Worldwide Genetic Analysis of the CFTR Region

Eva Mateu,[1] Francesc Calafell,[1] Oscar Lao,[1] Batsheva Bonné-Tamir,[2] Judith R. Kidd,[3] Andrew Pakstis,[3] Kenneth K. Kidd,[3] and Jaume Bertranpetit[1]

[1]Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona; [2]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv; and [3]Department of Genetics, Yale University School of Medicine, New Haven

Mutations at the cystic fibrosis transmembrane conductance regulator gene (CFTR) cause cystic fibrosis, the most prevalent severe genetic disorder in individuals of European descent. We have analyzed normal allele and haplotype variation at four short tandem repeat polymorphisms (STRPs) and two single-nucleotide polymorphisms (SNPs) in CFTR in 18 worldwide population samples, comprising a total of 1,944 chromosomes. The rooted phylogeny of the SNP haplotypes was established by typing ape samples. STRP variation within SNP haplotype backgrounds was highest in most ancestral haplotypes—although, when STRP allele sizes were taken into account, differences among haplotypes became smaller. Haplotype background determines STRP diversity to a greater extent than populations do, which indicates that haplotype backgrounds are older than populations. Heterogeneity among STRPs can be understood as the outcome of differences in mutation rate and pattern. STRP sites had higher heterozygosities in Africans, although, when whole haplotypes were considered, no significant differences remained. Linkage disequilibrium (LD) shows a complex pattern not easily related to physical distance. The analysis of the fraction of possible different haplotypes not found may circumvent some of the methodological difficulties of LD measure. LD analysis showed a positive correlation with locus polymorphism, which could partly explain the unusual pattern of similar LD between Africans and non-Africans. The low values found in non-Africans may imply that the size of the modern human population that emerged "Out of Africa" may be larger than what previous LD studies suggested.

Introduction

The cystic fibrosis transmembrane conductance regulator gene (CFTR [MIM 602421]), also known as ABCC7 (member number 7 of subfamily C of the ATP-binding cassette [ABC] transporter gene family), was identified and cloned in 1989 (Kerem et al. 1989; Riordan et al. 1989; Rommens et al. 1989). Since then, >900 mutations in CFTR that cause cystic fibrosis (CF [MIM 219700]) have been reported (Cystic Fibrosis Mutation Data Base). Cystic fibrosis is the most common severe autosomal recessive disease in patients of European descent, affecting 1/2,500 newborns, which implies a gene frequency for the disease of $q = .02$ and a carrier frequency of 1/25. The CFTR gene comprises 27 exons, spanning 230 kb on the long arm of chromosome 7 (7q31.2), that encode a 1,480–amino acid protein with chloride-channel activity regulated by cyclic AMP. The most frequent CF mutation is a deletion of 3 bp at codon 508 (ΔF508 mutation), and it accounts for almost 67% of the global CF chromosomes. Only four other mutations (G542X, N1303K, G551D, and W1282X) have overall allele frequencies among CF chromosomes >1% (Estivill et al. 1997). Most of the remaining mutations are rare or are confined to specific populations.

Several short tandem repeat polymorphisms (STRPs, also known as microsatellites) and single-nucleotide polymorphisms (SNPs) have been described within the CFTR gene. Both types of markers can be used to trace the origin and evolution of the different CF mutations (Morral et al. 1994; Bertranpetit and Calafell 1996; Slatkin and Rannala 1997). SNPs can be used to define the haplotypic frameworks on which CFTR mutations occurred. Faster-mutating STRPs can be used to estimate ages of mutations from the variability accumulated in CF-mutated chromosomes.

The combination of several polymorphisms and the determination of haplotypes allows the estimation of linkage disequilibrium (LD)—that is, the departure from the haplotypic frequencies expected under independent inheritance of the different markers. The study of the distribution of LD patterns in different populations can yield valuable information on population history (Tishkoff et al. 1996, 1998; Kidd et al. 1998, 2000). The power of LD as a tool for gene mapping in relation to population demography can be explored as

Received September 22, 2000; accepted for publication November 1, 2000; electronically published December 4, 2000.

Address for correspondence and reprints: Dr. Jaume Bertranpetit, Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain. E-mail: jaume.bertranpetit@cexs.upf.es

0002-9297/2001/6801-0010$02.00

well, since it is currently under debate whether small and isolated populations are more suitable for LD mapping (Eaves et al. 2000; Jorde et al. 2000) and how far, in physical distance, LD extends (Ott 2000). Moreover, the effect on LD of the mutation rate and pattern is analyzed (i.e., slowly mutating SNPs vs. fast, stepwise-mutating STRPs).

The aims of this paper are to analyze the genetic variation in CFTR polymorphisms; to estimate allele and haplotype frequencies and describe their geographic distribution; and to measure LD within the CFTR gene, to describe its genomic patterns in relation to physical distances and marker variability as well as its population patterns, which then can be used to infer population history. In 1,944 chromosomes from healthy individuals from 18 worldwide populations, we have analyzed six polymorphisms, four STRPs, and two SNPs located within the CFTR gene.

Material and Methods

Polymorphic Sites

The polymorphisms studied are located within CFTR, as shown in figure 1. We have typed four STRPs—one of which is practically diallelic, whereas the other three are highly polymorphic—as well as two SNPs. Listed from the 5′ to the 3′ end of the gene, the polymorphisms typed are as follows: IVS1CA is a CA dinucleotide with high allelic variability, located in the first intron of the gene (Moulin et al. 1997, Mateu et al. 1999). IVS6aGATT is a mostly dimorphic 4-bp tandem repeat located in intron 6a (Chehab et al. 1991; Gasparini et al. 1991). IVS8CA is also a CA dinucleotide with high allelic variability, located in intron 8 of the gene (Morral et al. 1991). T854/*Ava*II is a silent T→G nucleotide substitution located in exon 14a (Zielenski et al. 1991*b*). IVS17bTA is a highly polymorphic TA dinucleotide, located in intron 17b (Zielenski et al. 1991*a*). Finally, TUB20/*Pvu*II is a G→A nucleotide substitution located in intron 20 (Quere et al. 1991).

Population Samples

We have studied 972 random, unrelated autochthonous individuals from 18 populations, representing all major world geographic areas. Sub-Saharan African populations comprised Mbuti Pygmies (from the Ituri Forest, in the former northeast Zaire), Biaka Pygmies (from the village of Bagandu, in the southwest corner of the former Central African Republic), and Tanzanians (from Ifakara, Kilombero district, Morogoro region, in southeastern Tanzania). North Africans were represented by the Saharawi (from the former Western Sahara). Samples from the Middle East were the Druze (a Moslem community from Galilee in northern Israel) and Yemenite Jews (Yemenite immigrants to Israel); the European populations comprised Basques (unrelated individuals of rural origin living in the Gipuzkoa province of the Basque country in Spain), Catalans (from rural villages of north Girona in Catalonia, Spain), Finns (unrelated individuals from Finland who are not of Swedish origin), Russians (from the Zuevsky district northeast of Moscow), and Adygei (from north of the Caucasus mountains in the Krasnodar region in southeast Russia). Asian samples included Kazakhs (from the village of

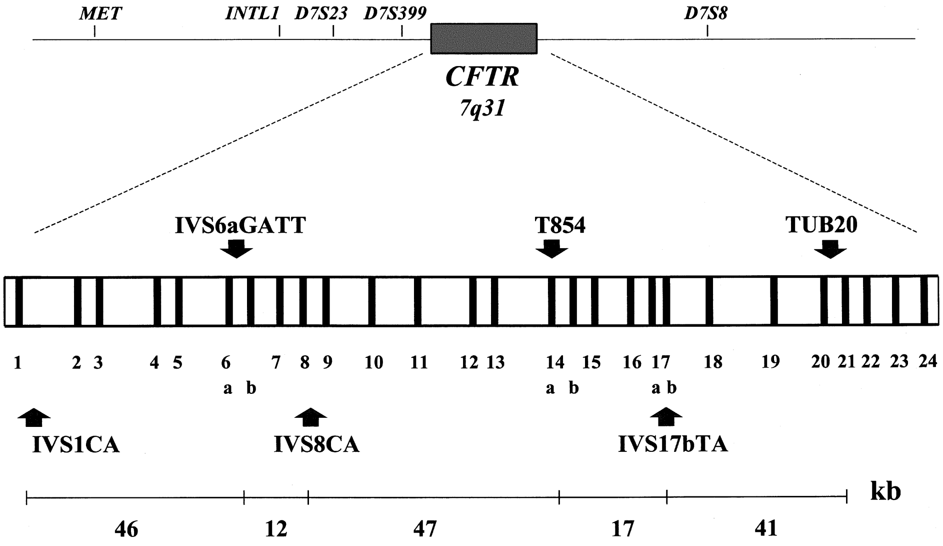

Figure 1 CFTR gene with all six polymorphic genetic markers studied (IVS1CA, IVS6aGATT, IVS8CA, T854, IVS17bTA, TUB20), showing the physical distances (in kb) between them. Gene exons are denoted by numbers (1–24).

