RESULTATS

CAPÍTOL I

Allele frequencies in a worldwide survey of a CA repeat in the first intron of the CFTR gene

Eva Mateu, Francesc Calafell, Batsheva Bonné-Tamir,
Judith R. Kidd, Teresa Casals, Kenneth K. Kidd i
Jaume Bertranpetit

Human Heredity (1999) 49: 15-20

Allele Frequencies in a Worldwide Survey of a CA Repeat in the First Intron of the CFTR Gene

Eva Mateu[a], Francesc Calafell[a], Batsheva Bonné-Tamir[b], Judith R. Kidd[c], Teresa Casals[d], Kenneth K. Kidd[c] and Jaume Bertranpetit[a]

[a] Unitat d'Antropologia. Facultat de Biologia. Universitat de Barcelona. Catalonia. Spain
[b] Sackler Faculty of Medicine. Tel Aviv University. Israel.
[c] Department of Genetics. Yale University School of Medicine. New haven. USA.
[d] Dep. Genètica Molecular. Institut de Recerca Oncològica. Barcelona. Catalonia. Spain.

Short title: CFTR CA Repeat Allele Frequencies

Key words: DNA polymorphism, dinucleotide STR, allele frequency, CFTR gene

Author to whom correspondence should be sent:

Jaume Bertranpetit
Unitat d'Antropologia
Facultat de Biologia
Diagonal 645
08028 Barcelona
Catalonia, Spain
Phone: +34-93-402 14 61
Fax: +34-93-411 08 87
e-mail: jaumeb@porthos.bio.ub.es

Abstract

A dinucleotide CA repeat within intron 1 of the CFTR gene was recently identified. We have determined the allele frequencies of this polymorphism in samples from 18 populations covering all major geographical areas, with a total of 1816 chromosomes. When considering allele distributions, African populations presented a wider range of alleles than other geographic areas and also presented higher expected heterozygosities. Analysis of Molecular Variance (AMOVA) showed that 8.04% of the genetic variance in this locus could be attributed to differences among populations. We concluded that the polymorphism in the CA repeat in intron 1 of the CFTR gene is highly informative in populations from all geographical regions of the world, and, thus, can be applied to family studies of unknown CF-causing mutations, and can provide valuable information for genetic counseling. Moreover its analysis should be included in haplotypic analysis of known CF mutations.

Introduction

Cystic fibrosis (CF) is the most common severe recessive disorder in patients of European descent, affecting 1 in 2,000 to 4,000 individuals. Since the isolation of the cystic fibrosis transmembrane conductance regulator (CFTR) gene in 1989, more than 700 CF mutations have been identified [1-4]. Of all CF mutations, a deletion of three base pairs at codon 508 (the F508 mutation) is the most frequent, accounting for approximately 67% of global CF chromosomes. Among all other mutations, most are rare and often confined to one or a few populations. The internal diversity of CF-causing mutations can vary widely across populations. Some populations (Bretons, Northern French, Belgian, Welsh, Hutterite or Ashkenazi Jews) present an apparently high homogeneity, and, in those, the mutation causing the disease has been identified in a large proportion of CF chromosomes, although the prevalent mutations found are not the same for each group. In contrast, other population groups (Southern French, Spanish) show a high heterogeneity for CF mutations [5], and a larger proportion of CF mutations remains to be identified. Several highly polymorphic short tandem repeats (STRs; also known as microsatellites) and biallelic markers (SNPs) have been described within the CFTR gene. Both types of markers present interesting features for segregation analysis and genetic counseling and can be used to trace the origin and evolution of the different CF mutations [6,7]. The analysis of rare CF mutations has been facilitated by the use of microsatellite haplotypes, as strong linkage disequilibrium exists between mutations and microsatellite haplotypes.

Different techniques are available to detect specific CF mutations. In some populations with an average detection rate of 85%, only 72% of the couples who seek genetic counseling will receive fully informative answers; the remaining cases will have to rely on closely linked DNA markers to obtain a more complete diagnosis. The haplotyping of several microsatellites within the CFTR gene [8], offers an alternative strategy to RFLPs in tracking the parental chromosomes in a family, as STR polymorphisms are often found to be excellent markers for family studies [9]. Chehab et al., (1991) [10] reported a polymorphic GATT tetranucleotide repeat in the 5' flanking region of exon 6b (intron 6a); Morral et al., (1991) [11] noted a polymorphic CA/GT dinucleotide repeat in intron 8; Zielenski et al., (1991b) [9] reported a cluster of two polymorphic dinucleotide repeats (a TA and a CA repeat) in intron 17b, and Moulin et al., (1997) [12] identified a CA repeat within intron 1 of the CFTR gene. The aim of this paper is to present a global survey of the variation for the newest marker (intron 1 CA repeat), which is likely to be of high interest in genetic diagnosis and counseling and also in population genetics. This marker has been described very recently and the allelic frequencies in population samples other than a sample containing Scots and Greek Cypriots [12] are, to the best of our knowledge, unknown.

Samples and Methods

The intron 1 CA repeat in the CFTR gene was analyzed in 913 autochthonous individuals from 18 populations representing all major world geographic areas (Figure 1). Sub-Saharan African populations comprised Mbuti Pygmies (former Zaire), Biaka Pygmies (Central African Republic), and Tanzanians. North Africans were represented by the Sahraui from the former Western Sahara. Samples from the Middle East were the Druze (Northern Israel) and Yemenite Jews; the European populations comprised Basques, Catalans, Finns, Russians, and Adygei (northern Caucasus). Asian samples included the Kazakhs (Central Asia), Yakut (Siberia), Han Chinese, and Japanese; the Pacific was represented by the Nasioi from Bougainville in Melanesia. The Native Americans sampled were the Maya (from Yucatan) and the Surui (NW Amazonia). Sample sizes ranged from 46 (Nasioi) to 212 (Basque) chromosomes, with a median of 97.

DNA for five populations (Basques, Catalans, Tanzanian, Kazakhs and Sahraui) was extracted from fresh blood of blood donors. DNA samples for the other populations were obtained from lymphoblastoid cell lines in the laboratory of J.R. and K.K. Kidd (Yale University, USA). These populations are genetically well characterized and many other markers have been previously typed in them [13-16, among many others].

PCR amplifications were performed using 50 ng of genomic DNA in a final volume of 10 l. The amplifications were carried out in a Perkin Elmer 9600 thermal cycler. The CA repeat was amplified with flanking primers TSR12 (fluorescently labeled) and TSR13 [12], with 30 PCR cycles, denaturing at 95 for 30sec, annealing at 50 for 30sec and extending at 65 for 45sec. PCR products 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 377™ sequencer. GeneScan 672™ was used to collect the data, track lanes, and measure fragment sizes. The number of CA repeats was estimated by sequencing three homozygous individuals of different fragment sizes. The sequencing reaction was performed with flanking primer TSR13 with the DNA Sequencing Kit™ (Perkin Elmer) according to manufacturerÆs specifications. The product of the sequence reaction was run in an ABI 377™ sequencer.

Allele frequencies were estimated by direct gene counting. Expected heterozygosity was estimated as $1- p_i^2$, where p_i is the frequency of the *i*th allele in the locus. Hardy-Weinberg equilibrium was tested by applying a hidden Markov chain with 100,000 steps. Overall genetic heterogeneity among populations was tested through Analysis of Molecular Variance (AMOVA) [17]. An exact pairwise test of population differentiation [18] was carried out. All these analyses were performed with the Arlequin package [19].

