, provide *direct* information about the range of rare variability. More recently, few studies[10–12] have adopted a less extreme approach by studying some genes on hundreds of random individuals from the major human groups (Africans, Europeans, Asians).

The present investigation adopted the '2nd-type loophole approach'. An average sample of 1550 European genomes was studied by examining the pattern of common (=polymorphic; $q > 0.005$) as well as rare (=sub polymorphic) variability, of a single gene (the *CFTR* gene, *Cystic Fibrosis Transmembrane conductance Regulator*, consisting of 4.5 coding kb) which belongs to the class of autosomal nonduplicated genes performing an essential function, whose deleterious alleles are mainly recessive.

The low limit of the range of variation, which one can consider as reliably explored with a given $n_g$ is that where the variants with the lowest $q$ are expected to be found at least ca. 5 times ($q \times n_g \approx 5$): with $n_g \approx 1550$, as in the present investigation, the lowest end of the range corresponds to $q \approx 0.003$, instead of 0.05 (corresponding to $n_g = 100$).

The large body of data generated by this investigation allowed us to study directly, for the first time, an issue of general relevance in molecular evolution, namely the pattern of nonsynonymous (NS) and synonymous (S) substitutions in the range of *rare* variability; it has also allowed us to evaluate the rate of misclassification (rare or common) of variants as a function of sample size.

## Materials and methods

In this study, we systematically explored all the 27 exons of the *CFTR* gene, using denaturing gradient gel electrophoresis (DGGE) or denaturing high-performance liquid chromatography (DHPLC), genotyping methods with a very high efficiency (about 98%, see Bombieri *et al*[13]) in detecting molecular variation. Preliminary results on a subset of 400 individuals, and details on the method, are in Bombieri *et al*.[13]

### The sample

All the individuals studied (present investigation and previous data[13]) come from six geographical areas, namely, Northern Italy (Verona), Central Italy (Rome), Southern France (Montpellier), Northern France (Brest), the Czech Republic (Prague) and Spain (Barcelona). All individuals gave their informed consent.

### Mutation analysis

Genomic DNA was extracted from blood samples, amplified *in vitro* by PCR and analyzed by DGGE as previously reported[13] or DHPLC. Every mutant discovered by these methods was sequenced with the ABI PRISM 377 or 310 Sequence Analyser. The following cSNSs, numbered as given in Table 1, were studied by RFLP analysis: no. 1, see Ghanem *et al*;[14] nos 20 and 59 see Fanen *et al*;[15] nos 37, see Chillon *et al*[16] and nos. 12, 24–26, 28, 29, 45, 48, 56, 60, methods available on request (cristina.bombieri@medgen.univr.it).

### Subdivision of the total number of the theoretically possible cSNSs into NS and S cSNSs

The *CFTR* coding sequence consists of 4443 bp (1480 sense codons + a stop codon); thus the total number of possible cSNSs is $(4443 \times 3) = 13\,329$. To compute exactly how many of these 13 329 cSNSs would be NS and how many would be S, we considered the *CFTR* codon usage, rather than simply its amino-acid composition, as it had been done commonly. The total number of *possible* NS cSNSs, $N_{NS}$, is 10 408 (including the eight stop codon→aa codon, and the 650 aa codon→stop codon) and that of the synonymous cSNSs, $N_S$, is 2921 (including the TGA stop codon→TAA stop codon). If the mutation rate $\mu$ were the same for NS and S substitutions and, once the mutations occurred, both types of substitutions have the same probability of being sampled, the expected ratio $N_{NS}/N_S$ would be 3.56 (=10 408/2921). This ratio should be valid for both subpolymorphic and polymorphic cSNSs.

These figures can also be used to estimate the probability of being polymorphic or subpolymorphic for the NS ($P^{NS}_{Poly}$ and $P^{NS}_{SubPoly}$) and for the S ($P^{S}_{Poly}$ and $P^{S}_{SubPoly}$) substitutions. For example, the estimate of $P^{S}_{Poly}$ is the ratio between the number of S substitutions that have been shown to be polymorphic and 2921. It may be worth to point out that the herein adopted definition of polymorphic for an allele is only based on its frequency, disregarding any of its possible phenotypic effects.

### Estimate of $\theta$ and $\pi$ and of the distribution of variant frequencies

Under neutrality, both $\pi$ (mean heterozygosity per site) and $\theta$ (number of segregating sites) are expected to be independent from the sample size and equal to $4N_e\mu$.[17–19] These parameters can be estimated as follows:

$$\pi = \sum_{i=1}^{n} Hi/n_{bp},$$

where $H_i$ is the $2pq$ observed for each of the $n$ cSNS detected (see, the last two columns of Table 1) and $n_{bp}$ is the length of sequence studied;

$$\theta = \left[ n_{cSNS} / \left( \sum_{i=1}^{n-1} i^{-1} \right) \right] \Big/ n_{bp},$$

where $\left(\sum_{i=1}^{n-1} i^{-1}\right)$ is a factor that should counterbalance the expected increase of $n_{cSNS}$ associated with the increase of $n_g$.