Aktasty in the Almaty region in Kazakhstan, Central Asia), Yakut (from the Yakutian Autonomous Republic of Russia, in eastern Siberia), Han Chinese (from southern China, collected in the San Francisco area), and Japanese (also collected in the San Francisco Bay area); the Pacific was represented by the Melanesian Nasioi (from Bougainville in the Solomon Islands, Melanesia). The American Indians sampled were Mayans (from Yucatan, Mexico) and Rondônia Surui (southwest Amazon, Brazil). Sample sizes ranged from 46 (Nasioi) to 222 (Basque) chromosomes, with a median of 108. T854/*Ava*II and TUB20/*Pvu*II were also typed in six primates: two gorillas (*Gorilla gorilla*), one orangutan (*Pongo pygmaeus*) and three common chimpanzees (*Pan troglodytes*).

DNA from five populations (Basques, Catalans, Tanzanian, Kazakhs, and Saharawi) was extracted from fresh blood of blood donors. Appropriate informed consent was obtained from human subjects. DNA samples for the other populations were obtained from lymphoblastoid cell lines maintained in the laboratory of J.R.K. and K.K.K. at Yale University. Fresh primate blood samples were supplied by the Barcelona Zoo.

STRP Analysis

Typing methods for microsatellite IVS1CA are as described elsewhere (Mateu et al. 1999), where we reported allele frequencies for most of the current population set. The GATT tetranucleotide in intron 6a (IVS6aGATT) (Chehab et al. 1991) was analyzed by PCR amplification and electrophoresis of the products in a 8% acrylamide gel. Microsatellites IVS8CA and IVS17bTA (Morral et al. 1991; Zielenski et al. 1991*a*) were analyzed in a multiplex reaction using the primers described by Morral and Estivill (1992). PCR amplifications were performed using 50 ng of genomic DNA in a final 10-μl volume. The CA repeats were amplified with flanking primers I9D3 and I9R4, and the TA repeats were amplified with flanking primers AT17D1.2 and AT17R1.2. Markers I9D3 and AT17D1.2 were fluorescently labeled. Amplification conditions for 30 cycles were as follows: denaturing at 95° for 30 s, annealing at 50° for 30 s, and extension at 65° for 45 s. PCR products were pooled, were combined with a size standard (ABI GS500 ROX) and a bromophenol blue– and formamide-based loading buffer, and were loaded on a standard 6% denaturing sequencing gel. Electrophoresis was conducted using an ABI 377TM sequencer. GeneScan 672TM was used to collect the data, track lanes, and measure fragment sizes. The number of CA and TA repeats was estimated by sequencing two CA- and four TA-homozygous individuals with different fragment sizes for each loci. The sequencing reaction was performed with flanking primers I9R4 and AT17R1.2 and the DNA Sequencing KitTM (PE Biosystems) according to manufacturer's specifications.

Analysis of SNPs

The T854/*Ava*II (2694 T/G) and TUB20/*Pvu*II (4006-200 G/A) SNPs were analyzed by PCR amplification and digestion with the appropriate restriction enzyme, as described by Dörk et al. (1992).

Statistical Analysis

Allele frequencies were estimated by direct gene counting. Maximum-likelihood estimates of haplotype frequencies and their standard errors (jackknife method) were calculated from the multisite marker typing data, using the HAPLO program (Hawley and Kidd 1995), which implements the EM algorithm (Dempster et al. 1977; Slatkin and Excoffier 1996). Tishkoff et al. (2000) confirmed, by direct haplotype typing, that the frequencies estimated with the EM algorithm were quite precise for the common haplotypes.

Expected heterozygosities for loci and for the haplotypes were estimated as $1\text{-}\Sigma p_i^2$, where p_i stands for allele or haplotype frequencies. Analysis of molecular variance (AMOVA) (Excoffier et al. 1992) was performed with the Arlequin package (Schneider et al. 2000).

In order to quantify the portion of the possible haplotype space that was not recovered in the population samples, we computed the $\underline{f}$raction of $\underline{e}$xtra haplotypes (FE) statistic suggested by Slatkin (2000), with some modifications. As defined by Slatkin (2000),

$$FE = \frac{(k_H - k_{min})}{(k_{max} - k_{min})} ,$$

where k_H is the number of haplotypes found in the sample, k_{min} is the minimum possible number of haplotypes (i.e., the number of alleles at the locus with the maximum number of different alleles), and k_{max} is the maximum possible number of different haplotypes—that is, the product of the number of different alleles at each site. However, in our case, k_{max} greatly exceeds sample size for each population, and sample size becomes a limiting factor in the number of different haplotypes that can be actually found. Therefore, we have used as k_{max} the expected number of different haplotypes under linkage equilibrium, given the sample size and allele frequencies (k_e). This value was obtained by sampling—at random and independently—one allele at each locus, with probabilities equal to their population frequencies. This way, a number of random haplotypes equal to the original sample sizes was reconstructed, and the number of different haplotypes was counted. This procedure was repeated 100,000 times, and the average number of different haplotypes at each iteration was used to estimate

Table 1

Expected Heterozygosity, by Locus and Haplotype

Population	IVS1CA	IVS6aGATT	IVS8CA	T854	IVS17bTA	TUB20	Haplotype
Sub-Saharan Africa:							
Biaka	.90	.48	.83	.45	.81	.40	.966
Mbuti	.84	.50	.82	.49	.79	.24	.954
Tanzanians	.85	.35	.60	.46	.91	.18	.968
North Africa:							
Saharawi	.82	.36	.35	.50	.79	.44	.961
Middle East:							
Yemenites	.70	.32	.48	.33	.84	.24	.958
Druze	.64	.33	.40	.35	.82	.30	.902
Europe:							
Adygei	.70	.28	.41	.42	.83	.34	.951
Russians	.77	.32	.50	.50	.70	.42	.948
Finns	.76	.40	.54	.41	.87	.32	.957
Catalans	.75	.37	.42	.44	.79	.38	.953
Basques	.73	.37	.45	.41	.88	.29	.967
Asia:							
Kazakhs	.73	.47	.59	.49	.86	.14	.962
Chinese	.67	.50	.53	.50	.78	.05	.926
Japanese	.68	.41	.48	.40	.78	.00	.934
Yakut	.76	.39	.58	.34	.72	.05	.936
Pacific:							
Nasioi	.79	.36	.75	.50	.78	.04	.901
America:							
Maya	.59	.46	.53	.46	.84	.06	.933
Surui	.43	.32	.28	.26	.78	.00	.854

k_e. Since we are interested in relating *FE* to LD, and since *FE* as formulated by Slatkin (2000) should decline with LD, we have used instead $FNF = 1 - FE$, which can be interpreted as the fraction of haplotypes not found.

Overall disequilibrium and all 15 pairwise disequilibria were evaluated using the program HAPLO/P (Zhao et al. 1997, 1999), which uses a permutation test to evaluate significance of deviation from random assortment of alleles and calculates the ξ coefficient to quantify the deviation from randomness. Zhao et al. (1999) proposed the following estimate for ξ:

$$\hat{\xi} = \sqrt{2\nu}\frac{1}{n}\left(\frac{t-\mu}{\sigma}\right) ,$$

where μ and σ^2 are the mean and variance of the empirical distribution of the likelihood-ratio test statistics from the permuted samples, and t is the likelihood ratio statistic for the observed sample. Asymptotically, the ξ coefficient allows quantitative comparisons of deviation from randomness, in different populations and between different genetic systems. Physical distances (in kb) between the six loci were based on the CFTR gene sequence published in GenBank (accession numbers AC000111 and AC000061).

Results

Six polymorphisms (four STRPs and two SNPs) in the CFTR gene were typed in 972 individuals (1,944 chromosomes) from 18 populations. Allele frequencies for each population and marker, as well as for the 770 haplotypes estimated to be present, are available on request and have been deposited in ALFRED, the Allele Frequency Database.

Allele Frequencies and Geographic Distribution

IVS1CA allele frequencies had been reported for a subset of the current data (Mateu et al. 1999), and here we present an increased data set for some populations (i.e., Biaka and Mbuti Pygmies, Druze, Yemenites, Kazakhs, Basques, and Catalans). The overall allele frequency distribution is unimodal, with a sharp mode at 22 repeats and a smooth, left-skewed decline towards the ends of the distribution. The most extreme alleles found were the 12 and 28 repeats; allele 13 was found in a single Biaka individual and is reported here for the first time. Alleles 22 and 23 have been found in all populations studied. African populations presented a larger number of alleles and higher heterozygosities at this locus (table 1), because of a higher frequency of peripheral alleles, which seems to be a common feature in many

STRP allele frequency distributions in Africans (Calafell et al. 1998).

