European Journal of Human Genetics (2003) 11, 385–394

www.nature.com/ejhg

ARTICLE

Spatial patterns of cystic fibrosis mutation spectra in European populations

Oscar Lao[1], Aida M Andrés[1], Eva Mateu[1], Jaume Bertranpetit[1] and Francesc Calafell*,[1]

[1]*Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain*

Cystic fibrosis (CF) is the most frequent severe recessive disorder in European populations. We have analyzed its mutation frequency spectrum in 94 European, North African and SW Asian populations taken from the literature. Most major mutations as well as the incidence of CF mutations showed clinals patterns as demonstrated by autocorrelogram analysis. More importantly, measures of mutation diversity did also show clinal patterns, with mutation spectra being more diverse in southern than in northern Europe. This increased diversity would imply roughly a three-fold long-term effective population size in southern than in northern Europe. Distances were computed among populations based on their CF mutation frequencies and compared with distances based on other genic regions. CF-based distances correlated with mtDNA but not with Y-chromosome-based distances, which may be a consequence of the relatively homogeneous CF mutation frequencies in European populations.
European Journal of Human Genetics (2003) **11,** 385–394. doi:10.1038/sj.ejhg.5200970

Keywords: cystic fibrosis; spatial autocorrelation; mutation spectra; effective population size; genetic diversity; spatial patterns; Europe

Introduction

Cystic fibrosis (CF) is the most frequent severe recessive disorder in European populations. The mutation spectrum of this disease, that is, the pool of alleles that cause CF in homozygosis or in heterozygosis with another mutation, is made up of more than 1000 different mutations;[1] of those, a three-basepair deletion, F508del, accounts for roughly two-thirds of cases, and only four others are found at average frequencies $>1\%$.

The geographic distribution of mutation frequencies is heterogeneous among European populations, showing a high geographic variation; F508del is ubiquitous, although its frequency ranges from less than 50% to almost 100%; other mutations reach significant frequencies only in part of the European continent, and many are rare and population specific. Most of the knowledge about frequency distributions has been obtained on a mutation-by-mutation basis and has consisted of mapping mutation frequencies,[2] and the consequent association of their geographic distribution with hypothetic demographic expansions of historical populations such as the Celts or the Phoenicians. However, this single-mutation approach fails to capture the complexity of both the mutation spectrum and the population history. Gene flow among populations tends to disperse a number of mutations simultaneously and not just one. Moreover, too often these *ad hoc* explanations are put forward without any quantitative considerations about the extent of gene flow required or for the basic population genetics involved.

A different, sounder approach is possible: considering the whole mutation spectrum at once and, beyond the mere description and subsequent story telling, applying to it the basic toolkit of population genetics, which can lead to more accurate interpretations of the history of the populations in relation to the natural history of genetic diseases. One basic requirement for that approach is the reasonable assumption that all CF mutations are selectively equivalent. That is, they cause the same loss of fitness in

*Correspondence: Dr F Calafell, Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain. Tel: +34 93 542 28 41; Fax: +34 93 542 28 02; E-mail: francesc.calafell@cexs.upf.es
Received 3 September 2002; revised 21 January 2003; accepted 24 January 2003

homozygosity and they give the same advantage to heterozygotes.[3–6] Under this assumption, Reich and Lander[7] modelled mutation diversity as a function of population history and of the overall frequency of disease-causing alleles. However, they considered one average spectrum per disease, and, as we show below, allele diversity can be very different among populations for the same disease. A basic property of a mutation set is its frequency spectrum, that is, how many mutations are found and at frequency in a particular population. It can be understood acumulatively, that is, how many mutations are required to explain a given relative frequency (say 90%) of CF chromosomes. This has an obvious interest in clinical testing, but it is also crucial in population genetics, since a frequency spectrum may be the result of population history and of other evolutionary factors.

We have compiled CFTR mutation frequencies for a wide set of populations from Europe, North Africa and SW Asia from the literature and have described their spatial patterns, both of a few single mutations and of mutation diversity. We present an interpretation of our findings grounded on the population genetics theory that relates CF mutation frequency spectra to the patterns of effective population size across European populations and to selection.

Materials and methods

Databases

Mutation frequency spectra for CF were compiled from the bibliography for populations (countries or main regions within countries) from Europe, North Africa, and SW Asia, that is, areas with a relatively high prevalence of CF and where mutation spectra can be reported for a reasonable number of patients. Since we are interested in the analysis of geographical patterns in CF mutation spectra in a historical frame, data for Americans and Australians of European descent were not taken into account. These populations are clearly related to northern Europeans but are located at a large distance, which would distort the spatial analyses. For each population, the frequency of each allele associated to the disease, the geographical location and the sample size were considered. In each population, a fraction of CF chromosomes remained with an unrecognized mutation; this frequency of unknown mutations depends on technical limitations that may vary across studies. Those populations with an indetermined location, or with a sample size of less than 20 chromosomes, or with a frequency of unknown mutations over 60%, were excluded from further analysis, which gave a total in the database of 94 populations with 32 mutations found in more than one population and 109 private (ie population specific) mutations. CF incidences were taken from,[8–23] where incidences were estimated either from neonatal screening programs or from general population screens; incidence values were available for 16 populations, 13 of which referred to whole countries. In those cases, data such as mutation frequencies that were compared to incidences were first pooled within the country.

Genetic diversity estimates

Two different genetic diversity estimates were obtained for each population using the Arlequin package:[24] the expected heterozygosity and an estimate of θ based on expected heterozygosity.[25,26] The latter is an estimate of $4N_e\mu(1-f_0)$, where N_e is the effective population size, μ is the mutation rate and f_0 is the frequency of the overall disease allele class.[7] Computation of both the expected heterozygosity and θ requires the complete specification of the frequencies of all alleles. However, an average 24.8% (ranging from 0 to 51.3%) of the CF chromosomes carried unknown CFTR mutations. Estimates of allele diversity are bracketed between two extremes: they would be minimal if all unknown mutations in a population were, in fact, the same allele, and they would be maximal if all unknown mutations were different from each other. Intensive mutation-detection efforts by means of denaturing high-performance liquid chromatography[27] have shown that chromosomes bearing previously unknown mutations carried each a new, different, unique mutation. Thus, it is likely that the 'unknown' portion of the mutation spectrum contains a wide diversity of alleles. Under that assumption, and assuming as well the absence of phenocopies (non-CF patients diagnosed as having CF) and genocopies (CF caused by mutations in genes other than CFTR), mutation diversity has been estimated by assuming that all chromosomes in the 'unknown mutation' category carry each a different mutation. This implies that, given the frequency of the 'unknown mutation' category, we have considered the maximum allele diversity estimate.