Results and Discussion

A CA repeat in intron 1 of the CFTR gene was typed in 913 individuals from 18 world populations. Tables 1 and 2 show the allele frequencies and heterozygosities for the CA repeat, distributed by major geographic areas. Populations are named as in Figure 1 and alleles are named by the estimated number of dinucleotide repeats of the core unit, according to the sequencing results of three homozygous individuals of 261, 273 and 275bp, which corresponded to 15, 21 and 22 repeats respectively. These results do not match those reported by Moulin et al., (1997) [12] where the fragment of 26 CA repeats corresponded to 293bp, instead of 283bp. These authors used the same primers with 5 extra bases at the 5' end of each primer (personal communication). The range found is from 12 to 28 repeats, except for the 13-repeat allele, which was not found. Moulin et al., (1997) [12] described alleles from 17 to 26 repeats in a mixed European sample, and thus we report six new alleles. The number of different alleles varies from one population to another. We have found 15 different alleles in African populations, 13 in Europeans, nine in Asian populations and only five in Native Americans. Since the sample size for Europeans was larger that for other geographic areas (606 chromosomes as opposed to 272 for Sub-Saharan Africans, 394 for East Asians, and 192 for Native Americans), the relative number of different alleles in Europeans may be slightly overestimated and in Native Americans slightly underestimated. Then, this data set would be yet another example of the higher genetic diversity within African populations [15, 20-21, among many others].

When considering allele distributions, African populations presented a bimodal distribution, showing peaks at 15-18 and 21-23 repeats. In Europeans, East Asians and Native Americans we found only this last group of alleles at high frequencies, restricted to alleles with 22-23 repeats in Native Americans, where the 21-repeat allele might have become very rare or been lost by drift during the colonization of the continent or at least in the studied populations. Drift may also explain the high frequencies of the shorter mode of alleles (16-18 repeats) in the Melanesian Nasioi. The small effective size of this population may explain the upward drift in frequency of the short alleles, which may have been brought in from mainland Asia (where the 16-repeat allele is found at low frequencies); it is unlikely that short alleles have been regenerated by mutation from the large alleles as mutation has had a small effect compared to drift in the making of allele distributions for human populations [16].

Expected heterozygosities vary from 0.431 in the Surui population to 0.918 in the Sahraui (Tables 1 and 2). Sub-Saharan African populations show higher expected heterozygosities (0.842-0.900) than Europeans (0.708-0.758), Middle Easterners (0.639-0.710), East Asians (0.677-0.741), and Native Americans (0.431-0.584). This pattern of decreasing expected heterozygosity was also observed in a set of 45 CA repeat markers by Calafell et al., (1998) [15]. The fact that the highest heterozygosity was found in the Sahraui

can be interpreted as the result of admixture in this population between North Africans, which are closely related to Europeans and Middle Easterners [22] and Subsaharan Africans. A similar effect may explain the relatively high heterozygosity of the Kazakh, which may result from an admixture between East Asians and Europeans [unpublished data]. With the sole exception of the Surui, the expected heterozygosity was always above 0.5, which makes this CA repeat more informative for family studies than any biallelic marker.

All populations were found to conform to Hardy-Weinberg expectations except for Basques (p=0.006), Catalans (p=0.023), Mbuti (p=0.041), Russians (p=0.028) and Surui (p=0.004); except for the latter population, all others showed an excess of homozygotes. However, when the Bonferroni correction for multiple tests was applied, no statistically significant test remained.

As previously discussed, allele frequencies varied across populations. Analysis of Molecular Variance (AMOVA) showed that 8.04% of the genetic variance in this locus could be attributed to differences among populations. This fraction is equivalent to F_{ST}=0.0804, and is significantly different from zero (p<0.0001). A fraction of 6.47% (p<0.0001) of the genetic variance could be apportioned to differences among geographic areas as defined in Tables I and II, and also a significant fraction of genetic variance (1.57%, p<0.0001) was found among the populations belonging to the same geographic area.

An exact test of population differentiation [18] was applied. Most pairwise tests were statistically significant (p<0.0001), and only a few comparisons, involving always a pair of populations from the same continent, were not statistically significantly heterogeneous.

We have shown that the polymorphism in the CA repeat in intron 1 of the CFTR gene is highly informative in populations from most geographical regions of the world, and, thus, can be applied to family studies of unknown CF-causing mutations, and can provide valuable information for genetic counseling, especially in populations, such as South America and the Mediterranean, where a relatively large proportion of CF-causing mutations remains unknown. In the usual clinical practice, the three microsatellite loci routinely typed (IVS8CA, IVS17BTA and IVS17BCA) may provide an informativeness of 99.5%. The additional typing of the intron 1 CA would have mostly a confirmation value, and, since it provides an additional point in the CFTR gene map, it would allow the detection of recombination events, although they are rare in the CFTR gene. However, in some special situations, additional information is needed and the typing of the intron 1 CA can be particularly useful. For instance, if a false paternity is suspected, the confirmation value of the intron 1 CA may become crucial. In prenatal testing, the additional information would facilitate the detection of maternal contamination of the specimen. Finally, the characterization of microsatellite haplotypes has proved useful in the study of the origin and history of CF-causing mutations [6]. The inclusion of one additional marker to those haplotypes may improve the resolution of

the method and may give additional insights in the understanding of the origin and evolution of cystic fibrosis.

Acknowledgments

This research was supported by the Direcci n General de Investigaci n Cient fico T cnica (Spain) grant PB95-0267-CO2-01, by the Human Capital and Mobility contracts to J.B. (network ERCHRXCT92-0032 and ERB-CHRX-CT920090), by Direcci General de Recerca, Generalitat de Catalunya (1995SGR00205 and 1996SGR00041), by Institut dÆEstudis Catalans, and by U.S.N.S.F. grant SBR9632509 to J.R.K. The work was also possible thanks to a fellowship to E.M. (University of Barcelona).

Samples from Tanzania were kindly supplied by Dr. Clara Men ndez from Unitat dÆEpidemiologia i Bioestad stica (Hospital Cl nic, Barcelona). We also thank M.D. Ramos for technical assistance and X. Estivill for useful comments, both from Departament Gen tica Molecular, Institut de Recerca Oncol gica (Barcelona); and the staff of the Servei de Seq enciaci , Serveis Cient fico T cnics, University of Barcelona, for their invaluable technical support.

References

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

2 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: Identification of the cystic fibrosis gene:cloning and characterization of complementary DNA. Science 1989;245:1066-1073.

3 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: Identification of the cystic fibrosis gene; chromosome walking and jumping. Science 1989;245:1059-1065.

4 http://www.genet.sickkids.on.ca

5 Estivill X, Morral N: Evolution of cystic fibrosis alleles; in Dodge JA, Brock DJH, Widdicombe JH (eds): Cystic fibrosis-Current topics., John Wiley and Sons Ltd, 1996, pp 141-164.

6 Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gim nez J, Reis A, Varon-Mateeva R, Macek Jr M, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo MP, Garnerone S, Restagno G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwarz M, Dallapiccola B, Novelli G, F rec C, de Arce M, Nemeti M, Kere J, Anvret M, Dahl N, Kadasi L: The origin of the major cystic fibrosis mutation (DF508) in European populations. Nature Genet 1994;7:169-175.

7 Bertranpetit J, Calafell F: Genetic and geographical variability in cystic fibrosis; evolutionary considerations; in John Wiley & Sons (ed): Variation in the human genome. Chichester, Ciba Foundation Symposium 197, 1996, pp 97-118.

8 Zielenski J, Markiewicz D, Rininsland F, Rommens J, Tsui LC: A cluster of highly polymorphic dinucleotide repeats in intron 17b of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Am J Hum Genet 1991a;49:1256-1262.

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

10 Chehab FF, Johnson J, Louie E, Goossens M, Kawasaki E, Erlich H: A dimorphic 4-bp repeat in the cystic fibrosis gene is in absolute linkage disequilibrium with the DF508 mutation; implications for prenatal diagnosis and mutation origin. Am J Hum Genet 1991;48:223-226.

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

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

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

14 Castiglione CM, Deinard AS, Speed WC, Sirugo G, Rosenbaum HC, Zhang Y, Grandy DK, Grigorenko EL, Bonn -Tamir B, Pakstis AJ, Kidd JR, Kidd KK: Evolution of haplotypes at the DRD2 locus. Am J Hum Genet 1995;57:1445-1456.

15 Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK: Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 1998;6:38-49.