In the Tajima's test,[19] the null hypothesis of neutrality is rejected if a statistically significant difference between $\pi$

**Table 1** List of the 61 cSNSs[a] encountered in the present survey

| | | | *cSNS variants found* | | *The random samples of genes (and the technique utilized)* | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *NE Italy (DGGE)* | | *Central Italy (DGGE)* | | *Southern France (DGGE)* | | *Northern France (DHPLC)* | | *Spain (SSCA)* | *Czechia (DGGE)* | | | | $H^b \times 10^4$ | |
| *Exon* | *Exon Length (bp)* | *Ref. no.* | *SNS* | *SAS*[c] | *1st 100*[d] | *2nd 500* | *1st 100*[d] | *2nd*[e] | *1st 100*[d] | *2nd 500* | *1st 100* | *2nd*[e] | *82*[d] | *72* | *Abs. Freq.* | *Total sample size* | $q \times 10^4$ | $se \times 10^4$ | *NS*[f] | *S*[f] |
| 1[g] | 53 | | | | 0 | | 0 | | 0 | | 0 | 0/452 | | 0 | | 924 | | | | |
| 2 | 111 | 1 | 223C>T | R31C | 1 | | 1 | 1/500 | 1 | 1 | 0 | 0/450 | 0 | | 5 (11) | 1 932 (2 432) | 45.23 | 13.61 | 90 | |
| | | 2 | 224G>T | R31L | 0 | | 0 | 0/500 | 0 | 0 | 0 | 1/450 | 0 | | 1 | 1 932 | 5.17 | 5.17 | 10 | |
| | | 3 | 257C>T | S42F | 0 | | 0 | 1/500 | 0 | 0 | 0 | 0/450 | 0 | | 1 | 1 932 | 5.17 | 5.17 | 10 | |
| 3 | 109 | 4 | 334A>G | K68E | 1 | 0 | 0 | 0/498 | 0 | 0 | 0 | 0/452 | 0 | 0 | 1 | 2 504 | 3.99 | 3.99 | 8 | |
| | | 5 | 352C>T | R74W | 0 | 0 | 0 | 0/498 | 0 | 0 | 0 | 1/452 | 0 | 0 | 1 | 2 504 | 3.99 | 3.99 | 8 | |
| | | 6 | 356G>A | R75Q | 1 | 7 | 1 | 7/498 | 2 | 9 | 2 | 9/452 | 0 | 2 | 40 (40) | 2 504 (2 544) | 157.23 | 24.66 | 310 | |
| | | 7 | 386G>A | G85E | 0 | 0 | 1 | 1/498 | 0 | 0 | 0 | 0/452 | 0 | 0 | 2 | 2 504 | 7.99 | 5.65 | 16 | |
| 4 | 216 | 8 | 482G>A | R117H | 0 | 0 | 0 | 0/292 | 0 | 2 | 0 | 1/456 | 0 | 0 | 3 | 2 302 | 13.03 | 7.52 | 26 | |
| | | 9 | 528T>G | I132M | 0 | 0 | 0 | 0/292 | 0 | 0 | 0 | 1/456 | 0 | 0 | 1 | 2 302 | 4.34 | 4.34 | 8 | |
| | | 10 | 575T>C | I148T | 1 | 2 | 0 | 1/292 | 0 | 0 | 0 | 1/456 | 0 | 1 | 6 | 2 302 | 26.06 | 10.63 | 52 | |
| 5 | 90 | 11 | 640C>T | R170C | 0 | 0 | 0 | 0/6 | 0 | | 0 | 1/448 | 0 | | 1 | 1 436 | 6.96 | 6.96 | 14 | |
| | | 12 | 641G>A | R170H | 1 | 1 | 0 | 0/6 | 0 | | 0 | 2/448 | 0 | | 4 (4) | 1 436 (1 930) | 20.73 | 10.35 | 41 | |
| 6a | 164 | | | | 0 | | 0 | 0/6 | 0 | | 0 | 0/432 | 0 | 0 | | 992 | | | | |
| 6b | 126 | | | | 0 | | 0 | 0/6 | 0 | | 0 | 0/454 | 0 | | | 942 | | | | |
| 7 | 247 | | | | 0 | | 0 | 0/6 | 0 | | 0 | 0/796 | 0 | | | 1 284 | | | | |
| 8 | 93 | 13 | 1281G>A | L383 | 0 | 0 | 0 | 0/6 | 0 | | 0 | 1/456 | 0 | 0 | 1 | 1 516 | 6.60 | 6.60 | | 13 |
| 9 | 183 | 14 | 1402G>A | G424S | 0 | | 0 | 0/6 | 0 | | 0 | 1/454 | 0 | | 1 | 940 | 10.64 | 10.64 | 21 | |
| | | 15 | 1459G>T | D443Y | 0 | | 0 | 0/6 | 0 | | 0 | 1/454 | 0 | | 1 | 940 | 10.64 | 10.64 | 21 | |
| 10 | 192 | 16 | 1540A>G | M470V[h] | 42 | 197 | 30 | 37/96 | 39 | 199 | (i) | (i) | 27 | | 571(736) | 1 484 (1 912) | 3849.37 | 111.28 | 4 735 | |
| | | 17 | 1598C>A | S489X | 0 | 0 | 0 | 0/96 | 0 | 0 | 0 | 1/796 | 0 | | 1 | 2 374 | 4.21 | 4.21 | 8 | |
| | | 18 | 1648A>G | I506V | 1 | 0 | 0 | 0/96 | 0 | 0 | 0 | 0/796 | 0 | | 1 | 2 374 | 4.21 | 4.21 | 8 | |