Geographical patterns

A geographical description of allele frequency patterns was obtained by drawing maps of gene frequencies for the most common CF mutations (namely F508del, G452X, G551D, N1303 K, and W1285X, which are the ones found at average frequencies $>1\%$), as well as for genetic diversity estimates, by using Surfer 7.0 (http://www.goldensoftware.com) with the inverse-squared distance method. A regular grid covering Europe, North Africa, and the Middle East and limited between 30°N and 64°N and between 10°W and 42°E was used. Interpolation points were spaced 0.1°. For each interpolation point, only data points within the same landmass (island or continent) were considered. It should be noted that interpolation was used only to map allele frequencies and diversities, and that interpolated values were not used in any other analysis.

The frequency of the most common mutations, maximum genetic diversity estimates and countrywide CF

incidences were subjected to spatial autocorrelation analysis[28] by means of the SAAP program (http://www.exeter-software.com/). Autocorrelation analysis, which consists in plotting a measure of correlation among pairs of populations classified according to the geographical distance between them, allows to characterize geographical patterns such as clines (gradients), depressions (clines irradiating from the center of the area considered), and isolation by distance, since each of these patterns leads to autocorrelation plots that are statistically significantly different from each other (see examples in Barbujani[29]). Thus, spatial autocorrelation analysis allows one to describe objectively the spatial patterns and to compare them with the expectations derived from demographic hypotheses, such as growth and migration.

Genetic distances

Reynolds' genetic distances[30] based on CF mutation relative frequencies were computed. The same measure of genetic distance was used by Cavalli-Sforza *et al*[31] to estimate genetic distances among European populations based on *classical* genetic polymorphisms (ie blood groups, protein polymorphisms, and HLA). Both distance matrices were compared by means of a Mantel test, which calculates a nonparametric index of matrix correlation;[32] correction by a geographic distance matrix was performed in order to make partial Mantel tests controlled by this variable.[33] Reynolds' distances were also computed and compared to CF mutation distances for Y-chromosome haplogroup frequencies.[34] Finally, CF mutation distances were compared to corrected pairwise distances for hypervariable region I mitochondrial DNA sequences;[35–37] since some slightly negative values were obtained, a small positive quantity was added to all distance values so that all would be positive. Genetic distance calculations and Mantel tests were performed with Arlequin 2.000.[24]

Results

CF mutation frequency spectra were gathered for 94 populations from Europe, SW Asia, and North Africa. Information about populations, sample sizes, main mutation frequencies, and mutation diversities can be found in Table 1. In all, 32 different mutations were considered; complete mutation frequencies and other information can be found at http://www.upf.es/cexs/recerca/bioevo/index.htm. As seen in Figures 1–5, the most common CF mutations display clinal or largely clinal patterns as determined by their spatial correlograms. As for the direction of the clines, F508del peaks in NW Europe and declines towards SE Europe, G542X declines from SW to NE Europe, G551D is almost restricted to NW Europe, and N1303 K and W1282X show gradients from SW Asia and SE Europe towards NW Europe.

The richness of a mutation frequency spectrum can be summarized by several genetic parameters, such as allele diversity and θ. Maximum allele diversity (see Material and methods) ranges from less than 0.4 to over 0.9, and shows a significant correlogram with a negative peak around 3000 km ($P<0.0001$, Figure 6(a)), with maxima in southern Europe (Turkey, Italy, Spain) and minima in northern and NW Europe (Denmark, Britain). Given the predominant S–N orientation of the cline, the maximum CF mutation diversity is strongly and negatively correlated with latitude ($r=-0.587$, $P<0.001$), although it is also slightly correlated with longitude ($r=0.220$, $P=0.033$). θ shows a similar spatial pattern, with a three-fold decrease from southern to northern/NW Europe, a significant correlogram ($P<0.001$, Figure 6(b)), and a high correlation with latitude ($r=-0.394$, $P<0.001$).

The previous analyses were performed on the mutation frequencies relative to the total CF chromosomes. For instance, on an average 70% of CF chromosomes carry F508del, but, given the average CF incidence (1/2500 newborns), which implies that on an average 2% of chromosomes carry a CF mutation, 1.4% of all chromosomes carry F508del. It should be noted that the prevalence of CF in Europe is irregular and slightly correlated with longitude ($r=-0.533$, $P=0.028$, growing from east to west), and although the overall correlogram is significant ($P=0.001$), the pattern it shows is not significant at high geographical distance classes, and therefore can be interpreted as reflecting only isolation by distance.[38] Furthermore, we did not find any correlation between CF incidence and F508del mutation frequency ($r=-0.002$; $P=0.994$).

The relation among different CF mutation spectra across European populations can be investigated using standard population genetic methods, in order to relate it to the population history of the continent as reflected in other genome regions.

Genetic distances among CF mutation pools (ie a measure of the difference in relative CF mutation frequencies among pairs of populations) in different European populations were computed. By a Mantel test, we found a significant correlation between distances based on CF mutations and geographic distances ($r=0.329$, $P<0.0005$). For a subset of 23 populations, genetic distances based on classical polymorphisms[31] were available. This was also the case for 27 populations and mtDNA control region sequences[35–37] and 28 populations and Y-chromosome haplogroups.[34] Classic and CF distances were not correlated ($r=-0.039$, $P=0.48$), even when controlling for geographic distance ($r=-0.044$, $P=0.52$). When known outliers such as Sardinians and Basques were removed from the analysis, the correlations increased, although without reaching statistical significance ($r=0.083$ and after controlling for geographical distance, $r=0.090$). The correlations with Y-chromosome-based

Table 1 94 middle Eastern, North African and European populations used in the analysis