The IVS6aGATT tetranucleotide has only two common alleles with six and seven repeats, as described elsewhere (Chehab et al. 1991; Gasparini et al. 1991), and both are present in all populations studied. Allele 7 is the most frequent on average and in most populations; its frequencies range from .20 in the Surui and .24 in the Nasioi to >.75 in many European, Middle Eastern, and African populations. Only three chromosomes in the worldwide sample did not bear alleles 6 or 7; alleles 4, 5, and 8 were found in one Adygei and two Basque chromosomes, respectively. Allele 4 is described for the first time.

The IVS8CA dinucleotide STRP has a highly right-skewed distribution, with a mode at 16–17 repeats (which, together, account for ~75% of the global chromosomes) and a range of 14–25 repeats. Allele 16 is found at frequencies .5–.8 in Tanzanians, North Africans, Middle Easterners, and Asians other than the Chinese. In the latter population and among the Mayans, allele 17 is slightly more frequent, and its frequency reaches 0.8 in the Surui. Allele 23 is found at frequencies 0–.12, except among the Nasioi, in whom it is the most frequent allele (.41). Again, the Biaka and the Mbuti present flatter allele distributions, with a larger number of alleles and high heterozygosity; in contrast, among the American Indians, only three different alleles were found.

The T854 SNP (Zielenski et al. 1991*b*) was found to be polymorphic in all populations studied. Allele 1 (i.e., absence of the restriction site for the *Ava*II enzyme) was found at frequencies ranging from .5 to .79, in Europeans and Asians, and from to .34 to .49, in Africans, Nasioi and Mayans. The lowest frequency was found at .16, in the Surui.

The IVS17bTA dinucleotide STRP is extremely polymorphic, with expected heterozygosities that range from .72 to .91 (table 1). The alleles found range from 7 to 53 repeats and are distributed in four discontinuous groups: allele 7 (and 8 in one chromosome), alleles 15–25, 27–38 and 39–53. As discussed below, the mutation pattern at IVS17bTA may be responsible for this multimodal distribution. Overall, the 27–38 group is the most frequent (global average frequency .63), and, within this group, alleles 30–32 are the most frequent. This group of alleles is most frequent in Asians and American Indians (.74–.96), though it is quite common in Europeans and Africans too (.24–.65). Allele 7 is found in almost all populations, and—except for allele 8 in one chromosome—the next allele in size is allele 15. Allele 7 is found at low frequencies in East Asians and American Indians (it is absent in the Japanese and in the Surui, and its frequency reaches .11 in the Kazakhs), and it is more frequent in Europeans, southwest Asians, and Africans (.14–.50). Within the allele group 15–25, alleles 19–22 are the most frequent; the overall frequency of this group is .15, although it has a wide populational variation. The group was not found in the Yemenites. Its frequency reaches .02 in Europeans and Asians; it is somewhat more frequent in the Africans and in the Maya (.15–.22), and it reaches high frequencies in two populations: the Mbuti (.61) and the Nasioi (.67). Finally, the right-hand tail of the allele distribution extends from allele 39 to 53; most of those alleles are rare or absent, except for 45–47. The average frequency of this group of alleles is .03; all alleles in the group are absent in six populations, and the group reaches frequencies of .10 in the Basques and .13 in the Druze.

The TUB20 SNP (Quere et al. 1991) was detected by a restriction enzyme assay, and in all human populations typed, the presence of the *Pvu*II restriction site (i.e., allele 2) is the most frequent allelic state. Its frequencies range from .7 in the Saharawi to fixation in the Surui and Japanese.

Haplotype Frequencies and Geographic Distribution

The total number of possible haplotypes is 146,880, of which 770 were estimated in the analysis to have occurred at least once. The full set of frequencies for these 770 haplotypes is available in the ALFRED database. Results using HAPLO were very close to those using ARLEQUIN (results not given). Frequencies estimated for the 43 six-locus haplotypes having an estimated frequency of ⩾.05 are given in table 2. The T854 and TUB20 markers can be used to define the core haplotypes since they are diallelic, have presumably much lower mutation rates than the other polymorphisms and the ancestral state can be inferred for them. The most common haplotypes defined with these polymorphisms are: 1-2 in Middle Eastern, European, and Asian populations, with the lowest frequency in Chinese (.51) and the highest in Yakut (.80); and 2-2 in American Indians (0.65 in Maya and 0.85 in Surui). Haplotype 1-1 is scarcely represented in the worldwide sample (table 3). Haplotype backgrounds for the major CF mutations are: 1-2, for ΔF508, G542X, and N1303K mutations, and 2-1, for G551D and W1282X mutations (Morral et al. 1996).

Expected haplotype diversities in all populations are shown in table 1. It is remarkable that, although sub-Saharan African populations have high allele diversities at the STRP loci when compared with other populations, haplotype diversities for Africans are not noticeably higher than haplotype diversities in other populations. This could be caused by higher LD in Africans. We have computed the fraction of haplotypes not found (*FNF*) for each population, a quantity that should grow with LD. Number of haplotypes found, their theoretical

Table 2

Frequency of CFTR Haplotypes

		Frequency of Haplotype[a]																				
		15	16	16	17	17	17	18	18	21	21	21	21	21	21	21	22	22	22	22	22	22
		6	6	7	6	6	6	6	7	6	7	7	7	7	7	7	6	6	6	7	7	7
	Frequency	24	23	17	19	20	22	23	16	23	14	16	16	17	17	17	17	17	22	16	16	16
	of	1	2	2	2	2	2	2	2	1	2	1	2	1	2	2	2	2	2	1	1	1
	Residual	19	23	27	22	19	7	22	7	20	38	31	7	20	7	7	31	37	19	7	7	29
Population (2*N*)	Class	2	2	2	2	2	1	2	2	2	2	2	1	2	1	2	2	2	2	1	2	2
Biaka (122)	.675	0	0	.033	0	0	.060	0	.008	0	.082	0	.008	0	0	0	0	0	0	.096	.011	0
Mbuti (66)	.647	.106	0	0	0	.091	0	0	0	0	0	0	0	.076	0	0	0	0	.061	0	0	0
Tanzanians (64)	.709	0	0	.063	0	0	0	0	.078	0	0	0	0	0	0	0	0	0	0	0	0	0
Saharawi (106)	.576	0	0	0	0	0	0	0	0	0	0	.038	.131	0	0	0	0	0	0	.038	.019	.047
Yemenites (80)	.563	0	0	0	0	0	0	0	0	0	0	.038	.012	0	0	0	0	0	0	0	.038	.100
Druze (126)	.424	0	0	0	0	0	0	0	0	0	0	0	.103	0	0	0	0	0	0	0	.056	.024
Adygei (98)	.410	0	0	0	0	0	0	0	0	0	0	.011	.095	0	0	0	0	0	0	0	.010	.020
Russians (60)	.433	0	0	0	0	0	0	.017	0	0	0	0	.117	0	0	.050	0	0	0	.050	0	0
Finns (62)	.481	0	0	0	0	0	0	.065	0	0	0	.065	.032	0	.081	0	0	0	0	0	0	0
Catalans (166)	.508	0	0	0	0	0	0	.006	0	0	0	.006	.125	0	0	0	0	0	0	.014	.125	.024
Basques (216)	.500	0	0	0	0	0	0	.014	0	0	0	.026	.083	0	0	0	0	0	0	.022	.016	.036
Kazakhs (60)	.498	0	0	0	0	0	0	0	0	.004	0	.033	.017	0	0	0	.052	0	0	.017	.017	.050
Chinese (86)	.501	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.024	.012	0	0	0	0
Japanese (86)	.332	0	0	0	0	0	0	0	0	0	0	.035	0	0	0	0	0	0	0	0	0	.012
Yakut (78)	.458	0	0	0	0	0	0	0	0	.002	0	.050	0	0	0	0	0	0	0	0	0	0
Nasioi (46)	.332	0	.065	0	.152	0	0	0	0	.217	0	0	0	0	0	0	0	0	0	0	0	0
Maya (92)	.364	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Surui (84)	.119	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.056	.077	0	0	0	0