16 P rez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J: Microsatellite variation and the differentiation of modern humans. Hum Genet 1997;99:1-7.

17 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.

18 Raymond M, Rousset F: An exact test for population differentiation. Evolution 1995;49:1280-1283.

19 Schneider S, Kueffer JM, Roessli D, Excoffier L: Arlequin (ver. 1.0): a software environment for the analysis of population genetics data. Geneva, Genetics and Biometry Lab, 1996.

20 Armour JA, Anttinen T, May CA, Vega EE, Sajantila A, Kidd

JR, Kidd KK, Bertranpetit J, P bo S, Jeffreys AJ: Minisatellite diversity supports a recent African origin for modern humans. Nature Genet 1996;13:154-160.

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

22 Bosch E, Calafell F, P rez-Lezaun A, Comas D, Mateu E, Bertranpetit J: Population history of North Africa; evidence from classical genetic markers. Hum Biol 1997;69:295-311.

Number of repeats	TAN 2n=80	BIA 2n=126	MBU 2n=66	Sub-Saharan Africa 2n=272	NorthAfrica (SAH) 2n=118	DRU 2n=110	YEM 2n=88	Middle East 2n=198	MAY 2n=106	SUR 2n=86	America 2n=192	Pacific (NAS) 2n=46
12	1.3	2.4	0	1.2	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	4.8	0	1.6	0	0	0	0	0	0	0	0
15	0	1.6	24.2	8.6	1.7	0	0	0	0	0	0	0
16	12.5	6.3	0	6.3	2.5	0	0	0	0.9	0	0.4	13
17	3.8	11.9	16.7	10.8	8.5	0	0	0	0	0	0	28.3
18	15	7.9	4.5	9.1	7.6	0.9	1.1	1	0	0	0	4.3
19	1.3	4.8	0	2	3.4	0	2.3	1.1	0	0	0	0
20	5	7.9	9.1	7.3	4.2	0	12.5	6.2	0.9	0	0.4	0
21	15	15.9	6.1	12.3	22	23.6	19.3	21.5	0	0	0	26.1
22	20	13.5	21.2	18.3	29.7	53.6	46.6	50.3	35.8	31.4	33.6	10.9
23	15	12.7	4.5	10.7	11	8.2	13.6	10.9	52.8	68.6	60.7	17.4
24	8.8	2.4	9.1	6.8	6.8	10	0	5	9.4	0	4.7	0
25	1.3	3.2	4.5	3	2.5	0	2.3	1.1	0	0	0	0
26	0	4.8	0	1.6	0	3.6	2.3	2.9	0	0	0	0
27	1.3	0	0	0.4	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
Het*	0.864	0.9	0.842		0.918	0.639	0.71		0.584	0.431		0.791

* Expected heterozygosity

Number of repeats	BAS	CAT	FIN	RUS	ADY	Europe	KAZ	YAK	CHI	JPN	Asia
	2n=212	2n=134	2n=68	2n=88	2n=104	2n=606	2n=76	2n=100	2n=124	2n=94	2n=394
12	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	1.1	0	0.2	0	1	0.8	0	0.4
17	0.5	3.7	0	0	0	0.8	0	0	0	0	0
18	2.4	1.5	7.4	1.1	0	2.5	2.6	0	0	0	0.6
19	0.9	0	0	0	1.9	0.6	0	0	0	0	0
20	2.4	4.5	10.3	6.8	1	5	5.3	6	4.8	7.4	5.9
21	25.9	23.9	22.1	34.1	26	26.4	14.5	12	3.2	6.4	9
22	40.6	40.3	39.7	27.3	43.3	38.2	39.5	44	43.6	50.1	44.4
23	13.7	9.7	11.8	20.5	18.3	14.8	26.3	8	34.7	19.1	22
24	9.4	9	1.5	3.4	0	4.7	1.3	18	8.9	8.5	9.2
25	1.4	3.7	0	1.1	2.9	1.8	9.2	11	4	7.4	7.9
26	1.9	3.7	7.4	3.4	4.8	4.2	1.3	0	0	1.1	0.6
27	0.5	0	0	0	1	0.3	0	0	0	0	0
28	0.5	0	0	1.1	1	0.5	0	0	0	0	0
Het*	0.739	0.757	0.758	0.76	0.708		0.741	0.737	0.677	0.691	

*Expected heterozygosity

Table 1. Percent allele frequencies in African, Middle Eastern, America and Pacific populations

Table 2. Percent allele frequencies in European and Asian populations

Figure 1. Geographic location of the populations sampled. Abbreviations: ADY, Adygei; BAS, Basques; BIA, Biaka Pygmies; CAT, Catalans; CHI, Han Chinese; DRU, Druze; FIN, Finns; JPN, Japanese; KAZ, Kazakhs; MAY, Maya; MBU, Mbuti Pygmies; NAS, Nasioi; RUS, Russians; SAH, Sahraui; SUR, Surui; TAN, Tanzanian; YAK, Yakut; YEM, Yemenite.


FIN
YAK
RUS
BAS
CAT
ADY
KAZ
DRU
JPN
SAH
CHI
YEM
MAY
MBU
BIA
TAN
SUR
NA

CAPÍTOL II

Worldwide genetic analysis of the CFTR region

Eva Mateu, Francesc Calafell, Oscar Lao,
Batsheva Bonné-Tamir, Judith R. Kidd, Andrew Pakstis,
Kenneth K. Kidd i Jaume Bertranpetit

American Journal of Human Genetics (2001) 68: 103-117

Am. J. Hum. Genet. 68:103–117, 2001

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

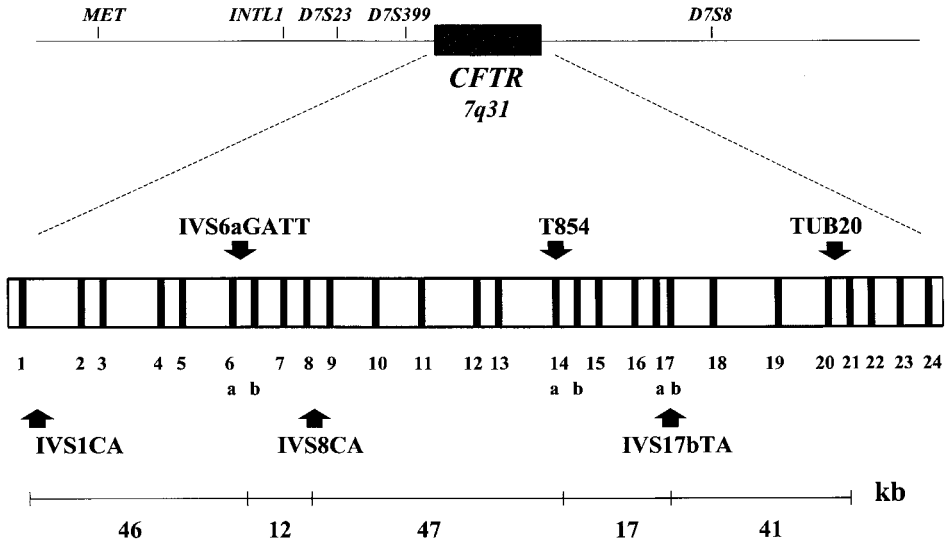


Figure 1 CFTR gene with all six polymorphic genetic markers studied (IVS1CA, IVS6aGATT, IVS8CA, T854, IVS17bTA, TUB20), showing the physical distances (in kb) between them. Gene exons are denoted by numbers (1–24).