Population	*2N*	*F508del*	*G542X*	*G551D*	*N1303K*	*W1282X*	*Rare*	*Other*	*Unknown*	H_{max}	Θ_{max}	*Incidence*	*References*[a]
Austria	592	0.660	0.022	0.012	0.005	0.002	0.064	0.019	0.216	0.562	0.96		49
Belgium	646	0.752	0.025	0.002	0.028	0.012	0.053	0.046	0.082	0.430	0.56		50
Bulgaria	208	0.654	0.034	0.000	0.067	0.000	0.096	0.067	0.082	0.563	0.97		51
Crete	26	0.462	0.077	0.000	0.038	0.000	0.231	0.038	0.154	0.785	2.87		
Czech Republic	584	0.697	0.021	0.034	0.026	0.005	0.051	0.021	0.146	0.512	0.78		
Denmark	678	0.872	0.006	0.001	0.010	0.001	0.034	0.035	0.040	0.239	0.23	0.000210	
Great Britain												0.000414[b]	
North England	4111	0.772	0.008	0.023	0.005	0.001	0.032	0.011	0.148	0.403	0.50		
Scotland	1167	0.751	0.033	0.061	0.003	0.003	0.033	0.023	0.093	0.430	0.56	0.000504	
South England	3679	0.769	0.020	0.029	0.005	0.002	0.032	0.009	0.133	0.407	0.51		
Wales	372	0.659	0.024	0.030	0.005	0.000	0.134	0.065	0.083	0.557	0.94		
Estonia	25	0.640	0.000	0.000	0.000	0.000	0.160	0.080	0.120	0.577	1.02		
Former Yugoslavia	203	0.700	0.030	0.000	0.010	0.005	0.039	0.069	0.148	0.506	0.76		
Finland	52	0.462	0.019	0.000	0.000	0.000	0.288	0.019	0.212	0.713	1.90		
France												0.000232	
Alsace	126	0.595	0.024	0.000	0.016	0.008	0.040	0.008	0.310	0.646	1.38		
Aquitaine	116	0.612	0.034	0.000	0.017	0.000	0.043	0.009	0.284	0.626	1.26		
Auvergne	102	0.725	0.039	0.000	0.029	0.010	0.020	0.000	0.176	0.474	0.67		
Burgundy	168	0.702	0.024	0.000	0.006	0.000	0.060	0.006	0.202	0.507	0.77		
Brittany	582	0.744	0.009	0.024	0.017	0.003	0.064	0.002	0.137	0.444	0.60	0.000343	
Centre	218	0.716	0.050	0.000	0.023	0.000	0.023	0.000	0.188	0.486	0.71		
Champagne	182	0.665	0.049	0.000	0.016	0.000	0.055	0.005	0.209	0.556	0.94		
Franche-Comte	118	0.746	0.085	0.000	0.085	0.025	0.059	0.000	0.000	0.431	0.56		
Languedoc	90	0.700	0.022	0.011	0.033	0.000	0.044	0.000	0.189	0.511	0.78		
Llimousin	44	0.545	0.023	0.000	0.068	0.000	0.023	0.023	0.318	0.705	1.83		
Loire Valley	308	0.737	0.006	0.019	0.013	0.003	0.032	0.000	0.188	0.457	0.63		
Lorraine	286	0.717	0.031	0.000	0.000	0.000	0.042	0.000	0.210	0.486	0.70		
Lower Normandie	174	0.644	0.017	0.023	0.017	0.000	0.069	0.000	0.230	0.585	1.06		
Midi-Pyrenees	114	0.649	0.035	0.000	0.018	0.009	0.018	0.000	0.272	0.580	1.03		
Nord	468	0.660	0.019	0.004	0.015	0.002	0.053	0.006	0.239	0.563	0.97		
Paris Region	830	0.643	0.027	0.007	0.010	0.012	0.035	0.000	0.266	0.585	1.06		
Picardie	200	0.650	0.040	0.000	0.040	0.010	0.080	0.000	0.180	0.574	1.01		
Poitou	100	0.770	0.030	0.000	0.020	0.000	0.020	0.000	0.160	0.408	0.51		
Provence-Cote d'Azur	178	0.674	0.028	0.000	0.051	0.017	0.028	0.006	0.197	0.544	0.89		
Rhone-Alpes	668	0.668	0.036	0.001	0.027	0.009	0.018	0.009	0.232	0.552	0.92		
Upper Normandie	248	0.645	0.020	0.008	0.012	0.004	0.048	0.004	0.258	0.584	1.05		
Germany													
Baden-Wurttemberg	59	0.763	0.000	0.000	0.034	0.000	0.051	0.102	0.051	0.412	0.52		
Bavaria	177	0.740	0.017	0.017	0.000	0.000	0.040	0.011	0.175	0.453	0.62		
Berlin[c]	132	0.773	0.015	0.000	0.023	0.000	0.038	0.015	0.136	0.403	0.50		
Bremen[d]	74	0.689	0.014	0.014	0.000	0.000	0.054	0.014	0.216	0.528	0.84		
Lower Saxony	198	0.803	0.005	0.005	0.015	0.000	0.015	0.030	0.126	0.355	0.41		
North-Rhine/ Westphalia	174	0.736	0.006	0.006	0.000	0.006	0.069	0.034	0.144	0.458	0.63		
Saxony[e]	83	0.639	0.012	0.012	0.024	0.000	0.036	0.036	0.241	0.594	1.10		
Rhineland-Palatina[f]	59	0.525	0.017	0.000	0.051	0.000	0.085	0.068	0.254	0.721	1.99		
Greece													
Ipiros/Ionian Islands	46	0.609	0.000	0.000	0.043	0.000	0.087	0.043	0.217	0.632	1.30		
Peloponese/Attica	89	0.573	0.000	0.022	0.045	0.000	0.146	0.045	0.169	0.667	1.52		
Thesalia/Macedonia/ Thrace	61	0.672	0.066	0.000	0.033	0.000	0.033	0.082	0.115	0.543	0.89		
Hungary	57	0.439	0.018	0.000	0.018	0.018	0.070	0.018	0.421	0.811	3.43		
Italy													
Abruzzo	66	0.500	0.061	0.000	0.091	0.076	0.030	0.000	0.242	0.739	2.19		
Basilicata	75	0.467	0.107	0.000	0.067	0.027	0.067	0.013	0.253	0.769	2.61		
Calabria	149	0.430	0.034	0.000	0.047	0.020	0.054	0.034	0.383	0.813	3.46		

Table 1 (*continued*)