| | | 19 | 1655T>G | F508C | 0 | 1 | 0 | 0/96 | 0 | 0 | 0 | 1/796 | 0 | | 2 | 2 038 | 8.42 | 5.96 | 17 | |
| | | 20 | 1716G>A | Q528 | 2 | 16 | 1 | 0/96 | 0 | 19 | i | i | 5 | | 43 (58) | 1 478 (2 024) | 286.56 | 37.08 | | 557 |
| 11 | 95 | 21 | 1756G>T | G542X | 0 | 2 | 0 | 0/134 | 0 | | 0 | 0/796 | 0 | 0 | 2 | 1 984 | 10.08 | 7.12 | 20 | |
| | | 22 | 1764T>G | G544 | 0 | 0 | 0 | 0/134 | 0 | | 0 | 1/796 | 0 | 0 | 1 | 1 984 | 5.04 | 5.04 | | 10 |
| | | 23 | 1784G>A | G551D | 0 | 0 | 0 | 0/134 | 0 | | 0 | 1/796 | 0 | 0 | 1 | 1 984 | 5.04 | 5.04 | 10 | |
| 12 | 87 | 24 | 1816G>A | V562I | 0 | 0 | 0 | | 0 | 1 | 0 | 0/450 | 0 | 0 | 1 (1) | 2 004 (2 504) | 3.99 | 3.99 | 8 | |
| | | 25 | 1816G>C | V562L | 0 | 0 | 0 | | 1 | 0 | 0 | 1/450 | 0 | 0 | 2 (3) | 2 004 (2 504) | 11.98 | 6.91 | 24 | |
| | | 26 | 1859G>C | G576A | 1 | 2 | 0 | | 1 | 11 | 0 | 8/450 | 0 | 0 | 23 (27) | 2 004 (2 538) | 106.38 | 20.36 | 213 | |
| 13 | 724[j] | | | | | | | | | | | | | | | | | | | |
| | 449 | 27 | 1997G>A | G622D | 0 | 0 | 0/80 | 0/96 | 1 | 0 | 0 | 0/444 | 0 | | 1 | 2 002 | 5.00 | 5.00 | 10 | |
| | | 28 | 2082C>T | F650 | 1 | 0 | 0/80 | 0/20 | 0 | 0 | 0 | 0/444 | 0 | | 1 (1) | 1 926 (2 412) | 4.15 | 4.15 | | 8 |
| | | 29 | 2134C>T | R668C | 1 | 2 | 0/80 | 0/96 | 1 | 11 | 0 | 12/444 | 0 | | 27(32) | 2 002 (2 558) | 125.10 | 21.98 | 247 | |
| | 275 | 30 | 2377C>T | L748 | 0 | | 0 | 0/6 | 0 | | | | 1 | | 1 | 388 | 25.77 | 25.77 | | 52 |
| 14a | 129 | 31 | 2670G>A | W846X | 0 | | 0 | 0/6 | 0 | | 1 | 0/452 | 0/80 | 0 | 1 | 1 010 | 9.90 | 9.90 | 20 | |
| | | 32 | 2694T>G | T854 | 33 | | 23 | 0/6 | 33 | | 38 | 149/452 | 14/80 | 11 | 301 | 1 010 | 2980.20 | 143.92 | | 4 184 |
| | | 33 | 2695G>A | V855I | 0 | | 0 | 0/6 | 0 | | 0 | 1/452 | 0/80 | 0 | 1 | 1 010 | 9.90 | 9.90 | 20 | |
| 14b | 38 | | | | 0 | 0 | 0 | 0/520 | 0 | 0 | 0 | 0/446 | 0 | | | 2 448 | | | | |
| 15 | 251 | 34 | 2816G>C | S895T | 0 | | 0 | 0/6 | 0 | | 0 | 2/436 | 0 | 0 | 2 | 996 | 20.08 | 14.18 | 40 | |
| | | 35 | 2831A>C | N900T | 0 | | 0 | 0/6 | 0 | | 0 | 1/436 | 0 | 0 | 1 | 996 | 10.04 | 10.04 | 20 | |
| | | 36 | 2988G>C | M952I | 0 | | 0 | 0/6 | 0 | | 0 | 1/436 | 0 | 0 | 1 | 996 | 10.04 | 10.04 | 20 | |
| | | 37 | 3030G>A | T966 | (2)[k] | | (1)[k] | | | | 0 | 6/436 | 0 | | 6 (25)[k] | 618 (1814)[k] | 137.82 | 27.37 | | 272 |
| | | 38 | 3032T>C | L967S | 0 | | 0 | 0/6 | 0 | | 0 | 1/436 | 0 | 0 | 1 | 996 | 10.04 | 10.04 | 20 | |
| 16 | 80 | | | | 0 | | 0 | 0/498 | 0 | | 0 | 0/450 | 0 | 0 | | 1 502 | | | | |
| 17a | 151 | 39 | 3123G>C | L997F | 0 | 2 | 2 | 1/494 | 0 | 7 | 1 | 4/454 | 0 | 0 | 17 | 2 502 | 67.95 | 16.42 | 135 | |
| | | 40 | 3157G>A | A1009T | 0 | 2 | 0 | 0/494 | 0 | 0 | 0 | 0/454 | 0 | 0 | 2 | 2 502 | 7.99 | 5.65 | 16 | |
| | | 41 | 3212T>C | I1027T | 1 | 0 | 0 | 0/494 | 0 | 0 | 0 | 0/454 | 0 | 0 | 1 | 2 502 | 4.00 | 4.00 | 8 | |
| 17b | 228 | 42 | 3286T>G | F1052V | 1 | 1 | 0 | 1/194 | 0 | 0 | 0 | 0/452 | 0 | 0 | 3 (3) | 2 200 (2 240) | 13.39 | 7.73 | 27 | |
| | | 43 | 3337G>A | G1069R | 0 | 1 | 0 | 0/194 | 0 | 0 | 0 | 0/452 | 0 | 0 | 1 | 2 200 | 4.55 | 4.55 | 9 | |