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 fraction of extra 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 $\geqslant .05$ 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

Population (2N)	Frequency of Residual Class	Frequency of Haplotype[a]																				
		15 6 24 1 19 2	16 6 23 2 23 2	16 7 17 2 27 2	17 6 19 2 22 2	17 6 20 2 19 2	17 6 22 2 7 1	18 6 23 2 22 2	18 7 16 2 7 2	21 6 23 2 20 2	21 7 14 2 38 2	21 7 16 2 31 2	21 7 16 2 7 1	21 7 17 1 20 2	21 7 17 2 7 1	21 7 17 2 7 2	21 7 17 2 7 2	22 6 17 2 31 2	22 6 17 2 37 2	22 6 22 2 19 2	22 7 16 1 7 1	22 7 16 1 7 2	22 7 16 1 29 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

Population (2N)	Frequency of Haplotype																					
	22 7 16 1 30 2	22 7 16 1 31 2	22 7 16 1 32 2	22 7 16 1 33 2	22 7 16 1 44 2	22 7 16 2 7 1	22 7 21 1 23 2	23 6 17 2 15 2	23 6 17 2 20 2	23 6 17 2 23 2	23 6 17 2 31 2	23 6 17 2 32 2	23 6 17 2 33 2	23 6 17 2 35 2	23 6 17 2 37 2	23 7 16 1 30 2	23 7 16 1 32 2	23 7 16 2 7 2	24 7 16 1 30 2	24 7 16 1 31 2	24 7 17 1 31 2	25 6 17 2 31 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. $2N$ = 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 $\sim 10^{-9}$ (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 (2*N*)	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 $\xi \sim 0$. 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

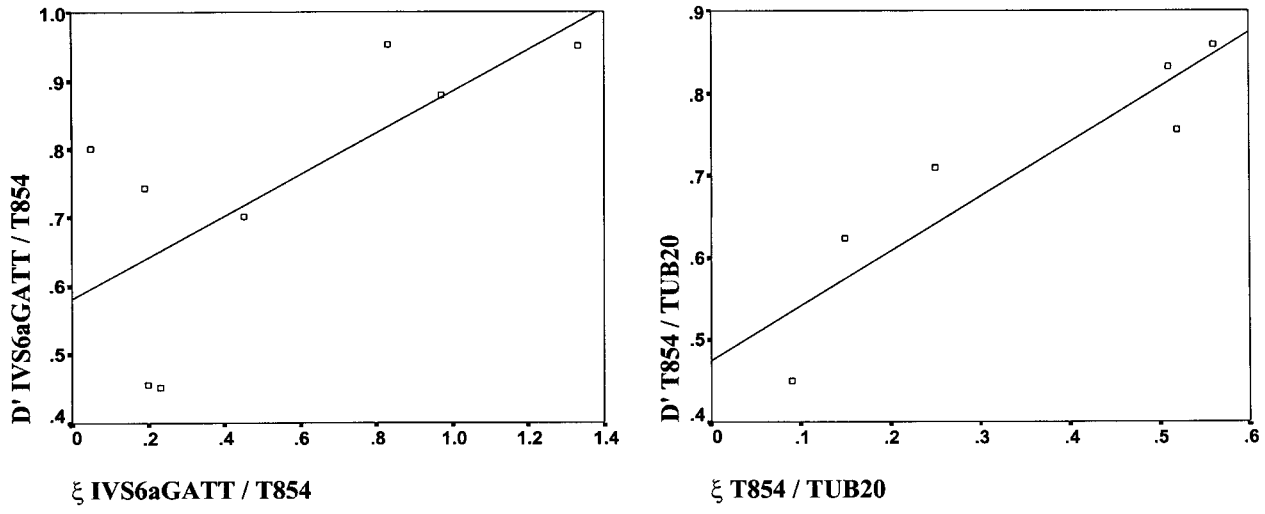


Figure 2 LD D′/ξ values correlation between loci IVS6aGATT / T854 and T854 / TUB20 (significance level for D′, $P < .05$)

Table 6

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

Variable and Population	Loci (Distance in kb)[a]														
	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).

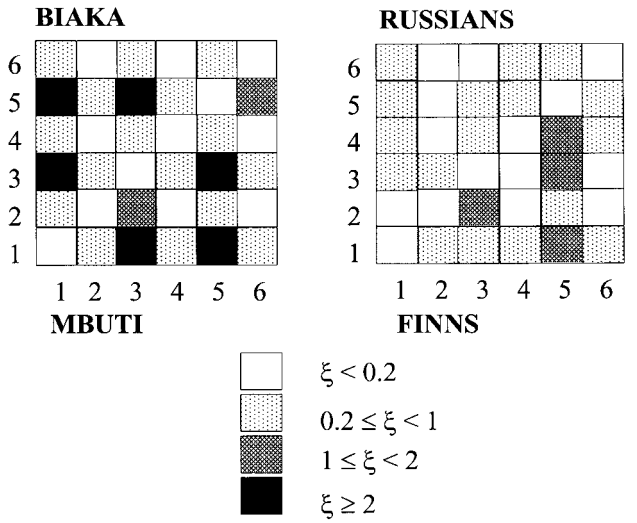


Figure 3 ξ values for four populations between all six loci (1 = IVS1CA, 2 = IVS6aGATT, 3 = IVS8CA, 4 = T854, 5 = IVS17bTA, and 6 = TUB20). Each square is a graphical representation of ξ levels for pairs of loci in two populations; each population is represented either above or below the diagonal (gray 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 ($\xi = 0.23$) and the Basques ($\xi = 0.20$), 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 contradic-

tion 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 se-

quences 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 disequi-

librium 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

CAPÍTOL II

APÈNDIX

Les freqüències al·lèliques, per *locus* i població, així com les freqüències haplotípiques estimades, pels polimorfismes del gen CFTR han estat dipositades a la base de dades ALFRED (ALlele FREquency Database, httt://info.med.yale.edu/genetics/kkidd). A continuació presentem, com a apèndix del Capítol II, les freqüències al·lèliques per *locus* i població.

1. IVS1CA

REPEATS		12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
POPULATION	**2N**																	
Biaka	124	.024	.008	.048	.008	.065	.121	.073	.040	.081	.169	.129	.129	.024	.032	.048	0	0
Mbuti	66	0	0	0	.242	0	.182	.045	0	.091	.076	.197	.030	.091	.045	0	0	0
Tanzanian	72	.014	0	0	0	.139	.014	.167	.014	.042	.125	.222	.153	.083	.014	0	.014	0
Saharawi	110	0	0	0	.018	.027	.073	.082	.027	.045	.227	.309	.109	.055	.027	0	0	0
Yemenites	80	0	0	0	0	0	0	.013	.025	.138	.200	.475	.100	0	.025	.025	0	0
Druze	126	0	0	0	0	0	0	.008	0	.008	.214	.540	.079	.095	0	.048	0	.008
Adygei	98	0	0	0	0	0	0	0	.020	.010	.255	.449	.173	0	.031	.051	0	.010
Russians	64	0	0	0	0	.016	0	.016	0	.063	.313	.281	.219	.031	0	.047	0	.016
Finns	66	0	0	0	0	0	0	.076	0	.106	.227	.394	.106	.015	0	.076	0	0
Catalans	166	0	0	0	0	0	.030	.018	.006	.036	.241	.410	.084	.090	.036	.048	0	0
Basques	216	0	0	0	0	0	.005	.023	0	.023	.264	.407	.144	.093	.014	.019	.005	.005
Kazakhs	66	0	0	0	0	0	0	.030	0	.030	.136	.409	.273	.015	.091	.015	0	0
Chinese	86	0	0	0	0	.012	0	0	0	.047	.047	.430	.372	.047	.047	0	0	0
Japanese	88	0	0	0	0	0	0	0	0	.080	.068	.523	.148	.091	.080	.011	0	0
Nasioi	46	0	0	0	0	.130	.283	.043	0	0	.261	.109	.174	0	0	0	0	0
Yakut	86	0	0	0	0	.012	0	0	0	.070	.128	.407	.093	.174	.116	0	0	0
Maya	98	0	0	0	0	.010	0	0	0	.010	0	.347	.531	.102	0	0	0	0
Surui	84	0	0	0	0	0	0	0	0	0	0	.310	.690	0	0	0	0	0