Population	*2N*	*F508del*	*G542X*	*G551D*	*N1303K*	*W1282X*	*Rare*	*Other*	*Unknown*	H_{max}	Θ_{max}	*Incidence*	*References*[a]
Campania	223	0.610	0.040	0.000	0.067	0.018	0.040	0.004	0.220	0.623	1.25		
Emilia-Romagna	242	0.541	0.058	0.000	0.025	0.008	0.050	0.000	0.318	0.704	1.82	0.000170	
Friuli	24	0.375	0.125	0.000	0.042	0.042	0.083	0.083	0.250	0.855	4.85		
Lazio	236	0.462	0.030	0.000	0.093	0.013	0.034	0.013	0.356	0.778	2.75		
Liguria	44	0.591	0.114	0.000	0.023	0.000	0.045	0.000	0.227	0.646	1.38		
Lombardia	399	0.499	0.038	0.000	0.038	0.010	0.090	0.050	0.276	0.743	2.24		
Marche	144	0.389	0.056	0.000	0.083	0.014	0.063	0.007	0.389	0.841	4.29		
Molise	27	0.481	0.037	0.000	0.074	0.000	0.037	0.000	0.370	0.775	2.70		
Piemonte	117	0.675	0.034	0.000	0.000	0.000	0.043	0.017	0.231	0.544	0.89		
Puglia	245	0.543	0.053	0.000	0.073	0.000	0.041	0.012	0.278	0.698	1.77		
Sardegna	141	0.582	0.057	0.000	0.028	0.000	0.028	0.142	0.163	0.641	1.35		
Sicilia	387	0.525	0.062	0.000	0.034	0.023	0.067	0.021	0.269	0.719	1.97		
Toscana	191	0.508	0.042	0.000	0.037	0.010	0.031	0.005	0.366	0.740	2.21		
Trentino	113	0.513	0.027	0.009	0.009	0.009	0.204	0.053	0.177	0.718	1.96		
Umbria	37	0.676	0.081	0.000	0.027	0.000	0.027	0.000	0.189	0.545	0.90		
Veneto	552	0.449	0.014	0.000	0.031	0.000	0.188	0.033	0.284	0.785	2.87	0.000370	
Ireland													
Republic of Ireland	509	0.727	0.010	0.069	0.004	0.000	0.037	0.014	0.139	0.467	0.65	0.000684	
Northern Ireland	876	0.619	0.021	0.045	0.001	0.000	0.063	0.047	0.205	0.612	1.19	0.000553	[52]
Israel	367	0.322	0.054	0.000	0.030	0.362	0.065	0.082	0.084	0.754	2.39	0.000304	
Lebanon	40	0.350	0.000	0.000	0.100	0.200	0.025	0.225	0.100	0.794	3.04	0.000390	[53]
Netherlands	1442	0.744	0.013	0.001	0.009	0.007	0.072	0.019	0.135	0.444	0.60	0.000252	
Norway	168	0.667	0.006	0.012	0.006	0.000	0.071	0.000	0.238	0.555	0.93	0.000152	
Poland	444	0.662	0.023	0.007	0.020	0.002	0.043	0.020	0.223	0.560	0.96		
Portugal													
Faro/Beja	25	0.680	0.000	0.000	0.000	0.000	0.040	0.000	0.280	0.547	0.90		
Lisboa[g]	100	0.480	0.030	0.000	0.000	0.000	0.080	0.060	0.350	0.767	2.57		
Setubal/Evora	33	0.485	0.000	0.000	0.000	0.000	0.121	0.091	0.303	0.767	2.57		
Russia	445	0.618	0.007	0.002	0.004	0.004	0.031	0.031	0.301	0.617	1.22	0.000051	
Slovakia	254	0.559	0.075	0.000	0.035	0.016	0.075	0.016	0.224	0.680	1.62		
Spain													
Andalucía	314	0.538	0.086	0.013	0.013	0.013	0.083	0.096	0.159	0.694	1.73		
Aragón	65	0.523	0.031	0.000	0.015	0.000	0.123	0.138	0.169	0.708	1.86		
Castilla la Mancha	69	0.478	0.058	0.000	0.043	0.000	0.014	0.029	0.377	0.771	2.63		
País Valencià	125	0.464	0.104	0.000	0.056	0.000	0.096	0.040	0.240	0.771	2.63		
Castilla León/ La Rioja	187	0.604	0.048	0.000	0.011	0.000	0.102	0.107	0.128	0.623	1.24		[54]
Catalonia	109	0.642	0.055	0.000	0.037	0.009	0.083	0.064	0.110	0.582	1.05	0.000187	
Extremadura	63	0.460	0.048	0.000	0.016	0.000	0.079	0.127	0.270	0.776	2.72		
Galicia	93	0.624	0.097	0.000	0.011	0.000	0.161	0.075	0.032	0.596	1.11		
Madrid	51	0.510	0.059	0.020	0.039	0.000	0.059	0.020	0.294	0.742	2.23		
Murcia	40	0.250	0.125	0.000	0.025	0.025	0.175	0.200	0.200	0.889	6.74		
Basque Country	31	0.710	0.000	0.000	0.000	0.000	0.065	0.097	0.129	0.497	0.74		
Sweden	165	0.733	0.006	0.000	0.000	0.000	0.103	0.085	0.073	0.448	0.60	0.000130	
Switzerland	95	0.432	0.032	0.000	0.011	0.000	0.263	0.168	0.095	0.732	2.11		
Tunisia	78	0.179	0.090	0.000	0.064	0.026	0.128	0.115	0.397	0.941	14.51		[55]
Turkey	263	0.274	0.038	0.000	0.042	0.004	0.087	0.114	0.441	0.907	8.44		[56]
Ukraine	396	0.543	0.000	0.005	0.000	0.000	0.018	0.000	0.434	0.706	1.84		[57]
Total	29131	0.674	0.025	0.015	0.017	0.009	0.053	0.024	0.182	0.586	1.06		

N, sample size (in number of CF chromosomes); F508del, G542X, G551D, 1303K, and W1282X, relative frequencies of the main mutations; rare, relative frequency of mutations not listed in Table 2 of reference 58; other, relative frequency of mutations listed in Table 2 of reference 58. unknown, fraction of chromosomes asociated to disease bearing unidentified mutations. H_{max} and Θ_{max}, allele diversity parameters (see text).[a]Except when it is indicated, data corresponds to Estivill et al.[58] [b]North & South England. [c]Berlin/Brandenburg/Mecklenburg-Vorpommern. [d]Bremen/Hamburg/Schleswig-Holstein. [e]Saxony/Saxony-Anhalt/Thuringia. [f]Rhineland-Palatina/Hesse/Saarland. [g]Lisboa/Santarem/Portalegre/Castelo Branco.

distances were similar ($r=0.147$, $P=0.116$ and after controlling for geographical distance $r=0.054$, $P=0.296$). A higher and significant correlation was obtained for mtDNA control region sequences ($r=0.423$, $P=0.004$, and after controlling by geographic distance $r=0.425$, $P=0.004$).

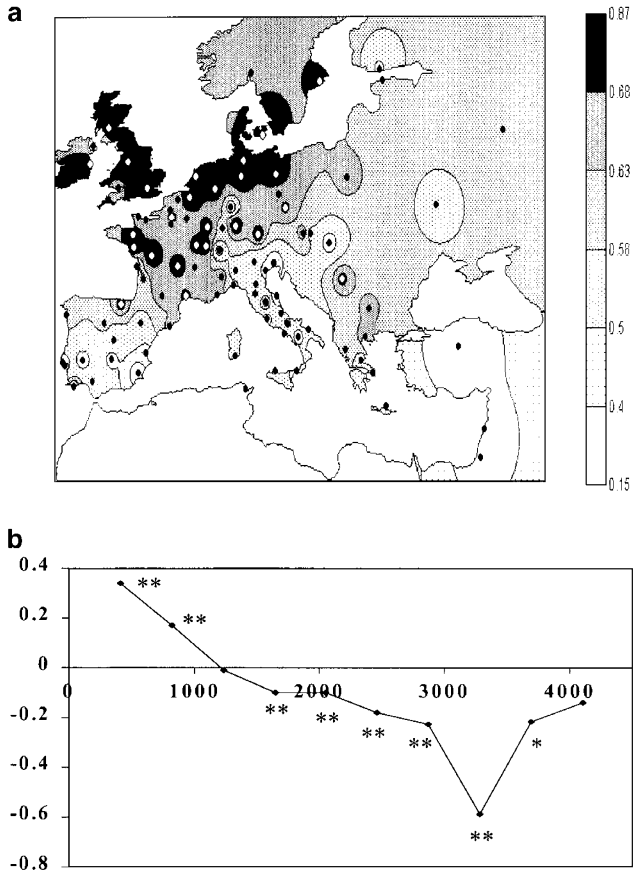

Figure 1 Geographical distribution (**a**) and spatial autocorrelogram (**b**) of F508del mutation in 94 middle Eastern, North African, and European populations. The *X*-axis represents geographic distance between samples; the *Y*-axis represents Moran's index; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$.