**Table 1** (Continued)

| | | | cSNS variants found | | The random samples of genes (and the technique utilized) NE Italy (DGGE) | | Central Italy (DGGE) | | Southern France (DGGE) | | Northern France (DHPLC) | | Spain (SSCA) | Czechia (DGGE) | | | | | $H^b \times 10^4$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exon | Exon Length (bp) | Ref. no. | SNS | SAS[c] | 1st 100[d] | 2nd 500 | 1st 100[d] | 2nd[e] | 1st 100[d] | 2nd 500 | 1st 100 | 2nd[e] | 82[d] | 72 | Abs. Freq. | Total sample size | $q \times 10^4$ | $se \times 10^4$ | NS[f] | S[f] |
| | | 44 | 3345G>T | Q1071H | 0 | 0 | 0 | 0/194 | 0 | 1 | 0 | 0/452 | 0 | 0 | 1 | 2 200 | 4.55 | 4.55 | 9 | |
| | | 45 | 3417A>T | T1995 | 1 | 3 | 0 | 0/194 | 1 | 1 | 0 | 0/452 | 0 | 0 | 6 (8) | 2 200 (2 506) | 31.92 | 11.27 | | 64 |
| | | 46 | 3419T>G | L1096R | 0 | 0 | 0 | 0/194 | 1 | 0 | 0 | 0/452 | 0 | 0 | 1 | 2 200 | 4.55 | 4.55 | 9 | |
| | | 47 | 3477C>A | T1115 | 0 | 0 | 0 | 0/194 | 0 | 0 | 0 | 1/452 | 0 | 0 | 1 | 2 200 | 4.55 | 4.55 | | 9 |
| 18 | 101 | 48 | 3523A>G | I1131V | 0 | 0 | 1 | 0/10 | 0 | | 0 | 0/448 | 0 | 0 | 1 (2) | 1 512 (1 908) | 10.48 | 7.07 | 21 | |
| | | 49 | 3586G>C | D1152H | 0 | 0 | 0 | 0/10 | 0 | | 0 | 1/448 | 0 | 0 | 1 | 1 512 | 6.61 | 6.61 | 13 | |
| 19 | 249 | 50 | 3617G>T | R1162L | 0 | 0 | 1 | 1/494 | 0 | 0/260 | 0 | 0/454 | 0 | 0 | 2 | 2 262 | 8.84 | 6.25 | 18 | |
| | | 51 | 3690A>G | Q1186 | 0 | 0 | 0 | 0/494 | 0 | 0/260 | 0 | 0/454 | 1 | 0 | 1 | 2 262 | 4.42 | 4.42 | | 9 |
| | | 52 | 3813A>G | L1227 | 0 | 1 | 0 | 0/494 | 0 | 0/260 | 0 | 0/454 | 0 | 0 | 1 | 2 262 | 4.42 | 4.42 | | 9 |
| | | 53 | 3837T>G | S1235R | 1 | 1 | 0 | 1/494 | 0 | 4/260 | 0 | 7/454 | 0 | 1 | 15 (15) | 2 262 (2 310) | 69.94 | 16.71 | 140 | |
| 20 | 156 | 54 | 4002A>G | P1290 | 2 | | 3 | 0/6 | 3 | | 5 | 18/454 | 3/80 | 2 | 36 | 1 012 | 357.73 | 58.22 | | 690 |
| 21 | 90 | 55 | 4009G>A | V1293I | 0 | | 0 | 0/6 | 0 | 0/300 | 0 | 1/456 | 0 | 0 | 1 | 1 316 | 7.60 | 7.60 | 15 | |
| | | 56 | 4029A>G | T1299 | 1 | | 0 | 0/6 | 0 | 1/300 | 0 | 1/456 | 0 | 0 | 3 (8) | 1 316 (2 330) | 34.33 | 12.12 | | 69 |
| | | 57 | 4041C>G | N1303K | 1 | | 0 | 0/6 | 0 | 0/300 | 0 | 0/456 | 0 | 0 | 1 | 1 316 | 7.60 | 7.60 | 15 | |
| | | 58 | 4085T>C | V1318A | 0 | | 0 | 0/6 | 0 | 0/300 | 0 | 1/456 | 0 | 0 | 1 | 1 316 | 7.60 | 7.60 | 15 | |
| 22 | 173 | | | | 0 | | 0 | 0/18 | 0 | | 0 | 0/450 | 0 | 0 | | 1 022 | | | | |
| 23 | 106 | | | | 0 | 0 | 0 | 0/6 | 0 | | 0 | 0/448 | 0 | | | 1 436 | | | | |
| 24[l] | 198+3 | 59 | 4404C>T | Y1424 | 1 | | 0 | 0/6 | 1 | | 2 | 5/420 | 0 | 2 | 11 (32) | 980 (2 516) | 127.19 | 22.34 | | 251 |
| | | 60[m] | 4521G>A | Q1463 | (21) | | (16) | (3/32) | (14/80) | | (30) | (94/420) | 15/76 | (17) | 15 (227) | 76 (1052) | 2142.86 | 131.07 | | 3 367 |
| | | 61 | 4563T>C | D1477 | 0 | | 0 | 0/6 | 0 | | 1 | 0/420 | 0 | 0 | 1 | 980 | 10.20 | 10.20 | | 20 |
| | Totals | | | | | | | | | | | | | | | | | | 6 525 | 9 584 |
| | | | | | | | | | | | | | | | | | | | 16 109 | |

The bracketed figures include also the RFLP analysis data (see Materials and methods); the NE Italy, Central Italy, Southern and Northern France are each subdivided into two samples where the 1st is made up of 100 genes.

[a]Single nucleotide substitutions in a coding sequence.

[b]Degree of heterozygosity = 2pq.

[c]Single aminoacid substitution; for the synonymous, the aminoacid is indicated.

[d]Data published in the previous survey.[13]

[e]The 2nd phase sample size (no. of genes) of Central Italy and Northern France was not the same for all the exons.

[f]NS = Nonsynonymous; S = synonymous.

[g]The whole exon length is 185 bp, but the coding sequence starts from the 133th bp.

[h]The frequency of the M allele is reported.

[i]Data not recorded and gels no longer available.

[j]Owing to of its large size (724 bp) this exon has been amplified in to two (partially overlapping) fragments.

[k]This variant was not detected in the previous survey[13] because it is undetectable with DGGE. It was specifically searched for following its detection in the Northern France sample.

[l]The whole exon length is 1755 bp (201 bp coding and 1554 bp 3′ UTR).