2. IVS6aGATT

REPEATS		4	5	6	7	8
POPULATION	**2N**					
Biaka	124	0	0	.411	.589	0
Mbuti	66	0	0	.515	.485	0
Tanzanian	72	0	0	.222	.778	0
Saharawi	110	0	0	.236	.764	0
Yemenites	80	0	0	.200	.800	0
Druze	126	0	0	.206	.794	0
Adygei	98	.010	0	.153	.837	0
Russians	64	0	0	.203	.797	0
Finns	66	0	0	.273	.727	0
Catalans	166	0	0	.241	.759	0
Basques	216	0	.005	.236	.755	.005
Kazakhs	66	0	0	.379	.621	0
Chinese	86	0	0	.477	.523	0
Japanese	88	0	0	.284	.716	0
Nasioi	46	0	0	.761	.239	0
Yakut	86	0	0	.267	.733	0
Maya	98	0	0	.633	.367	0
Surui	84	0	0	.798	.202	0

3. IVS8CA

REPEATS		14	15	16	17	18	19	20	21	22	23	24	25
POPULATION	**2N**												
Biaka	124	.129	.008	.234	.250	.032	.056	.016	.137	.105	.016	.016	0
Mbuti	66	0	.045	.273	.197	0	.015	.136	.015	.167	0	.121	.030
Tanzanian	72	.014	0	.583	.222	0	.014	0	.083	.014	.056	.014	0
Saharawi	110	0	.009	.791	.145	0	0	0	.009	0	.045	0	0
Yemenites	80	0	.038	.700	.150	0	.013	0	.025	.013	.038	0	.025
Druze	126	.056	0	.762	.103	0	0	.008	.016	.008	.048	0	0
Adygei	98	.020	0	.755	.143	.020	0	0	.020	.010	.031	0	0
Russians	64	0	.016	.672	.203	.031	0	0	0	0	.047	.031	0
Finns	66	0	.015	.636	.197	0	0	0	.015	.015	.121	0	0
Catalans	166	.006	0	.747	.114	.012	.006	.006	.018	.012	.060	.018	0
Basques	216	0	0	.718	.162	.023	0	0	0	0	.093	.005	0
Kazakhs	66	.015	.015	.530	.348	.015	0	0	.015	0	.045	.015	0
Chinese	86	.023	0	.442	.523	.012	0	0	0	0	0	0	0
Japanese	88	0	0	.659	.284	.034	0	0	.023	0	0	0	0
Nasioi	46	0	0	.217	.087	0	.152	.022	.043	.043	.413	0	.022
Yakut	86	0	.058	.547	.349	0	0	0	0	0	.023	.023	0
Maya	98	0	0	.408	.551	.041	0	0	0	0	0	0	0
Surui	84	0	0	.167	.833	0	0	0	0	0	0	0	0

4. T854

	ALLELE	1	2
POPULATION	**2N**		
Biaka	122	.336	.664
Mbuti	66	.439	.561
Tanzanian	64	.359	.641
Saharawi	110	.491	.509
Yemenites	80	.788	.213
Druze	126	.778	.222
Adygei	98	.704	.296
Russians	64	.500	.500
Finns	62	.710	.290
Catalans	166	.675	.325
Basques	216	.708	.292
Kazakhs	60	.583	.417
Chinese	86	.512	.488
Japanese	88	.727	.273
Nasioi	46	.478	.522
Yakut	86	.779	.221
Maya	96	.354	.646
Surui	84	.155	.845

5. IVS17bTA

REPEATS	7	8	15	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
POPULATION	**2N**																	
Biaka	124	.371	0	0	0	0	.056	.032	.089	.016	.008	.008	0	0	.089	0	0	.008	.008
Mbuti	66	.136	0	0	0	.015	.379	.197	.015	0	0	0	0	0	.015	.045	0	.061	.045
Tanzanian	72	.167	0	0	0	0	.069	.028	.042	0	.056	0	0	0	.097	.111	.028	.028	.125
Saharawi	110	.400	.009	0	0	0	.018	.009	.100	0	0	0	.027	0	0	.018	.045	.173	.073
Yemenites	80	.213	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.175	.138	.213
Druze	126	.270	0	0	0	0	.008	0	0	0	0	.008	0	0	0	0	.040	.286	.087
Adygei	98	.276	0	0	0	0	0	.031	0	0	0	0	0	0	0	.020	.031	.133	.133
Russians	64	.500	0	0	0	0	0	0	0	.016	0	0	0	0	0	.016	.031	.156	.125
Finns	66	.227	0	0	0	0	0	0	0	.076	.015	0	0	0	0	.030	.030	.121	.182
Catalans	166	.404	0	0	.012	0	0	.018	.018	.036	0	.006	.018	0	0	0	.066	.157	.102
Basques	216	.204	0	0	.009	0	.009	0	.009	.019	0	.005	.005	0	0	.023	.097	.218	.060
Kazakhs	66	.106	0	0	0	0	0	.061	.015	.030	0	0	0	0	0	.030	.061	.106	.288
Chinese	86	.012	0	0	0	0	0	0	.023	0	0	0	0	.012	0	0	.023	.105	.407
Japanese	88	0	0	0	0	.080	0	0	0	0	0	0	0	0	0	0	.011	.205	.352
Nasioi	46	.043	0	0	0	0	0	.326	0	.261	.087	0	0	0	.022	0	0	.152	.087
Yakut	86	.035	0	0	0	0	0	.035	.012	0	0	0	0	0	0	0	0	.093	.465
Maya	98	.031	0	.061	0	0	.020	0	.020	0	.071	.041	0	0	0	0	.010	.051	.224
Surui	84	0	0	0	0	0	0	0	0	0	0	0	.036	0	0	0	0	.024	.202

REPEATS	32	33	34	35	36	37	38	39	40	42	43	44	45	46	47	49	51	53
POPULATION																		
Biaka	.065	0	.008	.024	.081	.008	.129	0	0	0	0	0	0	0	0	0	0	0
Mbuti	.030	0	.015	.030	0	0	0	0	.015	0	0	0	0	0	0	0	0	0
Tanzanian	.069	.056	.028	.028	.028	0	.014	0	0	0	0	0	0	.014	.014	0	0	0
Saharawi	.055	.018	.027	.009	0	0	0	0	0	0	0	0	.009	.009	0	0	0	0
Yemenites	.100	.063	.013	0	.013	0	.025	0	0	0	0	0	.025	.013	.013	0	0	0
Druze	.016	.040	.048	.048	.008	.008	0	0	0	0	0	.063	.048	.008	.016	0	0	0
Adygei	.224	.092	.010	0	0	0	.010	0	0	0	0	0	.020	.020	0	0	0	0
Russians	.063	.031	.031	.016	0	0	.016	0	0	0	0	0	0	0	0	0	0	0
Finns	.076	.106	.015	.015	.015	0	.015	0	0	.030	.045	0	0	0	0	0	0	0
Catalans	.060	.036	.018	.018	0	.006	.012	0	0	0	.006	0	0	0	.006	0	0	0
Basques	.079	.097	.037	.023	0	0	.005	.009	0	0	0	0	.023	.046	0	.014	.005	.005
Kazakhs	.091	.091	.106	.015	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chinese	.163	.035	.105	0	.012	.035	.035	0	0	0	.012	.023	0	0	0	0	0	0
Japanese	.193	.068	0	.011	0	.011	.011	0	0	0	0	.045	.011	0	0	0	0	0
Nasioi	0	.022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Yakut	.198	.047	.047	.070	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maya	.296	.031	.041	.031	.041	0	.010	0	.020	0	0	0	0	0	0	0	0	0
Surui	.190	0	.012	.238	0	.298	0	0	0	0	0	0	0	0	0	0	0	0

6. TUB20

	ALLELE	1	2
POPULATION	**2N**		
Biaka	124	.274	.726
Mbuti	66	.136	.864
Tanzanian	72	.097	.903
Saharawi	106	.321	.679
Yemenites	80	.138	.863
Druze	126	.183	.817
Adygei	98	.214	.786
Russians	60	.300	.700
Finns	66	.197	.803
Catalans	166	.259	.741
Basques	216	.176	.824
Kazakhs	66	.076	.924
Chinese	86	.023	.977
Japanese	86	0	1.000
Nasioi	46	.022	.978
Yakut	78	.026	.974
Maya	94	.032	.968
Surui	84	0	1.000

CAPÍTOL III

Can a place of origin of the main CF mutations be recognized?