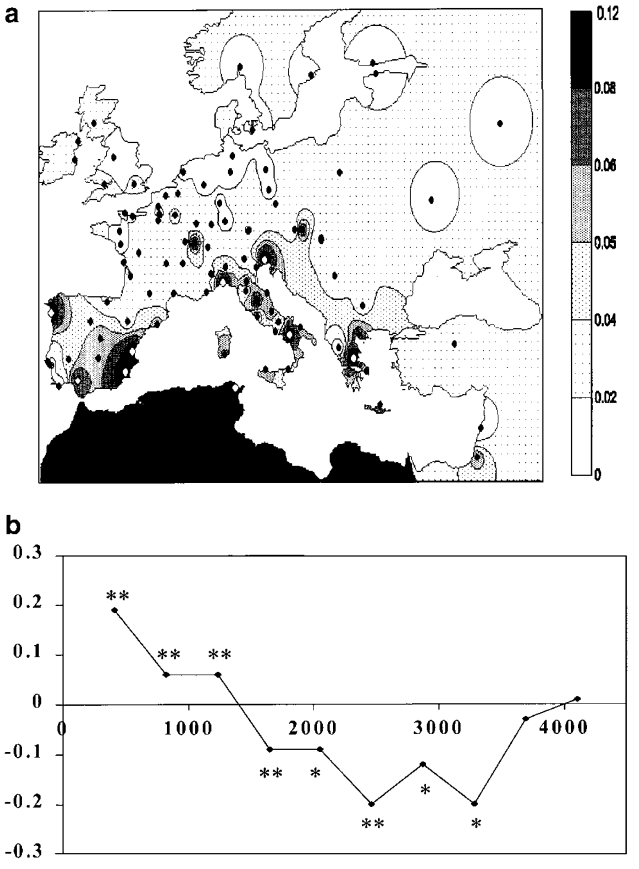

Figure 2 Geographical distribution (**a**) and spatial autocorrelogram (**b**) of the G542X mutation in 94 middle Eastern, North African, and European populations. The *X*-axis represents geographic distance between samples; the *Y*-axis represents Moran's index; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$.

We also analyzed the correlation between CF diversity and Y-chromosome and mtDNA control region diversity. In both cases, correlation was positive, although not statistically significant ($r=0.308$, $P=0.111$ for CF-Y chromosome and $r=0.319$, $P=0.105$ for CF-mtDNA diversity).

Discussion

We have described in detail the geographical variation pattern of the main CF mutations in Europe and in the immediately adjacent territories of Asia and Africa, and we have also characterized the diversity of mutation frequency spectra and their spatial pattern in this geographical frame. We have confirmed that the main CF mutations show frequency clines in Europe, and that this is also the case for mutation diversity. Mutation spectra are richer in southern than in northern Europe, but the pattern is not simply related to latitude. This has practical implications, since, on average, more mutations need to be assayed before finding the one(s) responsible for a particular case in southern than in northern Europe.

Measuring mutation diversity as θ opens the possibility of explaining this pattern. In a population at equilibrium, this parameter is an estimator of $4N_e\mu(1-f_0)$, where N_e is the effective population size, μ is the mutation rate, and f_0 is the prevalence of the overall disease allele class.[7] Mutation rate is highly unlikely to be higher in southern than in northern Europe; therefore, a higher θ may be caused by a higher N_e and/or lower f_0 in southern Europe. That is, we need to explore whether spatial patterns in incidence or effective population size may be related to CF mutation diversity. Estimates of CF incidence would indicate that the CF allele class is 3.8 times more frequent in Ireland than in Russia, the two extremes of the incidence of CF in Europe (see Table 1). This may pose a problem and may seem to have contributed to the observed spatial patterns in CF mutation diversity. However, the effects of the two parameters are not equivalent:

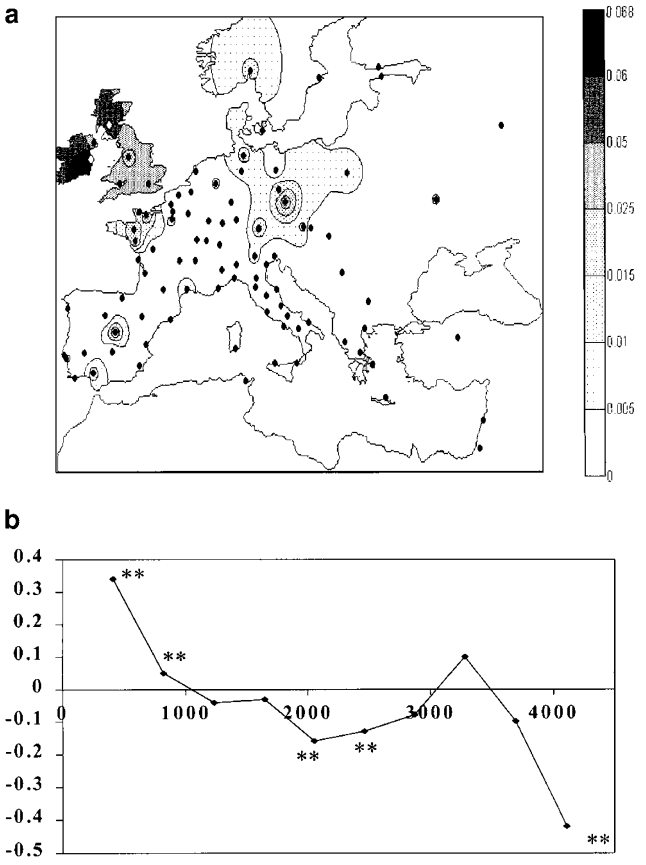

Figure 3 Geographical distribution (**a**) and spatial autocorrelogram (**b**) of the G551D mutation in 94 middle Eastern, North African, and European populations. The *X*-axis represents geographic distance between samples; the *Y*-axis represents Moran's index; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$.