[m]The genotype for this variant is difficult to identify with DGGE. Therefore, it was specifically examined with a restriction assay,[13] except for the Spain sample.

and $\theta$ is observed, because different selective regimes differently affect these two quantities.[20]

The distribution of gene frequencies expected by the neutral mutation model was calculated according to Glatt *et al*.[10]

## Results

In the present investigation, a gene of ca. 4.5 coding kb was explored on a mean random sample of 1550 European (Italian, French, Spaniard, Czech) genomes (total window $(n_g \times n_{bp}) \approx 7$ Mb). The results are presented in Table 1 and Figure 1. Table 1 reports the detailed results derived from each subsample and Figure 1 shows the length of the analyzed DNA sequences encompassing the variable nucleotide sites and the cSNSs found along with estimates of their frequencies and their position in the gene. A total of 61 cSNSs was found, 45 were NS and 16 S. Only 12 (six NS and six S) showed a polymorphic frequency ($q > 0.005$), corresponding to a mean density of 2.71 SNP per kb (being the total length of the analyzed sequence 4.4 kb), a figure that compares well with the mean density (1.77 per kb) published in the dbSNP website (http://www.ncbi.nlm.nih.gov/SNP).

Two parameters describing the nucleotide diversity ($\theta$, proportion of segregating sites; and $\pi$, mean heterozygosity) and the Tajima's $D$ were calculated for the whole sample (1550 genomes) and for two subsamples of 100 and 400 genomes, respectively [(Table 2, section (a)]. Furthermore $\theta$, $\pi$ and the Tajima's $D$ were calculated separately for the NS and the S cSNSs in the whole sample ($n_g = 1550$) [Table 2, section (b)].

The data concerning the four subsamples of 100 genes are comparable with each other and with those of the literature obtained on samples of similar size and relating to very numerous genes: present $\theta$ and $\pi$ are similar to those of the other authors[4,5] both for their values and for being in agreement with the Tajima's model (Tajima's $D$ not significantly different from zero). This is the first investigation in which a gene was studied on a very large ethnically homogenous sample so that a comparison of the present parameters $\theta$ and $\pi$, obtained on the whole sample (1550 genomes), is possible only with those of the few studies in which a large, even though ethnically heterogenous, sample has been analysed.[10–12] Both in the present and in the previous studies the values of $\pi$ obtained on large samples are not significantly different from those derived from small samples; on the contrary, $\theta$ turned out to be higher both than $\pi$ and than the $\theta$ observed with small samples.

## Discussion

### Effect of sample size on the apparent level of polymorphism

To evaluate the rate of misclassification of the variants detected as singletons with $n_g = 100$ (estimated $q = 0.01$), we examined 400 random genomes (200 individuals) subdivided into four subsamples of 100 genes each (marked by 1st in Table 1) and compared the data obtained with these four subsamples with those observed in the whole study. In particular, we recorded the number of what we refer to as 'false negatives' (cSNSs that failed to be detected in a subsample $n_g = 100$, even though their 'true' frequency (estimated in the whole sample of 1550 genomes, on the average) was certainly higher than 0.01) and 'false positives' (cSNSs that exhibited, in one or more subsamples of 100 genes, a $q = 0.01$ although their 'true' frequency was less than 0.005). There were six false negatives (0, 2, 1 and 3 in the four subsamples) and 14 false positives (7, 2, 3 and 2).

These data show that the number of cSNSs found with $n_g = 100$ is very likely to be an *over*estimate of the true number of cSNPs, since the false positives were more than the false negatives. Furthermore, the final $q$'s of the 34 singletons were largely dispersed (from 0.0004 to 0.0286, Figure 2) and many were lower than 0.01 with a final mean value of only 0.0052 instead of 0.01. In addition, the singletons found in each of the four subsamples were almost never the same although they were all derived from the same population (Europeans).

### Evidences suggesting a role of selection

The value of $\theta$ increased with the increase of sample size ($6.3 \times 10^{-4}$ with $n_g = 100$; $10.6 \times 10^{-4}$ with $n_g = 400$ and $17.3 \times 10^{-4}$ with $n_g = 1550$; see Table 2. section (a)), leading to a negative $D$ value statistically significant only with large $n_g$. Negative Tajima's $D$ are consistent with population expansion and/or negative selection. The observation that the present increase of $\theta$ is almost exclusively due to the NS cSNSs (see Table 2. section (b)) is against a pure neutrality model: demographic events, in fact, would have affected to the same extent both types of substitutions (NS and S). A selection process is also suggested by the pattern of NS and S substitutions (see later) and by the distribution of the variant frequencies: if one compares the observed distribution with that predicted by the neutral mutation model, there appears to occur a striking excess of rare variants, particularly of the NS: in the class of variants with $q < 0.005$, the expected number among the 45 NS variants was 14.8, while the observed was 39 ($P \approx 0$); for the 16 synonymous variants the number expected was 5.2 and the observed was 10 ($P \approx 0.04$). The slightly significant $P$ for the synonymous substitutions may be due to recent population expansion.

High values of $\theta$ and significantly negative Tajima's $D$ have been found also in the few investigations in which much larger samples than usual were studied.[10–12] In particular, Glatt *et al*.[10] compared the $\pi$, $\theta$ and $D$ obtained on a subsample of 180 genomes with those obtained on a sample of 900 genomes and found a similar increase of $\theta$

**Figure 1** The windows ($n_g \times n_{bp}$) utilized to study the 27 *CFTR* exons and the position of the 61 discovered cSNSs with the frequency of the minor allele. Upper graph: The figure within each box (■ and ◆) is the reference number of the cSNS as in Table 1. The *q* confidence limits are equal to 2.5 se. Lower graph: Each exon (or PCR-amplified segment (the exon 13 has been amplified into two segments)) is represented as a window with the horizontal and the vertical side indicating, respectively, the length of the fragment studied ($n_{bp}$) and the sample size ($n_g$) (some cSNSs have been examined on an additional sample (not shown in the graph) by a specific method, see Materials and methods). The two figures inside each window indicate the exon number and its distance, in kb, from the cap site (GDB accession nos AC000061-AC000111).