Eva Mateu, Francesc Calafell, Maria Dolors Ramos i Jaume Bertranpetit

(sotmès a consideració editorial)

CAN A PLACE OF ORIGIN OF THE MAIN CF MUTATIONS BE RECOGNIZED?

Eva Mateu[1], Francesc Calafell[1], Maria Dolors Ramos[2] and Jaume Bertranpetit[1]

[1]Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona.
[2] Departament de Genètica Molecular, Institut de Recerca Oncològica, Barcelona.

Correspondence address:

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.
Tel: +34-93-542 28 40
Fax: +34-93-542 28 02
e-mail: jaume.bertranpetit@cexs.upf.es

The genetic background of the most frequent mutations causing cystic fibrosis (CF) has long been recognized to be different from that of non-CF chromosomes[1,2]. An open question has been whether these haplotype backgrounds could be found at high frequencies in populations were CF is not common nowadays[3]. An analysis of CFTR haplotypes in a worldwide survey in normal chromosomes alongside with its variation in specific CF chromosomes shows:

i) A very low frequency or absence of the most common CF haplotypes in all populations analyzed.

ii) A strong genetic variability and divergence of the chromosomes carrying disease-causing mutations in relation to population differences in a worldwide perspective. For this genome region, genetic differences by CF mutation are much stronger than by population. These results give further support to previous estimates of an ancient, pre-Neolithic age for the most frequent CF mutation[4,5]. In fact, the depth of the gene genealogy associated with disease-causing mutations may be older than the evolutionary process that gave rise to present day human populations. The concept of "place of origin" lacks either spatial or temporal meaning for mutations that are likely to have been present in Europeans before the ethnogenesis of present populations.

The understanding of a genetic disease implies not only its genetic and functional identification, but also its natural history, that is, the place and time of origin of the genetic variant and an explanation of its presence, frequency and geographic distribution. It includes the genetic mechanisms that may explain its appearance (like mutation rate and pattern), the population factors that may explain its frequency and distribution (migration, population size and drift), and the interaction with the environment (selection in its rich variety).

Cystic fibrosis (CF) is the most common severe autosomal recessive disease in populations of European origin, where it affects one in 2,500 individuals. It is caused by mutations in the Cystic Fibrosis Transmembrane

Conductance Regulator (*CFTR*) gene, which was identified and cloned in 1989 (refs. 6-8). Since then, more than 970 mutations have been reported[9]. Among the different CF mutations, a deletion of 3 bp at codon 508 (ΔF508) is the most frequent, accounting for two thirds of the global CF chromosomes. Only four other mutations (G542X, N1303K, G551D and W1282X) have overall frequencies higher than 1% of the CF chromosomes. These five mutations are found throughout Europe, although with clear geographical patterns; for instance, ΔF508 shows a NW to SE gradient, with a maximum in Denmark (87.2% of all CF chromosomes) and a minimum in Turkey (21.3%)[10]. In addition, 17 other mutations have frequencies between 0.1 and 0.9% (ref. 11), and most of the remaining mutations are rare or confined to few populations.

Several polymorphic short tandem repeat (STRPs) 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[4,5,12,13]. SNPs can be used to define the haplotypic frameworks on which *CFTR* mutations occurred. Microsatellite markers can help to measure genetic variability within the *CFTR* locus and to estimate the age of CF mutations.

The most frequent CF-causing mutations are found in two groups of haplotype backgrounds. Considering six markers in its chromosomal order (Fig. 1); i.e., *IVS1CA, IVS6aGATT, IVS8CA, T854, IVS17bTA* and *TUB20*, these haplotype background groups are:

i) (20/21/22)-6-(16/17/18/22/23/24)-1-(30/31/32/33/34)-2, of which the most frequent are

21-6-23-1-31-2 for ΔF508 and N1303K mutations and 21-6-23-1-33-2 for G542X mutation; it is evident that these three different mutations are found in very closely related haplotypes, and ii) 7-(16/17)-2-7-1, in which G551D and W1282X are found. For these mutations no data are available for *IVS1CA*, the first marker.

When analyzing non-disease chromosomes for the same populations where CF is present (usually non-disease chromosomes of carriers) it became evident that the general genetic background was very different from that of CF chromosomes[1,2]. As CF is mostly confined to populations of European ancestry, it

seemed likely that the origin of the most frequent CF mutations should be non-European[5].

Moreover, as the overdominance hypothesis gained support to explain the high frequency of the overall CF mutations[14-16], places where putative positive selection could act on carriers seemed to be more likely to exist outside Europe. These should include places where diarrhoea-causing diseases were a major selective force, as cholera is known only from very recent times and cannot account by itself for the high frequency of CF[12].

We undertook a worldwide survey in order to find populations where specific haplotypes linked to CF mutations would be present at high frequency. It comprised 17 populations widely distributed (Fig. 2).

Considering the haplotypes associated specifically with ΔF508, even if several alleles for the STRs are included, no population has them in frequencies as to suggest a place of likely origin for ΔF508 (Table 1, columns A and B). Only minor traces of chromosomes bearing phylogenetically related haplotypes are found in several European and Asian populations (Table 1, columns C and D).

Haplotypes associated to CF mutations G542X and N1303K are closely related to those of ΔF508, and thus the situation is similar to that of ΔF508 backgrounds. When taking the three haplotype backgrounds together (and therefore a deeper branch in the gene tree of haplotypes), frequencies in normal chromosomes remain still at very low levels (Table 1). It should be kept in mind that the three mutations are independent unique events and that their occurrence in a similar background gives support to the hypothesis that the three mutations arose in a single population where these haplotypes were frequent. In that hypothetical population, demographic events (such as an expansion) or selective agents could have brought those deleterious mutations to relatively high frequencies and could have spread in other areas.

The situation is very different for the two other frequent mutations (G551D and W1282X). Both are associated to closely related haplotypes, now found at high frequency in populations of different continents: Africa (Mbuti, 12%; Saharawi, 20.8%), Western Asia (Adygei, 17.3%), Europe (Catalans, 16.7%) and in the Yakut (2.6%). It has not been found in other Eastern Asian populations or in

America. This is compatible with a local origin in Europe for those mutations, although a geographic structure that would allow to narrow the birthplace of these mutations is not evident.

To further explore the genetic-population relationships, a tree was built using maximum likelihood, including allele frequencies for chromosomes non-CF and with specific CF mutations; only in the case of ΔF508 it has been possible to consider different populations (Fig. 3). Results are clear: the main stratifying factor is genetic background (i.e., carrying or not CF mutations) rather than population, even when very distant populations are considered. These results stress the importance of the genetic background and the lesser importance of the population where the individual (or, better, a chromosome) belongs.

Age of ΔF508 is controversial, with estimates ranging from 3,000 (post-Neolithic)[17] to more than 40,000 (in the Upper Paleolithic, and clearly pre-Neolithic)[5]. A recent study using coalescence theory[18] has confirmed an old age for the mutation, with estimates ranging from 11,000 to 34,000 years ago. Even though the estimates are strongly dependent on genetic (such as mutation rate and selection) and demographic (expansion dynamics and population size) parameters, the mutation is clearly pre-Neolithic and it is an ancient mutation in human history. Other CF mutations have slightly younger ages[4]. These old ages are in agreement with the stronger differentiation of genetic variation by CF mutations.

The problem of the place or population of origin of CF mutations lacks sense as for several of them their origin in time is likely to be much earlier than the evolutionary process that gave rise to the present distribution of European populations[19]; in fact the structure of genetic variation is sharply distinct between non-CF and groups of disease-causing mutations. Being the ethnogenesis process much younger than the origin of those mutations, and given the possibility of undertaking genetic analyses through a molecular evolution perspective (including the coalescent process), a population approach may not result adequate for a detailed comprehension of the evolutionary processes that gave rise to CF mutations, as has been shown for other genome regions[20].