	Frequency of Haplotype																					
	22	22	22	22	22	22	22	23	23	23	23	23	23	23	23	23	23	23	24	24	24	25
	7	7	7	7	7	7	7	6	6	6	6	6	6	6	6	7	7	7	7	7	7	6
	16	16	16	16	16	16	21	17	17	17	17	17	17	17	17	16	16	16	16	16	17	17
	1	1	1	1	1	2	1	2	2	2	2	2	2	2	2	1	1	2	1	1	1	2
	30	31	32	33	44	7	23	15	20	23	31	32	33	35	37	30	32	7	30	31	31	31
Population (2*N*)	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Biaka (122)	0	0	0	0	0	0	.008	0	0	0	0	.033	0	0	0	0	0	.008	0	0	0	0
Mbuti (66)	0	0	.015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tanzanians (64)	0	.063	0	0	0	0	.063	0	0	0	0	.031	.002	0	0	0	0	0	0	.001	0	0
Saharawi (106)	.085	0	0	0	0	.020	0	0	0	0	0	0	0	0	0	.028	0	0	0	0	0	0
Yemenites (80)	.056	.107	0	.013	0	.063	0	0	0	0	0	.003	0	0	0	.013	0	0	0	0	0	0
Druze (126)	.269	.040	.008	.008	.056	0	0	0	0	0	0	0	0	0	0	0	.008	0	.008	0	0	0
Adygei (98)	.071	.102	.104	.050	0	.038	0	0	0	0	0	0	0	0	0	0	.061	.020	0	0	0	0
Russians (60)	.100	.067	0	0	0	0	0	0	0	0	0	0	0	0	0	.002	.033	.100	0	.033	0	0
Finns (62)	.055	.048	.048	.074	0	.032	0	0	0	0	0	0	0	0	0	0	.016	0	0	0	0	0
Catalans (166)	.084	.052	.018	0	0	0	0	0	0	0	.006	.006	0	0	0	.019	0	0	.014	0	0	.003
Basques (216)	.083	.013	.043	.051	0	.007	0	0	0	0	.005	.005	.009	0	0	.009	0	0	.082	0	0	.005
Kazakhs (60)	.017	.115	.033	0	0	0	0	0	0	0	.017	0	.065	0	0	.033	0	0	0	0	0	.033
Chinese (86)	.012	.230	.048	.023	0	0	0	0	0	0	.048	.065	0	0	.012	0	0	0	0	0	.035	0
Japanese (86)	.126	.095	.139	.070	.023	0	0	0	0	0	.044	0	0	.012	0	0	0	0	0	.058	.012	.047
Yakut (78)	.038	.155	.101	0	0	0	0	0	0	0	.013	0	.026	0	0	0	0	0	0	0	.103	.064
Nasioi (46)	0	.065	0	.022	0	0	0	0	.065	0	0	0	0	0	0	.087	0	0	0	0	0	0
Maya (92)	.011	.092	.168	0	0	0	0	.065	0	.065	.048	.104	0	.033	0	.022	0	0	0	.033	0	0
Surui (84)	0	0	.155	0	0	0	0	0	0	0	.134	.036	0	.238	.185	0	0	0	0	0	0	0

[a] Haplotype designations show, from top to bottom, alleles at the loci IVS1CA, IVS6aGATT, IVS8CA, T854, IVS17bTA, and TUB20. All haplotypes that have low frequency (<.05) across all samples are combined into a residual class. 2*N* = number of chromosomes.

Table 3

T854-TUB20 Haplotype Frequencies by Population

	Haplotype			
Population	1-1	1-2	2-1	2-2
Africa:				
Biaka	.087	.249	.192	.472
Mbuti	0	.439	.136	.424
Tanzanians	0	.359	.094	.547
North Africa:				
Saharawi	.088	.403	.227	.282
Middle East:				
Yemenites	.044	.744	.094	.119
Druze	.026	.752	.157	.066
Europe:				
Adygei	.022	.682	.192	.104
Russians	0	.533	.300	.167
Finns	0	.710	.177	.113
Catalans	.040	.634	.219	.107
Basques	.038	.670	.138	.154
Asia:				
Kazakhs	.019	.564	.064	.353
Chinese	0	.512	.023	.465
Japanese	0	.733	0	.267
Yakut	0	.795	.026	.179
Pacific:				
Nasioi	.022	.457	0	.522
America:				
Maya	.033	.315	0	.652
Surui	0	.155	0	.845

bounds, and *FNF* can be found in table 4. It can be seen that some European and Asian populations have lower *FNF* values than African populations.

Ancestral States for SNP Markers and Haplotype Phylogeny

Mutation rates for SNPs are estimated at ~10^{-9} (Li et al. 1996). Therefore, most SNPs are likely to represent a single mutational event. The nucleotide state in other hominoids at the homologous site can be used to infer the ancestral state for the SNP (Iyengar et al. 1998).

Neither T584 nor TUB20 biallelic markers are situated within mutation-prone CpG dinucleotides (Cooper and Krawczak 1990). The T854 biallelic marker (Zielenski et al. 1991*b*) has been typed in primate samples in order to infer the ancestral allele. In these samples (two gorillas, one orangutan, and three chimpanzees) we have found only the 1-allele—that is, the absence of the restriction site for the *Ava*II enzyme. For biallelic marker TUB20 (Quere et al. 1991), also typed in the same primate samples, we have found only the 2-allele, indicating that the presence of the *Pvu*II restriction site is ancestral.

Therefore, in the nonhuman primate samples analyzed, the T854-TUB20 haplotype is 1-2, which is likely to be the ancestral haplotype. This is also the most frequent haplotype (.55) in the present sample set. The other three haplotypes (1-1, 2-2, and 2-1, with respective frequencies of .03, .29 and .13) would have been produced through mutation and recombination. The relative ages of those haplotypes can be explored by measuring the amount of STRP haplotype diversity they carry. Thus, if we consider the 942 chromosomes that carry T854-TUB20 haplotype 1-2, they contain 211 different four-STRP haplotypes, with a haplotype diversity of .96. Four-STRP haplotypic diversities within 1-1, 2-2, and 2-1 chromosomes are, respectively, .79, .98, and .79. The high diversity in 2-2 chromosomes suggests that the derived allele 2 at the T854 site is older than the derived allele at TUB20 and that 2-2 could be a very ancient haplotype.

STRP Variability in a Haplotype Frame

STRP allele-size variances by population have been calculated and are shown in table 5. IVS1CA variance by population varies from 0.22 in Surui to 11.24 in Mbuti, with a global variance of 4.38. The variance in repeat size in IVS6aGATT is very low and ranges from 0.16 in Russians to 0.25 in both Pygmy samples and the

Table 4

Fraction of Haplotypes Not Found (FNF) Values for Each Population

Population ($2N$)	$k_{\min}$[a]	k_{H}[b]	$k_{\max}$[c]	k_{e}[d]	FNF
Africa:					
Biaka (122)	16	63	748	118.3	.5403
Mbuti (66)	13	35	391	63.3	.5623
Tanzanians (64)	18	42	323	61.1	.4430
North Africa:					
Saharawi (106)	16	59	387	89.3	.4134
Middle East:					
Yemenites (80)	12	42	283	63.8	.4213
Druze (126)	14	47	416	88.1	.5548
Europe:					
Adygei (98)	12	45	290	73.8	.4662
Russians (60)	11	33	311	51.4	.4558
Finns (62)	15	33	262	56.4	.5656
Catalans (166)	18	76	575	113.0	.3895
Basques (216)	22	89	583	142.8	.4454
Asia:					
Kazakhs (60)	12	39	182	54.0	.3578
Chinese (86)	13	41	281	65.0	.4612
Japanese (86)	11	31	249	61.6	.6049
Yakut (78)	9	33	274	60.1	.5303
Pacific:					
Nasioi (46)	8	20	161	42.6	.6536
America:					
Maya (92)	16	37	227	67.3	.5906
Surui (84)	7	13	97	34.4	.7810

[a] Minimum possible number of haplotypes.

[b] Number of haplotypes found in the sample.

[c] Maximum possible number of different haplotypes.

[d] Expected number of haplotypes under linkage equilibrium, given sample size and allele frequencies.

Table 5

Microsatellite Variance by Population and by T854-TUB20 Haplotype

Population	IVS1CA	IVS6aGATT	IVS8CA	IVS17bTA
Africa:				
Biaka	11.02	.25	7.20	145.45
Mbuti	11.24	.25	9.61	61.07
Tanzanians	8.25	.19	4.69	108.89
North Africa:				
Saharawi	5.26	.19	2.27	132.55
Middle East:				
Yemenites	1.78	.17	4.48	125.52
Druze	1.83	.17	3.26	165.22
Europe:				
Adygei	2.24	.20	2.31	135.91
Russians	2.63	.16	3.26	143.86
Finns	3.13	.21	5.60	126.90
Catalans	2.97	.19	4.42	139.11
Basques	2.13	.21	4.54	136.25
Asia:				
Kazakhs	2.05	.24	2.86	63.48
Chinese	1.37	.25	.41	26.16
Japanese	1.58	.21	.74	24.99
Yakut	2.31	.19	3.25	36.08
Pacific:				
Nasioi	7.01	.19	9.05	30.52
America:				
Maya	.88	.24	.32	47.33
Surui	.22	.18	.16	8.36
Global	4.38	.23	4.14	118.59
T854-TUB20 haplotype:				
1-1 (2.8%)	3.26	.36	5.97	67.90
1-2 (55.4%)	2.37	.13	4.31	64.98
2-1 (12.6%)	6.41	.14	2.27	36.04
2-2 (29.2%)	7.75	.18	4.52	78.50

Chinese, with a global variance of 0.23. The variance by population in allele size at IVS8CA ranges from 0.16 in Surui to 9.61 in Mbuti, with a global variance of 4.14. The variance in repeat length at IVS17bTA is quite high and ranges from 8.36 in Surui to 165.22 in Druze, with a global variance of 118.59. It should be noted that allele-size variances are determined to a much greater extent by locus than by population, probably because of heterogeneity in mutation rate and pattern. Population differences would explain only 2% of the variation in allele-size variances, as determined by ANOVA.