**Table 2** Dependence of the pattern of the CFTR cSNS variability on the size, $n_g$, of the random sample and on the type (NS or S) of cSNS

*(a) With respect to $n_g$ (pooled NS and S data)*

| $n_g$ | $\pi \pm se \times 10^4$ | $\theta \pm se \times 10^4$ | $D^{Tajima's}$ | $P$ |
|---|---|---|---|---|
| $\sim 100$[a] | $3.4 \pm 2.3$ | $6.3 \pm 2.3$ | $-1.257$ | $>0.1$ |
| $\sim 400$[b] | $3.4 \pm 2.3$ | $10.6 \pm 2.9$ | $-1.811$ | $\sim 0.04$ |
| $\sim 1550$ | $3.6 \pm 2.4$ | $17.3 \pm 3.6$ | $-2.082$ | $<0.01$ |

*(b) With respect to the type of cSNS (with $n_g = 1550$)*

| *Type of cSNS* | $\pi \pm se \times 10^4$ | $\theta \pm se\ 10^4$ | $D^{Tajima's}$ | $P$ |
|---|---|---|---|---|
| NS | $1.5 \pm 1.3$ | $12.8 \pm 2.8$ | $-2.262$ | $<0.001$ |
| S | $2.2 \pm 1.7$ | $4.5 \pm 1.3$ | $-1.134$ | $\gg 0.1$ |

[a]The data of this line are means from four subsamples of 100 genes each:
NE Italy ncSNS = 22; $\pi = 3.8 \times 10\text{-}4$; $\theta = 9.5 \times 10^{-4}$;
Central ItalyncSNS = 12; $\pi = 2.9 \times 10\text{-}4$; $\theta = 5.2 \times 10^{-4}$;
Southern France ncSNS = 13; $\pi = 3.3 \times 10\text{-}4$; $\theta = 5.6 \times 10^{-4}$;Northern France ncSNS = 10; $\pi = 3.5 \times 10\text{-}4$; $\theta = 4.3 \times 10^{-4}$.
[b]Pooled data.
The section (b) shows that the statistically significant increase of $\theta$ associated with the sample size increase (see section (a)) is to be ascribed only to the NS substitutions.

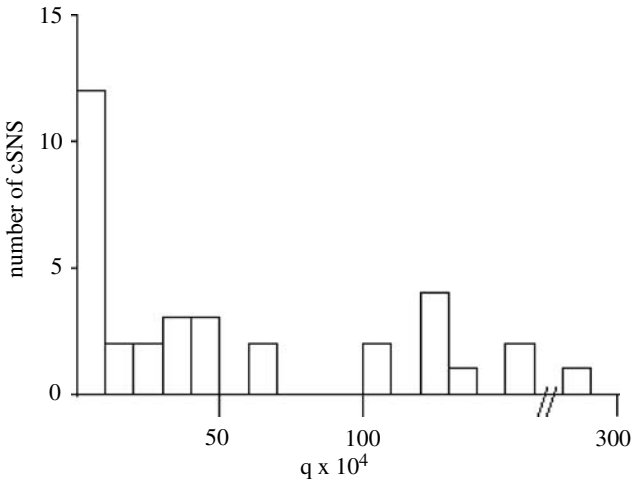
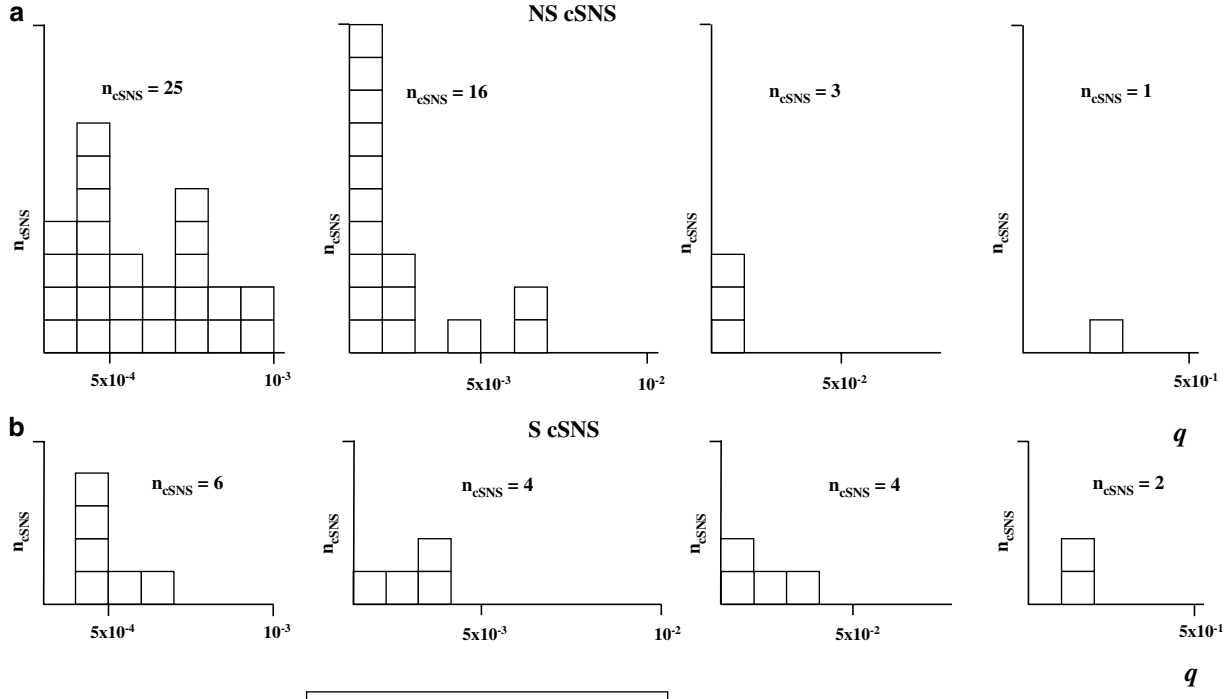



**Figure 2** The distribution of the final *q*'s (those observed on the sample of 1550 genomes) of the 34 'singletons' found on the four subsamples of 100 genomes (see text).




| Summary of the data on the 61 cSNSs | | |
|---|---|---|
| | NS | S |
| POLY<br>q > 0.005 | 6 | 6 |
| SUB-POLY<br>q < 0.005 | 39 | 10 |

$P < 0.005$

**Figure 3** The *q* distribution of the 45 NS and 16 S cSNSs.

almost exclusively due to the NS substitutions and a significantly negative *D* value only on the larger sample.