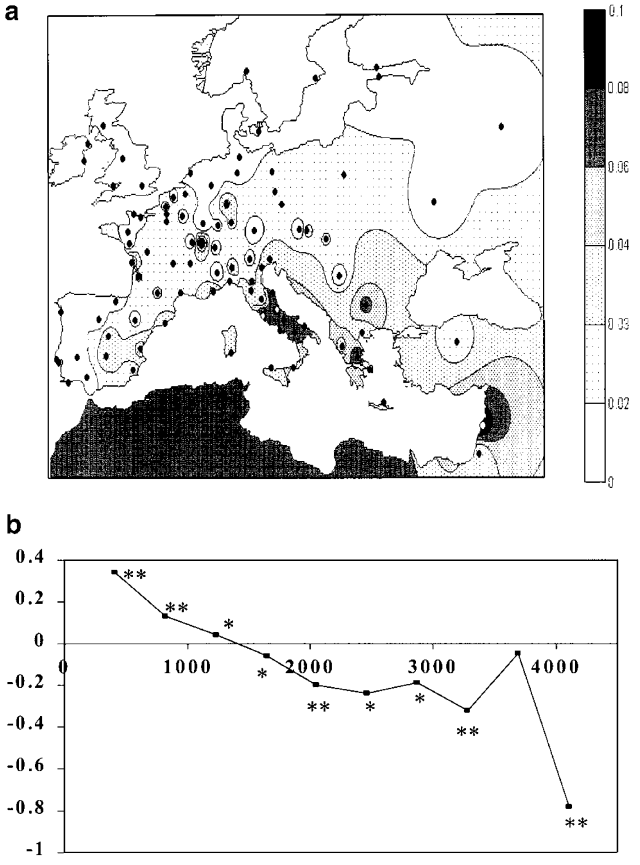

Figure 4 Geographical distribution (**a**) and spatial autocorrelograms (**b**) of the N1303 K mutation in 94 middle Eastern, North African, and European populations. The *X*-axis represents geographic distance between samples; the *Y*-axis represents Moran's index; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$.

for any N_e, doubling it would double θ; for a typical CF allele frequency $f_0=0.02$, doubling it would mean just a 2% reduction in θ. The nonmutated CF chromosomes are the repository from which new mutant alleles arise and contribute to mutant diversity (CF chromosomes already carrying F508del would normally go undetected since mutation testing would stop after finding F508del; although, for counterexamples, see Savov *et al*[39]). Since even a large increase in the incidence has a small impact in the pool of normal chromosomes, variation in incidence is unlikely to have a large effect on mutation diversity. And, as predicted, the correlation between θ and the frequency of alleles carrying CF mutations is small and nonsignificant ($r=0.061$, $P=0.821$). Then, the variation in CF incidence in Europe does not seem to contribute significantly to CF mutation diversity.

An additional factor may prevent incidence from modulating CF mutation diversity. The frequency of the CF allele class is likely to be in balancing selection equilibrium.[3–6] It has been suggested that CF heterozygotes could have a selective advantage against cholera and other diarrhoal diseases, even if Bertranpetit and Calafell[3] showed that cholera by itself is not enough to explain selection on a CF background; that advantage would be given by any mutation that disrupts CFTR function; that is, by any mutation that causes CF. The selective advantage determines the overall frequency of the mutant allele class: for a selective advantge s, the equilibrium frequency of the mutant allele class is $s/(1+s)$,[40] although, for small values of s such as those expected for CF, the mutant frequency can take hundreds of generations to reach the equilibrium value. In this process, selection will pull up the frequency of any mutant allele as long as it confers a selective advantage to the heterozygote, that is, as discussed above, any CF causing mutation. Thus, the initial spectrum of CF causing mutations is that found when the selection process started; selection will increase the frequency of those mutations already present in the different populations, and, in the end, the total frequency of the mutant class may be similar

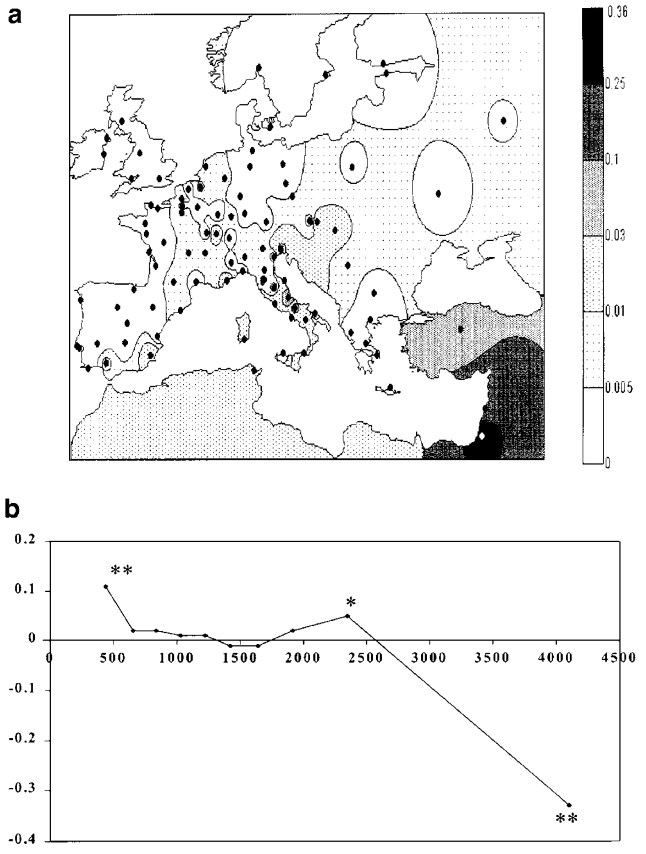

Figure 5 Geographical distribution (**a**) and spatial autocorrelogram (**b**) of the W1282X mutation in 94 middle Eastern, North African, and European populations. The *X*-axis represents geographic distance between samples; the *Y*-axis represents Moran's index; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$.

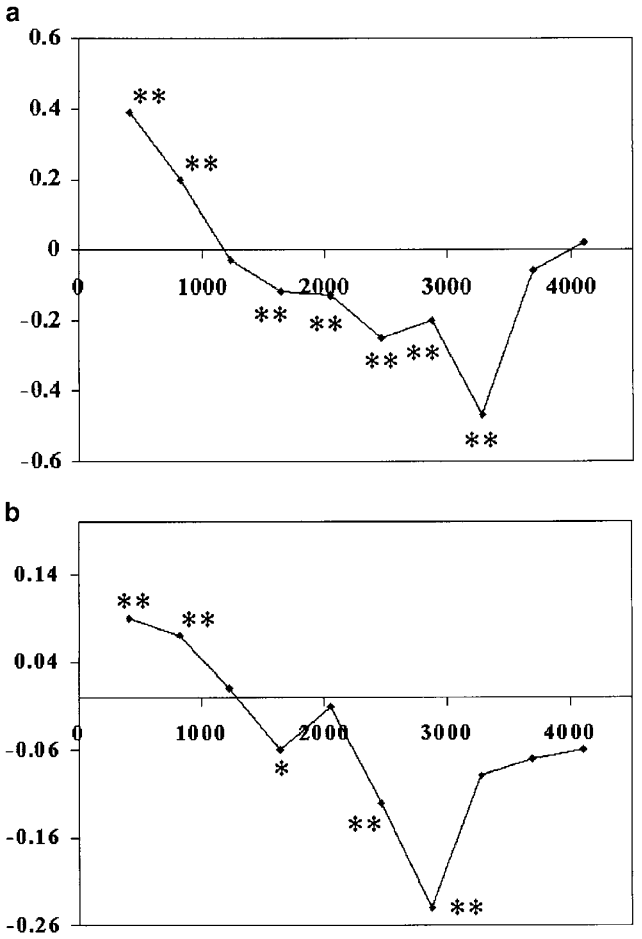

Figure 6 Spatial autocorrelogram of genetic diversity calculated as expected heterozygosity of CF mutations (**a**) and of genetic diversity calculated as Θ of CF mutations (**b**). The *X*-axis represents geographic distance between samples; the *Y*-axis represents Moran's index; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$.

across populations (as long as the selection pressure is similar), but the actual mutations that fill up this frequency determined by selection may be different across populations. Balancing selection will tend to increase the whole CF allele class frequency rather than reshape the spectrum itself, and, thus, it is not expected to have contributed to the pattern of CF mutation diversity in Europe. However, as discussed below, population history, through the action of gene flow and random drift, will then reshape the mutation spectra of the populations.