STR variance by T854-TUB20 haplotype is also represented in table 5. As discussed above, haplotype 1-2 is likely to be ancestral; however, 1-2 chromosomes do not carry the highest STR variances, as would be expected if this were the oldest haplotype. Given the weight of extreme-sized alleles in the variance, this parameter may have a larger evolutionary variance when compared to, for instance, haplotype diversity at each background. Moreover, it presents a large degree of heterogeneity among populations.

Alleles at each STR did tend to show preferential associations with T854-TUB20 haplotypes. Thus, .484 of all T854-TUB20 2-1 haplotypes carried allele 21 at IV1CA, while this allele is found at an overall frequency of .172. Conversely, 1-2 chromosomes tended to carry allele 22 (.561), and 2-2 seemed associated with allele 23 (.459). IVS6aGATT presented two alleles, 6 and 7, at frequencies of ~.33 and ~.66, respectively; however, among all 2-2 chromosomes the frequency of allele 6 was .766, whereas all other haplotypes carried allele 7 at frequencies of .72–.85. Haplotype 2-2 seemed to have also a preferential association with allele 17 at IVS8CA, which has an overall frequency of .257 but one of .623 in haplotype 2-2. Finally, the most striking associations at IVS17bTA were of 1-1 and 2-1 with allele 7; frequencies of allele 7 in those chromosomes were .787 and .939, respectively, whereas allele 7 has an overall frequency of .207.

The degree of association between the T854-TUB20 background and STR variability can be measured with AMOVA, which provides the fraction of genetic variability at each STR that is found within or among haplotype backgrounds. For IVS1CA, the fraction of genetic variability found among haplotype backgrounds obtained by weighting each allele independently of repeat number (also called F_{ST}) was .20; such values were .47 for IVS6aGATT, .34 for IVS8CA and .19 for IVS17bTA. It should be remarked that this kind of analysis is usually performed across populations rather than across chromosome backgrounds; the values by population are lower (.07–.13). A similar analysis in the Y chromosome (Bosch et al. 1999) yielded also much lower F_{ST} values across populations rather than across haplotype backgrounds, a result that implies that genetic backgrounds may be older than population origin. F_{ST} values by background would decrease with recombination and mutation rate. In particular, they can be used to draw inferences on mutation rate and patterns at IVS6aGATT. This STRP is practically diallelic, and at least two models can be suggested to explain this pattern: either (1) IVS6aGATT mutates at a high rate, but its allelic variation is strongly constrained to alleles 6 and 7, or (2) IVS6aGATT has an extremely low mutation rate. The two models predict different outcomes for F_{ST}: (1) it should be low in the first case, since mutation would have repeatedly placed both alleles in any background, and (2) it should be high if mutation rate is low. IVS6aGATT has the highest F_{ST} value among all four STRPs, which seems to be evidence for a low mutation rate at this locus.

F_{ST} by background, as a measure of association between stable chromosomal backgrounds and STRP alleles, can also be regarded as a function of LD. Next, we discuss other, more conventional measures of LD and apply them to CFTR polymorphisms.

Measuring LD

Recently, a new measure for LD among multiallelic loci, the ξ coefficient, has been suggested (Zhao et al. 1999). It is based on the standardized likelihood-ratio χ^2 statistic, and its significance can be obtained from the same permutation analysis used to generate it. We have assessed the performance of ξ by comparing it to one of the most-used LD measures for diallelic loci, D′ (Lewontin 1964). Three of the six loci we typed are diallelic (T854 and TUB20) or practically so (IVS6aGATT). We computed both ξ and D′ for the three pairs of loci among these diallelic loci for all the populations. Results showed that high, but nonsignificant D′ values, such as those obtained when allele frequencies at one allele are close to 0, always correspond to ξ ∼ 0. Given that, if one of the alleles at one of the loci is found at a low frequency, |D′| can easily reach 1 without any meaningful LD, we computed a correlation coefficient between ξ and |D′| by removing all cases with |D′| = 1. The correlation between the two SNPs (fig. 2) reached $r = .914$ ($P = .011$), and the correlation was $r = .697$ ($P = .054$) between IVS6aGATT and T854.

The performance of the significance of ξ was measured by comparing it to significance values obtained with the different likelihood-ratio test and permutation procedure previously suggested by Slatkin and Excoffier (1996) and implemented in Arlequin (Schneider et al. 2000). Both significance values were computed for each population and locus pair, and the correlation coefficient among them was $r = .943$ ($P < .001$). If we dichotomize the significance values according to an arbitrary significance level, we can compare how often both measures agree in accepting or rejecting LD. At a significance level of $P = .05$, both significance measures agreed in 94.6% of the cases. Of the 14 tests with discrepant results, 13 did not show significance by the Slatkin and Excoffier (1996) method but were significant according to ξ. At $P = .01$, results were very similar, with 93.9% agreement between both measures and 11 out of 16 cases in which ξ was less conservative than the method by Slatkin and Excoffier. Finally, if we corrected by multiple testing by using the Bonferroni correction ($P = .01$ divided by the number of loci pairs, i.e., 15; Sánchez-Mazas et al. 2000), agreement decreased to 90.4%, with the 25 discrepant cases divided into 13 cases in which ξ was more conservative and 12 cases in which ξ was less conservative. This result may be due to decreased precision at low P values.

In summary, for pairs of diallelic markers, D′ and ξ have similar values, although ξ is more robust to small allele frequencies, and the significance of ξ behaves much like that of the likelihood method devised by Slatkin and Excoffier (1996). This measure of LD seems suitable for comparisons among markers and genome regions.

LD Analysis: Relation to Physical Distance and Population Distribution

The ξ coefficient and its significance have been computed for each pair of loci and each population (table 6). In all but three populations (Yemenite, Adygei, and Chinese, $P < .05$), ξ is not correlated with physical distance–probably because of the alternation of markers with higher and lower levels of polymorphism, as discussed below. Figure 3 illustrates this situation by showing how high and low ξ values alternate among adjacent markers in four selected populations. This pattern can be explained, in part, by a correlation between ξ and locus polymorphism: the correlation between ξ and the

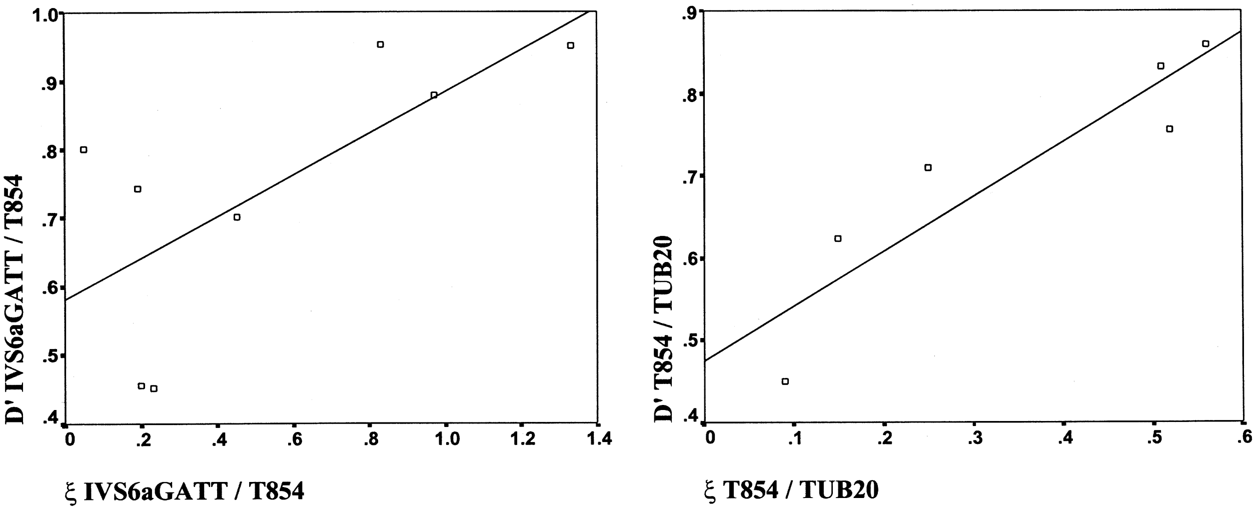

Figure 2 LD D′/ξ values correlation between loci IVS6aGATT / T854 and T854 / TUB20 (significance level for D′, $P < .05$)

Table 6

LD Pattern across Physical Size Intervals between Loci, in All Populations Analyzed