### Patterns of NS and S substitutions

A reasonably reliable knowledge of the frequency of an adequate number of rare alleles enabled us to compare for the first time the NS with the S pattern of variation also in the subpolymorphic frequency range (Figure 3).

The table inserted in Figure 3 shows that the observed relative fractions of the NS and S cSNSs depend dramatically on the range of variation: it turned out to be 6:6 among the 12 polymorphic substitutions ($q \geqslant 0.005$) and 39:10 among the 49 subpolymorphic ($q < 0.005$) substitutions ($P < 0.005$).

In other words, the large yield of cSNSs detected by expanding the sample size mainly consists of NS cSNSs. If one takes into account the number of NS and that of S cSNSs that may occur in the *CFTR* gene (10 408 and 2921, respectively (see: Materials and methods)) it turns out that

(1) The probability of the NS substitutions of being polymorphic, $P^{NS}_{poly}$, is *much lower* than that of the S substitutions ($6/10\,408 = 5.8 \times 10^{-4}$ *vs* $6/2921 = 20.5 \times 10^{-4}$) ($P < 0.03$), whereas
(2) The estimates of the probabilities of being subpolymorphic, $P_{SubPoly}$, are equal ($P^{NS}_{SubPoly} = 39/10\,408 = 37.5 \times 10^{-4}$; and $P^{S}_{SubPoly} = 10/2921 = 34.2 \times 10^{-4}$; $P > 0.75$); this implies that
(3) For a NS substitution the probability of being subpolymorphic is much higher than its probability of being polymorphic (39/10 408 *vs* 6/10 408), while this is not the case for the synonymous cSNSs (6/2921 *vs* 10/2921).

Obviously, the mean heterozygosity of the NS substitutions, $H^{NS}_{mean}$, turned out to be much lower than that of the S substitutions, $H^{S}_{mean}$ ($H^{NS}_{mean} = 0.65/45 = 0.014$; $H^{S}_{mean} = 0.96/16 = 0.060$) (last two columns of Table 1).

Similar findings that $P^{NS}_{Poly} \ll P^{S}_{Poly}$ were already reported for very many other loci[4,5] and the commonly accepted explanation for this is that the majority of the NS cSNSs are deleterious whereas the majority of the S cSNSs are neutral. This explanation is also supported by previous evidence that the relative shortage of NS cSNPs concerns mainly the drastic aa substitutions[21] and by the data in Figure 4, showing that the proportion of cSNPs found among the cSNSs detected in the present survey were 6/16, 5/30 and 0/14 in the classes of cSNSs with a 'score of drasticity' $dr = 0$ (the SS cSNSs) or with $0 < dr < 8$ or with $8 < dr < 24$, respectively.

However, another result of this study, which concerns the low range of variation, seems to deserve serious consideration. Although the majority of the NS substitutions are deleterious, their probability of being subpolymorphic turned out to be approximately equal to that of the S substitutions. This finding directly implies that the majority of the rare NS substitutions behave as if they were neutral despite being intrinsically deleterious. Besides having obvious relevance for Medical Genetics, this observation is evolutionarily important since it makes less vague and arbitrary the distinction between 'common' (ie polymorphic) and 'rare' (ie subpolymorphic) range of variation. In the common variation range, evolution proceeds deterministically, namely, selection overwhelms (or, more likely, overwhelmed) genetic drift; thus all the polymorphic alleles are either neutral or advantageous, at least in the heterozygous state. On the contrary, in the low range of variation evolution proceeds stochastically, namely, selection ceases (or ceased) to be effective so that rare alleles behave(d) as neutral, irrespective of being intrinsically neutral or deleterious. Further data are necessary to clarify if the threshold (more realistically a threshold range of *q* values) varies (for the position and/or steepness), as it would seem likely, between different genes and populations.

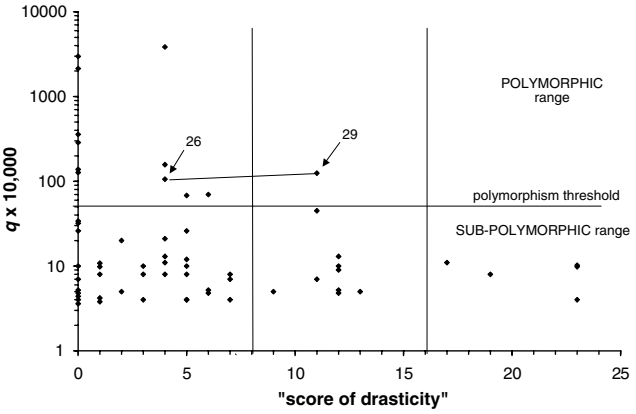



**Figure 4** Negative correlation between the 'score of drasticity', dr, of the cSNSs and the frequency *q* of their minor allele. The herein adopted term 'score of drasticity', dr, assigned to the cSNSs is a figure equal to the 'score' shown in the Matrix: (pam 10 (point accepted mutation)[22]) except for the synonymous cSNSs to which, as a first approximation, we assigned a $dr = 0$. The cSNSs 26 and 29 appear to be always associated.