Methods

Polymorphic sites and population samples. We analyzed four short tandem repeats (STRPs) and two single nucleotide substitutions (SNPs) located within *CFTR* (Fig. 1). These are, in 5' to 3' order, *IVS1CA*[21,22]; *IVS6aGATT*[23], *IVS8CA*[24], *T854* (ref. 25), *IVS17bTA*[26] and *TUB20* (ref. 27). Samples typed comprised 949 unrelated autochthonous healthy individuals (1898 chromosomes) from 17 populations representing all major world geographic areas (Fig. 2). Appropriate informed consent was obtained from human subjects. DNA from five populations (Basques, Catalans, Tanzanian, Kazakhs and Saharawi) was extracted from fresh blood of blood donors. DNA samples for the other populations were obtained from lymphoblastoid cell lines maintained in K.K Kidd's laboratory at Yale University.

A total of 126 patients (252 chromosomes), carrying the ΔF508, G542X and/or N1303K mutations were typed for the *IVS1CA* locus. The samples were obtained at the Institut de Recerca Oncològica, Barcelona.

STRP and SNP analysis. Typing methods for six loci are as in ref. 3, where we reported allele and haplotype frequencies for the population set.

Statistical analysis. Maximum likelihood estimates of haplotype frequencies and the jackknife standard errors were calculated from the multi-site marker typing data, using the HAPLO program[28]. We used PHYLIP[29] to produce maximum likelihood representations of allele frequencies of five polymorphisms (no data were available for *IVS1CA* for chromosomes carrying G551D or W1282X) for normal and CF chromosomes. *IVS6aGATT, IVS8CA, T854, IVS17bTA* and *TUB20* allele frequencies for CF chromosomes (with ΔF508, G542X, N1303K, G551D and W1282X mutations) were obtained from the literature[2,5]. *IVS8CA* and *IVS17bTA* allele frequencies for ΔF508 chromosomes in geographically defined samples for European populations were from ref. 5.

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”), by the Dirección General de Investigación Científica y Técnica (Spain) grant PB98-1064, by Direcció General de Recerca, Generalitat de Catalunya (1998SGR00009), and by Institut d’Estudis Catalans. This work was also possible thanks to fellowships to E.M. (Universitat de Barcelona and Universitat Pompeu Fabra).

Druze and Yemenite samples were kindly supplied by Dr.Batsheva Bonné-Tamir from Sackler Faculty of Medicine (Tel Aviv University, Tel Aviv), samples from Tanzania 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 also especially thank Judith R. Kidd and Kenneth K. Kidd (Department of Genetics, Yale University School of Medicine, New Haven) for supplying Pygmy, Finn, Russian, Adygei, Chinese, Japanese, Yakut, Maya and Surui DNA samples. We would like to thank Montgomery Slatkin for useful comments on an earlier version of the manuscript.

References

1. Estivill X, Morral N, Bertranpetit J. Age of the ΔF508 cystic fibrosis mutation. *Nature Genet* **8**,216-218 (1994).

2. Morral, N. et al. Haplotype analysis of 94 cystic fibrosis mutations with seven polymorphic CFTR DNA markers. *Hum Mut* **8**,149-159 (1996).

3. Mateu, E. et al. Worldwide genetic analysis of the CFTR region. *Am J Hum Genet* **68**,103-117 (2001).

4. Morral, N. et al. Microsatellite haplotypes for cystic fibrosis: mutation frameworks and evolutionary tracers. *Hum Mol Genet* **2**,1015-1022 (1993).

5. Morral, N. et al. The origin of the major cystic fibrosis mutation (ΔF508) in European populations. *Nature Genet* **7**,169-175 (1994).

6. Kerem, B.S. et al. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**,1073-1080 (1989).

7. Riordan, J.R. et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**,1066-1073 (1989).

8. Rommens, J.M. et al. Identification of the cystic fibrosis gene; chromosome walking and jumping. *Science* **245**,1059-1065 (1989).

9. Cystic Fibrosis Mutation Database; http://www.genet.sickkids.on.ca/cftr

10. European Working Group on CF Genetics. Gradient of distribution in Europe of the major CF mutation and of its associated haplotypes. *Hum Genet* **85**,436-441 (1990).

11. Estivill, X., Bancells, C., Ramos, C. & Biomed CF Mutation Analysis Consortium. Geographic distribution and regional origin of 272 Cystic Fibrosis mutations in European populations. *Hum Mut* **10**,135-154 (1997).

12. Bertranpetit, J. & Calafell, F. Genetic and geographical variability in cystic fibrosis: Evolutionary considerations. in *Variation in the Human Genome* (ed. Cardew, G., Ciba Foundation Symposium 197) 97-118 (Wiley & Sons, Chichester, 1996).

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

14. Gabriel, S.E., Brigman, K.N., Koller, B.H., Boucher, R.C. & Stutts, M.J. Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* **266**,107-109 (1994).

15. Guggino, S.E. Evolution of the ΔF508 CFTR mutation. *Trends Microbiol* **7**,55-56 (1999).

16. Pier, G.B. et al. *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* **393**,79-82 (1998).

17. Serre, J.L. et al. Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Hum Genet* **84**,449-54 (1990).

18. Wiuf, C. Do ΔF508 heterozygotes have a selective advantage?. *Genetical Res* (In the press).

19. Moore, J.H. Putting anthropology back together again: the ethnogenetic critique of cladistics theory. *Am Anthropologist* **96**,925-948 (1994).

20. Bosch, E. et al. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet* **65**,1623-38 (1999).

21. Moulin, D.S., Smith, A.N. & Harris, A. A CA repeat in the first intron of the CFTR gene. *Hum Hered* **47**,295-297 (1997).

22. Mateu, E. et al. Allele frequencies in a worldwide survey of a CA repeat in the first intron of the CFTR gene. *Hum Hered* **49**,15-20 (1999).

23. Chehab, E.F. et al. 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 (1991).

24. Morral, N., Nunes, V., Casals, T. & Estivill, X. CA/GT microsatellite alleles within the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene are not generated by unequal crossingover. *Genomics* **10**,692-698 (1991).

25. Zielenski, J. et al. Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. *Genomics* **10**,214-228 (1991a).

26. Zielenski, J., Markiewicz, D., Rininsland, F., Rommens, J.M. & Tsui, L.C. A cluster of highly polymorphic dinucleotide repeats in intron 17b of the CFTR gene. *Am J Hum Genet* **49**,1256-1262 (1991b).

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

28. Hawley, M.E. & Kidd, K.K.. HAPLO: a program using the EM algorithm to estimate frequencies of multi-site haplotypes. *J Hered* **86,**409-411 (1995).

29. Felsenstein, J. PHYLIP-Phylogeny inference package. *Cladistics* **5**,164-166 (1989).

30. The Allele Frequency Database; http://info.med.yale.edu/genetics/kkidd

Table 1. Frequencies in several populations of normal chromosomes of haplotypes associated to CF mutations ΔF508, G542X and N1303K

Haplotype:	A	B	C	D
Population	% ± SE*	% ± SE	% ± SE	% ± SE
Druze	2.4 ± 1.4	0	4 ± 1.7	0
Basques	0	0	4.2 ± 1.4	0
Catalans	0	0	1.4 ± 0.9	0.6 ± 0.6
Finns	0	1.6 ± 1.6	3.2 ± 2.2	1.6 ± 1.6
Russians	0	0	1.7 ± 1.7	1.7 ± 1.7
Adygei	0	0	2.0 ± 1.4	0
Japanese	0	0	0	1.2 ± 1.2

* Standard Error

The frequency in the populations not listed is zero.