	Loci (Distance in kb)[a]														
Variable and Population	2-3 (12)	4-5 (17)	5-6 (41)	1-2 (46)	3-4 (47)	4-6 (58)	1-3 (58)	2-4 (59)	3-5 (64)	2-5 (76)	3-6 (105)	1-4 (105)	2-6 (117)	1-5 (122)	1-6 (163)
ξ:															
Biaka	.63	.22	.96	.39	.59	−.01	4.01	−.02	3.56	.39	.25	.59	.01	4.25	.63
Mbuti	1.24	.23	1.18	.60	.51	.08	2.60	−.02	2.36	.91	.31	.28	.02	2.57	.38
Saharawi	.40	.47	.76	.58	.22	.09	.50	−.01	.31	.25	−.02	.26	.03	1.62	.40
Tanzanians	.39	.24	.19	.49	.06	.03	1.00	.05	2.34	.37	.02	.31	.03	2.06	.28
Yemenites	.77	.22	.77	.34	.28	.15	.47	.23	.37	.35	−.03	.20	−.02	.44	.17
Druze	.51	.22	.65	.52	.33	.51	.87	−.01	1.33	.33	.04	.51	.00	.96	.55
Adygei	.61	.74	1.01	.12	.38	.56	.52	−.01	.28	.00	−.02	.15	.04	.19	.06
Russians	.58	.98	.66	.19	.26	.57	.36	−.02	.55	.12	.14	.52	.08	.42	.23
Finns	1.01	1.18	.74	.39	.15	.58	.87	.01	1.32	.41	.12	.39	.09	1.63	.24
Basques	.99	.63	.72	.77	.18	.25	.61	.20	.55	.36	.15	.58	.01	1.00	.07
Catalans	.57	.52	.75	.24	.11	.52	.77	.01	.58	.09	−.01	.49	.09	.87	.12
Kazakhs	1.72	−.02	.55	.38	1.13	.01	1.36	.45	.42	−.03	−.10	.45	.05	.93	.17
Chinese	.75	.27	.13	.41	.83	−.03	.73	.83	.22	.23	−.03	.42	−.02	.14	−.05
Japanese	1.18	.44	NC	1.06	.73	NC	1.22	.97	.55	.49	NC	.71	NC	.79	NC
Yakut	.74	.31	.48	.46	.14	.09	1.64	.19	.69	.10	−.04	.43	−.02	.17	.07
Nasioi	1.51	.79	.29	.59	.72	−.06	3.08	.39	2.63	.60	.41	1.07	−.04	2.01	−.20
Maya	1.11	.33	.05	.47	1.14	.03	.62	1.33	.53	.47	−.01	.66	.02	.44	.11
Surui	.99	1.08	NC	.27	1.34	NC	.30	.97	.98	.69	NC	.36	NC	.25	NC
Pr:[b]															
Biaka	0	.011	0	0	0	.735	0	.810	0	0	.001	0	.232	0	0
Mbuti	0	.068	0	0	0	.059	0	.831	0	0	.011	.022	.169	0	.007
Saharawi	0	0	0	0	.001	.013	.004	.503	.064	.011	.548	.002	.068	0	0
Tanzanians	0	.075	.119	.001	.205	.143	0	.085	0	.012	.364	.010	.171	0	.022
Yemenites	0	.033	0	.003	.006	.004	.035	.001	.124	.004	.591	.025	.681	.790	.047
Druze	0	0	0	0	0	0	0	.697	0	.001	.189	0	.305	0	0
Adygei	0	0	0	.154	0	0	.004	.412	.119	.446	.539	.041	.117	.235	.212
Russians	0	0	0	.050	.003	0	.063	.886	.027	.141	.032	0	.045	.107	.026
Finns	0	0	0	.001	.066	0	.001	.210	0	.012	.102	.003	.050	0	.035
Basques	0	0	0	0	0	0	0	0	0	.003	0	0	.230	0	.065
Catalans	0	0	0	0	.021	0	0	.139	.006	.112	.541	0	.004	.001	.020
Kazakhs	0	.524	0	.002	0	.296	0	0	.125	.544	.843	0	.062	.007	.064
Chinese	0	.010	.875	0	0	.712	0	0	.098	.015	.879	0	.754	.272	.773
Japanese	0	0	NC	0	0	NC	0	0	.001	0	NC	0	NC	.001	NC
Yakut	0	.002	0	0	.029	.003	0	0	0	.109	.650	0	.772	.183	.136
Nasioi	0	0	.124	0	0	.476	0	.003	0	.004	0	0	.408	0	1.000
Maya	0	.004	.295	0	0	.197	0	0	.001	0	.517	0	.170	.017	.057
Surui	0	0	NC	0	0	NC	0	0	0	0	NC	0	NC	.010	NC

[a] 1 = IVS1CA, 2 = IVS6aGATT, 3 = IVS8CA, 4 = T854, 5 = IVS17bTA, 6 = TUB20. NC = not computable because of the absence of variation.

[b] Pr(ξ) for 1,000 permutations (so that Pr of 0 means <.001).

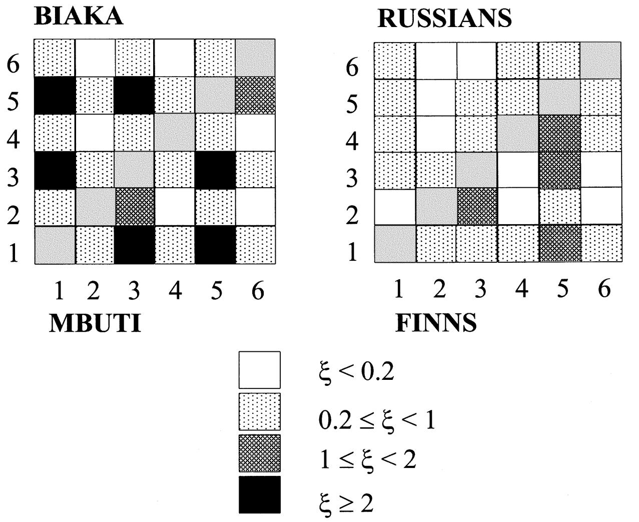

Figure 3 ξ values for four populations between all six loci (1 = IVS1CA, 2 = IVS6aGATT, 3 = IVS8CA, 4 = T854, 5 = IVS17bTA, and 6 = TUB20). Each square is a graphical representation of ξ levels for pairs of loci in two populations; each population is represented either above or below the diagonal (gray squares).

degrees of freedom in the corresponding haplotype table (which are equal to $k_1 - 1 \times k_2 - 1$, where k_1 and k_2 are the number of different alleles at the loci analyzed) is $r = .589$ ($P < .001$). This seems to be a property of LD, rather than a specific ξ bias. Slatkin (1994) found that the fraction of significant, nonrandom associations between alleles in mtDNA sequences grew with polymorphism; Sánchez-Mazas et al. (2000) found that LD increased with locus heterozygosity at the HLA region; and Ott and Rabinowitz (1997) showed that a more polymorphic marker provides increased statistical power to detect LD with a linked disease-causing mutation.

Average ξ was highest in the Africans, where it ranged from 0.52 in the Tanzanians to 1.10 in the Biaka. Middle Easterners, Europeans, and East Asians showed similar ranges of average ξ values, from 0.3 to 0.6. Among American Indians, average ξ was 0.48 in the Maya and 0.72 in the Surui; however, the latter value is not strictly comparable, since TUB20 was fixed in the Surui and ξ for 5 of the 15 loci could not be computed. Taken at face value, these results seem to indicate that LD is stronger in Africans than in other populations. However, we have shown that ξ grows with the number of alleles in the loci being compared, and that quantity is higher in Africans than in other populations. Thus, to assess the extent to which population history generated LD in Africans by means other than a higher STRP variation (Calafell et al. 1998), we should turn to ξ among the diallelic markers, which have practically the same number of alleles in all populations and should be free of the bias introduced by polymorphism on LD. IVS6aGATT and T854 show low, nonsignificant ξ values in Africans (−0.02 to 0.05), as well as in Middle Easterners and Europeans, who, except for the Yemenites (ξ = 0.23) and the Basques (ξ = 0.20), fall in the same range as the Africans. ξ is much higher in East Asians, Oceanians, and American Indians (0.19–1.33). T854 and TUB20 lie at a physical distance (59 kb) similar to that of the previous pair and likewise show low LD in Africans (−0.01 to 0.03), though it increases in Europeans and Middle Easterners (0.15–0.58) to decrease again in East Asians and the Maya (−0.06 to 0.09). Finally, global LD between IVS6aGATT and TUB20 is lower than in the two previous pairs—as expected, given the higher physical distance—and, in fact, ξ values are rather small and reach significance only in the Russians ($P = .045$) and in the Catalans ($P = .004$). In summary, pairs of diallelic markers at CFTR show reduced levels of LD in sub-Saharan Africans in comparison with other populations.