We turn now to the factor that may have determined to the largest extent of CF mutation diversity: effective population size. Taken at face value, a three-fold increase in θ from Britain (average $\theta=0.63$) to Italy (average $\theta=2.20$) should be the result of a historical effective population size three times larger in Italy than in Britain.[41] Besides the milder, more favorable living conditions, several (pre) historic processes may account for a higher N_e in southern Europe. In the harsher phase of the last glaciation (the so-called Last Glacial Maximum, 18 000 years ago), human populations may have retreated to three glacial refugia in S Europe: Iberia, Italy, and the Balkans, from where they re-expanded when the ice shield melted. It has been suggested that mtDNA[42] (although see the ensuing debate in Simoni *et al*[37] and Torroni *et al*[43]) and Y chromosome[44] diversity patterns bear the traces of the postglacial re-expansions. Later on, the Neolithic expansion that carried the new farming lifestyle to Europe and a 10-fold population increase[45] seems to have expanded faster along the Mediterranean shores,[37] where the population expansion may have started one to two millennia earlier than in northern Europe. All these factors may explain a higher long-term effective population size in the southeast.

In neutral genome regions, genetic distances can capture the shared history patterns among a set of populations. Thus, a correlation between distances based on CF mutation frequencies and distances based on other polymorphisms, once the shared effects of geographic distance are

discounted, implies that the same population history has shaped variation at both loci. We have observed such a significant correlation with mtDNA, although that was not the case with the Y chromosome. Females seem to be more mobile than males,[46,47] up to the point that mtDNA variation is much more homogeneous among European populations than Y-chromosome haplogroup frequencies.[34,37] Although they show clear spatial patterns, the frequency of the main CF mutations is relatively homogeneous across European populations; an analysis of the molecular variance[48] shows that differences among populations explain 3.34% of the variance in CF mutation frequencies; that fraction is 1.13% for mtDNA and 17.07% for Y-chromosome haplogroups in the panel of European populations used for comparison. Thus, the CF mutation landscape seems to correlate better with a homogeneous pattern such as that described by mtDNA than with a more structured pattern such as that found for the Y chromosome. The ubiquity of F508del and the wide areas of distribution of the other major mutations may be the result both of their antiquity and of the aid to dispersion contributed by heterozygote advantage.

The consideration of the whole mutation spectrum of CF has allowed us to interpret the natural history of this disease and its relation to demographic history in a much more comprehensive way than can be obtained with a single-mutation approach. Rather than a specific, *ad hoc* story per mutation, we can provide an overall frame to understand CF. We have shown how the interplay of population history and selection molded the mutation spectrum of CF; it can be expected that the same process has acted on other disease-related genes.

Acknowledgements

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'). O.L. was supported by a predoctoral fellowship from the Ministerio de Ciencia y Tecnología. We thank Teresa Casals for supplying us incidence values.

References

1 Cystic Fibrosis Mutation Data Base. 2002.
2 Lucotte G, Hazout S, De Braekeleer M: Complete map of cystic fibrosis mutation DF508 frequencies in Western Europe and correlation between mutation frequencies and incidence of disease. *Hum Biol* 1995; **67**: 797–803.
3 Bertranpetit J, Calafell F: Genetic and geographical variability in cystic fibrosis: evolutionary considerations. *Ciba Found Symp* 1996; **197**: 97–114.
4 Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ: Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* 1994; **266**: 107–109.
5 Pier GB, Grout M, Zaidi T *et al*: *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* 1998; **393**: 79–82.
6 Pier GB: Role of the cystic fibrosis transmembrane conductance regulator in innate immunity to *Pseudomonas aeruginosa infections*. *Proc Natl Acad Sci USA* 2000; **97**: 8822–8828.
7 Reich DE, Lander ES: On the allelic spectrum of human disease. *Trends Genet* 2001; **17**: 502–510.
8 Bombieri C, Pignatti PF: Cystic fibrosis mutation testing in Italy. *Genet Test* 2001; **5**: 229–233.
9 Scotet V, De Braekeleer M, Roussey M *et al*: Neonatal screening for cystic fibrosis in Brittany, France: assessment of 10 years' experience and impact on prenatal diagnosis. *Lancet* 2000; **356**: 789–794.
10 Asensio D, Cobos N, Seculi J *et al*: Programa de cribaje neonatal para la fibrosis quística en Cataluña. *Invest Clin* 2001; **4** (Suppl 1): 82–83.
11 Cristol P, Des GM, Levy A, Sahuc P: Value of neonatal screening for cystic fibrosis. Evaluation of a neonatal screening program including 34,522 neonates (author's transl). *Semin Hop* 1982; **58**: 499–455.
12 Nazer HM: Early diagnosis of cystic fibrosis in Jordanian children. *J Trop Pediatr* 1992; **38**: 113–115.
13 Cassio A, Bernardi F, Piazzi S *et al*: Neonatal screening for cystic fibrosis by dried blood spot trypsin assay. Results in 47 127 newborn infants from a homogeneous population. *Acta Paediatr Scand* 1984; **73**: 554–558.
14 Edminson PD, Michalsen H, Aagenaes O, Lie SO: Screening for cystic fibrosis among newborns in Norway by measurement of serum/plasma trypsin-like immunoreactivity. Results of a 2 1/2-year pilot project. *Scand J Gastroenterol Suppl* 1988; **143**: 13–18.
15 Roberts G, Stanfield M, Black A, Redmond A: Screening for cystic fibrosis: a four year regional experience. *Arch Dis Child* 1988; **63**: 1438–1443.
16 Cashman SM, Patino A, Delgado MG, Byrne L, Denham B, De Arce M: The Irish cystic fibrosis database. *J Med Genet* 1995; **32**: 972–975.
17 Nielsen OH, Thomsen BL, Green A, Andersen PK, Hauge M, Schiotz PO: Cystic fibrosis in Denmark 1945 to 1985. An analysis of incidence, mortality and influence of centralized treatment on survival. *Acta Paediatr Scand* 1988; **77**: 836–841.
18 de Vries HG, Collee JM, de Walle HE *et al*: Prevalence of delta F508 cystic fibrosis carriers in The Netherlands: logistic regression on sex, age, region of residence and number of offspring. *Hum Genet* 1997; **99**: 74–79.
19 Petrova NV, Ginter EK: Determination of the frequency of the deltaF508 mutation among newborns in the city of Moscow and evaluation of the frequency of cystic fibrosis in the European part of Russia. *Genetika* 1997; **33**: 1326–1328.
20 Brock DJ, Gilfillan A, Holloway S: The incidence of cystic fibrosis in Scotland calculated from heterozygote frequencies. *Clin Genet* 1998; **53**: 47–49.
21 Dodge JA, Morison S, Lewis PA *et al*: Incidence, population, and survival of cystic fibrosis in the UK 1968–1995. UK Cystic Fibrosis Survey Management Committee. *Arch Dis Child* 1997; **77**: 493–496.
22 Kerem E, Kalman YM, Yahav Y *et al*: Highly variable incidence of cystic fibrosis and different mutation distribution among different Jewish ethnic groups in Israel. *Hum Genet* 1995; **96**: 193–197.
23 Frequency of Inherited Disorders Database© (FIDD). 2002.
24 Schneider S, Roessli D, Excoffier L. Arlequin ver. 2000: *A Software for Population Genetics data ANALYSIS*. Switzerland: Genetics and Biometry Laboratory, University of Geneva, 2000.
25 Zouros E: Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics* 1979; **92**: 623–646.
26 Chakraborty R, Weiss KM: Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. *Am J Phys Anthropol* 1991; **86**: 497–506.
27 Le Marechal C, Audrezet MP, Quere I, Raguenes O, Langonne S, Ferec C: Complete and rapid scanning of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by denaturing high-performance liquid chromatography (D-HPLC): major implications for genetic counselling. *Hum Genet* 2001; **108**: 290–298.