The combined action of *recessivity* and the ensuing *genetic drift* appears to be the only reasonable explanation for the ineffectiveness of selection against the NS cSNSs in the subpolymorphic range of variation ($P^{NS}_{SubPoly} \approx P^{S}_{SubPoly}$).

*Recessivity* is implied by the strong frequency-dependence of the ratio $n_{NS}/n_{S}$ (see Table in Figure 3), which virtually excludes an appreciable share of dominant-negative alleles (provided any such alleles actually exist for the *CFTR* gene) among the deleterious *CFTR* alleles. This notion is also supported by the well-known formal genetics of CF and by the monomeric structure of the CFTR protein. In other words, for this autosomal gene, a

malfunction of a variant allele from a parent can have appreciable adverse phenotypic effects only in the few individuals where also the expression of the gene derived from the other parent is severely impaired.

The role of *genetic drift* in shaping the pattern of variation of the NS with respect to that of the S cSNSs is a direct consequence of the recessivity of the disadvantage associated with the deleterious alleles. As a matter of fact, for each deleterious allele, the effectiveness of the genetic drift in reducing the consequences of its selective disadvantage should be proportional, at least as a first approximation, to the ratio between $Se_q^{rel}(=se_q/q$, the relative extent of the random fluctuations of its frequency $q$) and $\Delta q_{rel}^{sel}(=\Delta q/q$, the expected relative decrease of its frequency per generation caused by selection), and recessivity reduces drastically $\Delta q_{rel}^{sel}$. In fact, for a deleterious *dominant* allele $\Delta q_{rel}^{sel}$ depends only on its severity (for a lethal dominant allele is 1 by definition, regardless of its frequency and that of the other deleterious alleles); on the contrary, for a deleterious *recessive* allele it also depends on the cumulative frequency of the other deleterious alleles (lethal or sublethal). If the ratio $se_q^{sel}/\Delta q_{rel}^{sel}$ is very high (because its numerator is high and/or because its denominator is very low) selection is quite ineffective. The numerator is a function of the demographic history of the population; the denominator is proportional to the cumulative frequency of the deleterious alleles (usually in the order of 0.01 or even less); thus its value is very close to 0. The finding that recessive deleterious or even lethal alleles are very numerous (owing to their behavior as neutral alleles, at least in the not extremely large populations) supports the old and commonly accepted idea of the existence, for many genes, of a reservoir of potentially useful rare variability.

## Acknowledgements

*This work was funded by the Italian Ministry of Health, CF Project, law 548/93; the Italian Ministry of University and Research; the Italian CNR Strategic Project Biotechnology; the CHU of Montpellier and the French Association against Cystic Fibrosis (VLM Vaincre la Mucoviscidose). Partial fulfillment of CB has been 'Assegno di Ricerca' by the University of Verona; FB was supported by the Italian CNR Strategic Project Biotechnology MM was supported by MZCR 00000064203, MSMT 111300003, LNOA079, EUCF Chip, CRMGEN.*

*We thank the European CF Network for providing mutation controls and the PCR mix for multiplex DGGE of the CFTR gene; and Drs Vassanelli and Gandini from the Verona blood Center for blood collection.*

*We are grateful to Guido Barbujani and Andrea Novelletto for helpful comments and to Maria Lo Ponte for her careful linguistic revision of the text.*

## References

1 Pagani F, Buratti E, Stuani C, Baralle FE: Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. *J Biol Chem* 2003; **278**: 26580–26588.
2 Pagani F, Stuani C, Tzetis M *et al*: New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. *Hum Mol Genet* 2003; **12**: 1111–1120.
3 Nickerson DA, Taylor SL, Weiss KM *et al*: DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet* 1998; **19**: 233–240.
4 Cargill M, Altshuler D, Ireland J *et al*: Characterisation of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet* 1999; **22**: 231–238.
5 Halushka MK, Fan JB, Bentley K *et al*: Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet* 1999; **22**: 239–247.
6 Nickerson DA, Taylor SL, Fullerton SM *et al*: Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res* 2000; **10**: 1532–1545.
7 The human genome. *Nature* 2001; **409**: 745–964.
8 The International SNP Map Working Group: A map of human genome sequence variation containing 1.42 milion single nucleotide polymorphisms. *Nature* 2001; **402**: 928–933.
9 Hirakawa M, Tanaka T, Hashimoto Y *et al*: JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res* 2002; **30**: 158–162.
10 Glatt EC, DeYoung JA, Delgado S *et al*: Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nature Genet* 2001; **27**: 435–438.
11 Leabman MK, Huang CC, DeYoung J *et al*: Natural variation in human membrane transporter genes reveals evolutionary and functional constrains. *Proc Natl Acad Sci* 2003; **100**: 5896–5901.
12 Smirnova I, Hamblin MT, McBride C *et al*: Excess of rare aminoacid polymorphisms in the Toll-like receptor 4 in humans. *Genetics* 2001; **158**: 1227–1234.
13 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.
14 Ghanem N, Costes B, Girodon E, Martin J, Fanen P, Goossens M: Identification of eight mutations and three sequence variations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. *Genomics* 1994; **21**: 434–436.
15 Fanen P, Ghanem N, Vidaud M *et al*: 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* 1992; **13**: 770–776.
16 Chillon M, Palacio A, Nunes V, Estivill X: A rare DNA variant in exon 15 of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. *Hum Genet* 1992; **90**: 474.
17 Watterson G: On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 1975; **7**: 256–276.
18 Tajima F: DNA polymorphisms in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* 1989; **123**: 229–240.
19 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.
20 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.
21 Stephens JC, Schneider JA, Tanguay DA *et al*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293**: 489–493.
22 Dayhof MO, Schwartz RM, Orcutt BC: A model of evolutionary change in proteins; In: Dayhoff MO (ed) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, 1978, vol. 5 (suppl 2), pp 345–352.