Discussion

The six polymorphisms in the CFTR region considered in the present paper have been thoroughly reported in CF chromosomes. Most of the existent marker and haplotype studies in clinical genetics are based on European populations (e.g., Hughes et al. 1995, 1996; Russo et al. 1995; Claustres et al. 1996; Morral et al. 1996); the study in our worldwide sample has shown important differences in allele and haplotype frequencies across populations. Several alleles have been reported for the first time.

It has been suggested that the high incidence of CF in Europeans (overall frequency of disease alleles ~2%) may be caused by heterozygote advantage against diarrheal diseases (Gabriel et al. 1994; Pier et al. 1998). However, such selection pressures are not expected to be a major factor in shaping worldwide CFTR variation, since >98% of chromosomes (100% in most continents) do not carry CF mutations.

Haplotype Phylogeny

Typing the T854 and TUB20 SNPs in nonhuman primate samples has allowed the inference of the ancestral states at those loci and the conclusion that 1-2 was the likely ancestral haplotype. It is also the most frequent haplotype and the chromosomes carrying it bear one of the highest STRP haplotype diversities. However, that diversity is slightly higher for 2-2, which may indicate that the mutation that produced the derived allele at T854 is older than its TUB20 counterpart. A simple prediction based on the ancestrality of 1-2 is that STRPs on that background should have the largest variance, which is not always the case. This apparent contradiction can be explained by at least three reasons. First, variance accumulation may not be linear with time and can even reach a plateau in which it ceases to grow with time (Goldstein et al. 1995). If that is the case for some of the oldest backgrounds, then STRP allele size variance may be a function of drift rather than of haplotype age (Di Rienzo et al. 1998). And, as we have seen that haplotype background determines STRP diversity to a greater extent than populations do, it is likely that haplotypes backgrounds are indeed older than populations. Second, the estimation of STRP allele-size variance has itself a large variance (Slatkin 1995), which may be the reason why variance-based genetic distances seem not to perform as well as those that do not take into account repeat size (Pérez-Lezaun et al. 1997; Calafell et al. 2000; Destro-Bisol et al. 2000). And third, an increase in STRP variance can be brought by repeated mutation at the presumed stable background or by recombination. In fact, a median network (Bandelt et al. 1995) constructed with the STRP haplotypes in the T854-TUB20 2-2 background showed two distinct and distantly related haplotype groups: a main group, with medium and large alleles at IVS17bTA, and a smaller group, with the 7 allele at IVS17bTA. This suggests that the extreme 7 allele could have been brought into a 2-2 background by recombination, thus greatly increasing repeat-size variance.

STRP Heterogeneity and CFTR: An STRP Spectrum

There is a vast heterogeneity in the diversity (as measured by heterozygosity or number of alleles) among STRPs, likely because of different mutation rates and patterns. See, for example, the ranges given by Calafell et al. (1998) for 45 CA-repeat polymorphisms. Several features, such as motif length (Chakraborty et al. 1997) and number of repeats (Brinkmann et al. 1998), have been suggested as contributors to mutation-rate variability across STRPs. Functional constraints can also play a role in determining number of repeats, as is exemplified by the disease-causing trinucleotide-repeat expansions. And yet, much of that heterogeneity is bound to be missed, given how most of the STRPs in the largest data sets (the linkage-mapping sets, for instance) were ascertained. Generally, libraries were screened with long $(CA)_n$ probes, as a rapid way of finding highly polymorphic markers. Thus, shorter, less polymorphic STRPs, or those with other motifs, may be underrepresented. In contrast, STRPs at CFTR were discovered from the whole sequence of the gene, and, although they are fewer, they may be a good, unbiased representation of STRP heterogeneity. We have typed all but one of the STRPs found in CFTR, and the range of polymorphism is remarkable: from two alleles accounting for 99.8% of the chromosomes at IVS6aGATT to 36 different alleles, ranging from 7 to 53 repeats, at IVS17bTA, with a corresponding 500-fold increase in allele-size variance.

STR variability depending on minihaplotype T854-TUB20 gives F_{ST} values higher than F_{ST} values depending on population. That is, STR allele frequency differences were greater between haplotype backgrounds than between populations. This suggests that the SNP mutation events that generated the haplotype backgrounds predate population differentiation processes.

The STRP analysis on a SNP haplotype background has allowed us to test two different models for mutation pattern at IVS6aGATT, and to reach the conclusion that it has a slow mutation rate, rather than a faster mutation rate and tight allele-size constraints. The fact that IVS6aGATT only has two alleles may be due to a low mutation rate, meaning it would be like an SNP, or to a normal mutation rate with constrictions in mutation pattern. Moreover, dinucleotides appear to have mutation rates 1.5–2 times higher than the tetranucleotides (Chakraborty et al. 1997). The variation pattern supports the first hypothesis, clarifying a debated point.

Genomic Effects on LD

LD, the nonrandom association of alleles at linked loci, is a powerful tool in gene mapping. It is often assumed that LD reflects genetic, and thus, physical distance (d), between a marker and a disease-causing mutation. However, differences in mutation rate can reverse the relation between LD and genetic distance among genetic markers (Calafell et al. 2001). Furthermore, it has been shown by Jorde et al. (1994) that, in a study of one locus in one population, there is a good correlation between LD and physical distance over 50–500 kb distances; but they do not correlate significantly when $d < 50–60$ kb. Kidd et al. (2000) showed that, in some populations, LD extended much farther than in others. Our results show a very complex pattern of LD, among the various sites, that is not a simple linear function of genetic distance. Part of this pattern may be caused by the relatively short genomic frame analyzed, in which recombination events may be rare and where the evolutionary variance of the effects of recombination may be large. In that situation, the effect of recombination becomes less predictable, particularly in relation to physical distance.

Allele diversity may also contribute to the LD pattern observed. Sánchez-Mazas et al. (2000) describes also a complex pattern of LD throughout the MHC region in a French population, where the significance of LD is not necessarily related to the physical distance between the loci they typed, but to allele diversity: pairs of loci with more alleles show stronger LD. This matches our findings, as well as the simulations by Ott and Rabinowitz (1997) and the analysis of mtDNA control-region sequences by Slatkin (1994). The combination of haplotypes with different degrees of polymorphism and with presumably different mutation rates has proved very fruitful in the understanding of different genome regions, such as CD4 (Tishkoff et al. 1996), DM (Tishkoff et al. 1998), DRD2 (Kidd et al. 1998), and the Y chromosome (Bosch et al. 1999). Such combinations provide both a stable background and markers that accumulate variation at a faster rate, which can then be used to date mutation events. However, care should be taken when measuring LD in such settings, particularly when SNP-SNP, SNP-STRP and STRP-STRP combinations are all found and the range of polymorphism across pairs of loci can determine LD to a much greater extent.

LD and Population History: How Many Went "Out of Africa?"

A number of studies of haplotypes consisting of several SNPs and, at most, one STRP (Tishkoff et al. 1996, 1998; Kidd et al. 1998; 2000) show a consistent population LD pattern: LD is small in Africans and grows stronger in Europeans, East Asians, and American Indians, up to the point that, at CD4 (Tishkoff et al. 1996), the authors found complete LD outside of Africa and conclude that there was only a single, *small* early migration of modern humans from Africa, which occurred <90,000 years before the present. It is also a recurrent result that Africans show higher allele diversity at STRPs (Bowcock et al. 1994; Jorde et al. 1997; Pérez-Lezaun et al. 1997; Calafell et al. 1998). This could explain why we find *stronger* LD in Africans, particularly among pairs of STRP loci. Is all LD at CFTR in Africans explainable by higher heterozygosity of STRPs? What was the underlying role of population history? A way around this conundrum is to consider the diallelic background, where Africans show little LD and, in some cases, are the population group with the lowest LD. A way of integrating STRP markers into this considerations would be through FNF, the fraction of possible different haplotypes that were not found in each population sample. Since the theoretical maximum (which depends on the number of different alleles at each locus) greatly exceeds sample size for each population, the effective maximum under linkage equilibrium is given by sample size and allele frequencies. By that measure, some European and Asian population samples cover the space of possible haplotypes more extensively than African samples do, which would indicate that the underlying LD is not lowest in Africans.

Some genetic studies suggest that the "Out of Africa" bottleneck was not so narrow (Ayala 1995; Helgason et al. 2000). If that were the case, it could be expected that LD in non-Africans in relation to Africans would follow a broad distribution, in which some loci, such as CD4, would show extreme LD only in non-Africans, whereas others, such as CFTR, should show similar amounts of LD in Africans and non-Africans. The combination of several types of markers at CFTR has allowed us to tackle the complicated interplay of genomic and population forces in creating and maintaining LD.