28 Sokal RR, Oden NL: Spatial autocorrelation in biology 1. Methodology. *Biol J Linn Soc* 1978; **10**: 199–228.
29 Barbujani G: Geographic patterns: how to identify them and why. *Hum Biol* 2000; **72**: 133–153.
30 Reynolds J, Weir BS, Cockerham CC: Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 1983; **105**: 767–779.
31 Cavalli-Sforza LL, Menozzi P, Piazza A: History and geography of human genes. Princeton: Princeton University Press, 1994.
32 Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; **27**: 209–220.
33 Smouse PE, Long JC, Sokal RR: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool* 1986; **35**: 627–632.
34 Rosser ZH, Zerjal T, Hurles ME *et al*: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 2000; **67**: 1526–1543.
35 Orekhov V, Poltoraus A, Zhivotovsky LA, Spitsyn V, Ivanov P, Yankovsky N: Mitochondrial DNA sequence diversity in Russians. *FEBS Lett* 1999; **445**: 197–201.
36 Richards M, Macaulay V, Hickey E *et al*: Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* 2000; **67**: 1251–1276.
37 Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G: Geographic patterns of mtDNA diversity in Europe. *Am J Hum Genet* 2000; **66**: 262–278.
38 Barbujani G: Autocorrelation of gene frequencies under isolation by distance. *Genetics* 1987; **117**: 777–782.
39 Savov A, Angelicheva D, Balassopoulou A, Jordanova A, Noussia-Arvanitakis S, Kalaydjieva L: Double mutant alleles: are they rare? *Hum Mol Genet* 1995; **4**: 1169–1171.
40 Cavalli-Sforza LL, Bodmer WF: The genetics of human populations. San Francisco: Freeman, 1971.
41 Przeworski M, Wall JD: Why is there so little intragenic linkage disequilibrium in humans? *Genet Res* 2001; **77**: 143–151.
42 Torroni A, Bandelt HJ, D'Urbano L *et al*: mtDNA analysis reveals a major late Pleistocene population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 1998; **62**: 1137–1152.
43 Torroni A, Richards M, Macaulay V *et al*: mtDNA haplogroups and frequency patterns in Europe. *Am J Hum Genet* 2000; **66**: 1173–1177.
44 Semino O, Passarino G, Oefner PJ *et al*: The genetic legacy of Paleolithic *Homo sapiens* sapiens in extant Europeans: a Y chromosome perpective. *Science* 2000; **290**: 1155–1159.
45 Ammerman AJ, Cavalli-Sforza LL: The Neolithic transition and the genetics of populations in Europe. Princeton: Princeton University Press, 1984.
46 Seielstad MT, Minch E, Cavalli-Sforza LL: Genetic evidence for a higher female migration rate in humans. *Nat Genet* 1998; **20**: 278–280.
47 Perez-Lezaun A, Calafell F, Comas D *et al*: Sex-specific migration patterns in Central Asian populations revealed by the analysis of Y-chromosome STRs and mtDNA. *Am J Hum Genet* 1999; **65**: 208–219.
48 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.
49 Stuhrmann M, Dork T, Fruhwirth M *et al*: Detection of 100% of the CFTR mutations in 63 CF families from Tyrol. *Clin Genet* 1997; **52**: 240–246.
50 Messiaen L, Van Loon C, Rossau R *et al*: Analysis of 22 cystic fibrosis mutations in 62 patients from the Flanders, Belgium, reveals a high prevalence of Nordic mutation 394delTT. *Hum Mutat* 1997; **10**: 236–238.
51 Angelicheva D, Calafell F, Savov A *et al*: Cystic fibrosis mutations and associated haplotypes in Bulgaria – a comparative population genetic study. *Hum Genet* 1997; **99**: 513–520.
52 Hughes DJ, Hill AJ, Macek Jr M, Redmond AO, Nevin NC, Graham CA: Mutation characterization of CFTR gene in 206 Northern Irish CF families: thirty mutations, including two novel, account for approximately 94% of CF chromosomes. *Hum Mutat* 1996; **8**: 340–347.
53 Desgeorges M, Megarbane A, Guittard C *et al*: Cystic fibrosis in Lebanon: distribution of CFTR mutations among Arab communities. *Hum Genet* 1997; **100**: 279–283.
54 Telleria JJ, Alonso MJ, Calvo C, Alonso M, Blanco A: Spectrum of CFTR mutations in the Middle North of Spain and identification of a novel mutation (1341G–>A). Mutation in brief no. 252. Online. *Hum Mutat* 1999; **14**: 89.
55 Messaoud T, Verlingue C, Denamur E *et al*: Distribution of CFTR mutations in cystic fibrosis patients of Tunisian origin: identification of two novel mutations. *Eur J Hum Genet* 1996; **4**: 20–24.
56 Kilinc MO, Ninis VN, Dagli E *et al*: Highest heterogeneity for cystic fibrosis: 36 mutations account for 75% of all CF chromosomes in Turkish patients. *Am J Med Genet* 2002; **113**: 250–257.
57 Livshits LA, Kravchenko SA: Cystic fibrosis in Ukraine: age, origin and tracing of the delta F508 mutation. *Gene Geogr* 1996; **10**: 219–227.
58 Estivill X, Bancells C, Ramos C: Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. *Hum Mutat* 1997; **10**: 135–154.