Acknowledgments

This research was supported by the Fundació La Marató de TV3-1998 (project "La Història Natural de la Fibrosi Quística: Interpretació Geogràfica de la Variació Genètica"). Further support was given by the Dirección General de Investigación Científica y Técnica (Spain; grant PB98-1064), by Direcció General de Recerca, Generalitat de Catalunya (grant 1998SGR00009), by Institut d'Estudis Catalans, and by United States National Science Foundation grant SBR9632509 (to J.R.K.). The work was also possible thanks to a fellowship to E.M. (Universitat de Barcelona and Universitat Pompeu Fabra). We would like to thank X. Estivill's staff (Departament de Genètica Molecular, Institut de Recerca Oncològica, Barcelona) for their technical support. Samples from Tanzania were kindly supplied by Dr. Clara Menéndez from Unitat d'Epidemiologia i Bioestadística (Hospital Clínic, Barcelona) and samples from Kazakhstan by Dr. Davide Pettener from Unità di Antropologia (Università di Bologna, Bologna). We would like to thank Montgomery Slatkin for useful comments on an earlier version of the manuscript.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Allele Frequency Database, http://info.med.yale.edu/genetics/kkidd (for allele and haplotype frequencies of the present study)

Arlequin package, http://anthropologie.unige.ch/arlequin (for analysis of molecular variance)

Cystic Fibrosis Mutation Data Base, http://www.genet.sickkids.on.ca/cftr

GenBank, http://www.ncbi.nlm.nih.gov/Genbank (for CFTR gene sequence [accession numbers AC000111, AC000061])

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim (for CFTR [MIM 602421], CF [MIM 219700])

References

Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270:1930–1936

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Bertranpetit J, Calafell F (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. In: Cardew G (ed) Variation in the human genome. Chichester, Wiley & Sons, pp 97–118

Bosch E, Calafell F, Santos FR, Pérez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, Bertranpetit J (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am J Hum Genet 65:1623–1638

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Brinkmann B, Klintschar M, Neuhuber F, Höhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62:1408–1415

Calafell F, Grigorenko EL, Chikanian AA, Kidd KK (2001) Haplotype evolution and linkage disequilibrium: a simulation study. Hum Hered 51:85–96

Calafell F, Pérez-Lezaun A, Bertranpetit J (2000) Genetic distances and microsatellite diversification in humans. Hum Genet 106:133–134

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism in humans. Eur J Hum Genet 6:38–49

Claustres M, Desgeorges M, Moine P, Morral N, Estivill X (1996) CFTR haplotypic variability for normal and mutant genes in cystic fibrosis families from southern France. Hum Genet 98:336–344

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94: 1041–1046

Chehab EF, Johnson J, Louie E, Goossens M, Kawasaki E, Erlich H (1991) A dimorphic 4-bp repeat in the cystic fibrosis gene is in absolute linkage disequilibrium with the ΔF508 mutation: implications for prenatal diagnosis and mutation origin. Am J Hum Genet 48:223–226

Cooper DN, Krawczak M (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Hum Genet 85:55–74

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Destro-Bisol G, Spedini G, Pascali VL (2000) Application of different genetic distance methods to microsatellite data. Hum Genet 106:130–132

Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, Petzl-Erler ML, Haines GK, Barch DH (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148: 1269–1284

Dörk T, Neumann T, Wulbrand U, Wulf B, Kälin N, Maass G, Krawczak M, Guillermit H, Férec C, Horn G, Klinger K, Kerem BS, Zielenski J, Tsui LC, Tümmler B (1992) Intra- and extragenic marker haplotypes of CFTR mutations in cystic fibrosis families. Hum Genet 88:417–425

Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat Genet 25:320–323

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

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science 266: 107–109

Gasparini P, Dognini M, Bonizzato A, Pignatti PF, Morral N, Estivill X (1991) A tetranucleotide repeat polymorphism in the cystic fibrosis gene. Hum Genet 86:625

Goldstein DB, Ruiz-Linares A, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate frequencies of multi-site haplotypes. J Hered 86:409–411

Helgason A, Sigurðardóttir S, Gulcher JR, Ward R, Stefánsson K (2000) MtDNA and the origin of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66:999–1016

Hughes D, Hill A, Redmond A, Nevin N, Graham C (1995) Fluorescent multiplex microsatellite used to identify haplotype association with 15 CFTR mutations in 124 Northern Irish CF families. Hum Genet 95:462–464

Hughes D, Wallace A, Taylor J, Tassabehji M, McMahon R, Hill A, Nevin N, Graham C (1996) Fluorescent multiplex microsatellites used to define haplotypes associated with 75 CFTR mutations from the UK on 437 CF chromosomes. Hum Mut 8:229–235

Iyengar S, Seaman M, Deinard AS, Rosenbaum HC, Sirugo G, Castiglione CM, Kidd JR, Kidd KK (1998) Analyses of cross-species polymerase chain reaction products to infer the ancestral state of human polymorphisms. DNA Seq 8: 317–327

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC (1997) Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 94:3100–3103

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

Jorde LB, Watkins WS, Kere J, Nyman D, Eriksson AW (2000) Gene mapping in isolated populations: new roles for old friends? Hum Hered 50:57–65

Kerem BS, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonné-Tamir B, Lu R-B, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, Kidd JR (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227

Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua FE, Odunsi A, Grigorenko E, Bonné-Tamir B, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet 66: 1882–1899

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49–67

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol 5:182–187

Mateu E, Calafell F, Bonné-Tamir B, Kidd JR, Casals T, Kidd KK, Bertranpetit J (1999) Allele frequencies in a worldwide survey of a CA repeat in the first intron of the CFTR gene. Hum Hered 49:15–20

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Giménez J, Reis A, et al. (1994) The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nat Genet 7:169–175

Morral N, Dörk T, Llevadot R, Dziadeek V, Mercier B, Férec C, Costes B, Girodon E, Zielenski J, Tsui L-C, Tümmler B, Estivill X (1996) Haplotype analysis of 94 cystic fibrosis mutations with seven polymorphic CFTR DNA markers. Hum Mut 8:149–159

Morral N, Estivill X (1992) Multiplex PCR amplification of three microsatellites within the CFTR gene. Genomics 13: 1362–1364

Morral N, Nunes V, Casals T, Estivill X (1991) CA/GT microsatellite alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene are not generated by unequal crossingover. Genomics 10:692–698

Moulin DS, Smith AN, Harris A (1997) A CA repeat in the first intron of the CFTR gene. Hum Hered 47:295–297

Ott J (2000) Predicting the range of linkage disequilibrium. Proc Natl Acad Sci USA 97:2–3

Ott J, Rabinowitz D (1997) The effect of marker heterozygosity on the power to detect linkage disequilibrium. Genetics 147:927–930

Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J (1997) Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1–7

Pier GB, Grout M, Zaidi T, Meluleni G, Mueschenborn SS, Banting G, Ratcliff R, Evans MJ, Colledge WH (1998) *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. Nature 393:79–82

Quere I, Guillermit H, Mercier B, Audrezet MP, Ferec C (1991) A polymorphism in intron 20 of the CFTR gene. Nucleic Acids Res 19:5453

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui LC (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066–1073

Rommens JM, Iannuzzi MC, Kerem BS, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui LC, Collins FS (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 245:1059–1065

Russo MP, Romeo G, Devoto M, Barbujani G, Cabrini G, Giunta A, D'Alcamo E, Leoni G, Sangiuolo F, Magnani C, Cremonesi L, Ferrari M (1995) Analysis of linkage disequilibrium between different cystic fibrosis mutations and three intragenic microsatellites in the Italian population. Hum Mut 5:23–27

Sánchez-Mazas A, Djoulah S, Busson M, Le Monnier de Gouville I, Poirier J-C, Dehay C, Charron D, Excoffier L, Schneider S, Langaney A, Dausset J, Hors J (2000) A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. Eur J Hum Genet 8:33–41

Schneider S, Kueffer JM, Roessli D, Excoffier L (2000) Arlequin (ver. 2.000): A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. Genetics 137:331–336

——— (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

——— (2000) Balancing selection at closely linked, overdominant loci in a finite population. Genetics 154:1367–1378

Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity 76:377–383

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Cheung K, Kidd JR, Bonné-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonné-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the DM locus: implications for the evolution of modern humans and the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518–522

Zhao H, Pakstis AJ, Kidd KK, Kidd JR (1997) Overall and segmental significance levels of linkage disequilibrium. Am J Hum Genet Suppl 61:A17

Zhao H, Pakstis AJ, Kidd JR, Kidd KK (1999) Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. Ann Hum Genet 63: 167–179

Zielenski J, Markiewicz D, Rininsland F, Rommens JM, Tsui LC (1991*a*) A cluster of highly polymorphic dinucleotide repeats in intron 17b of the CFTR gene. Am J Hum Genet 49:1256–1262

Zielenski J, Rozmahel R, Bozon D, Kerem BS, Grzelczak Z, Riordan JR, Rommens JM, Tsui LC (1991*b*) Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Genomics 10:214–228