Haplotype A: 21-6-23-1-31-2

Haplotype B: 21-6-17-1-31-2

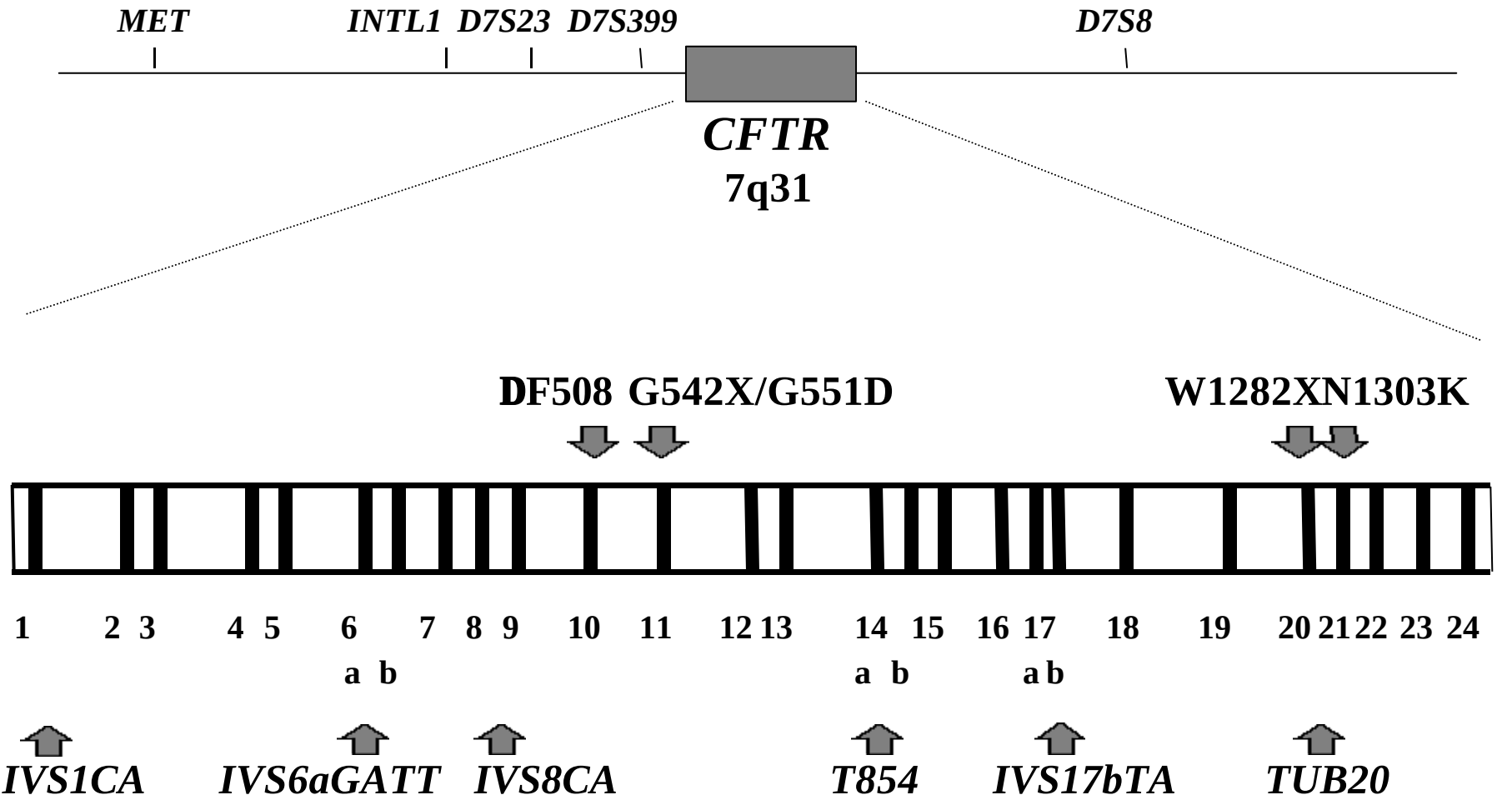
Haplotype C: (20/21/22)-6-(22/23/24)-1-(30/31/32/33/34)-2

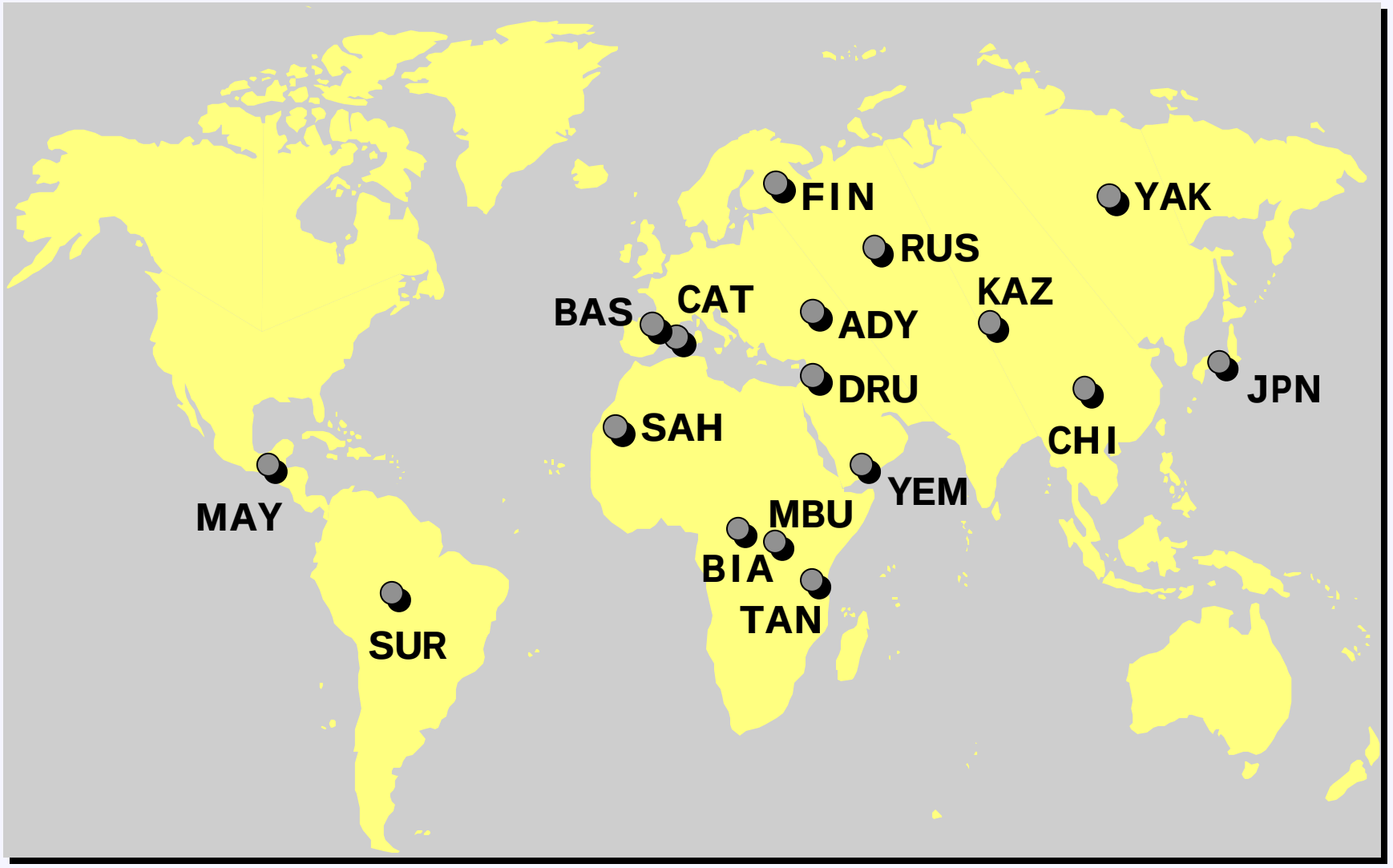
Haplotype D: (20/21/22)-6-(16/17/18)-1-(30/31/32)-2

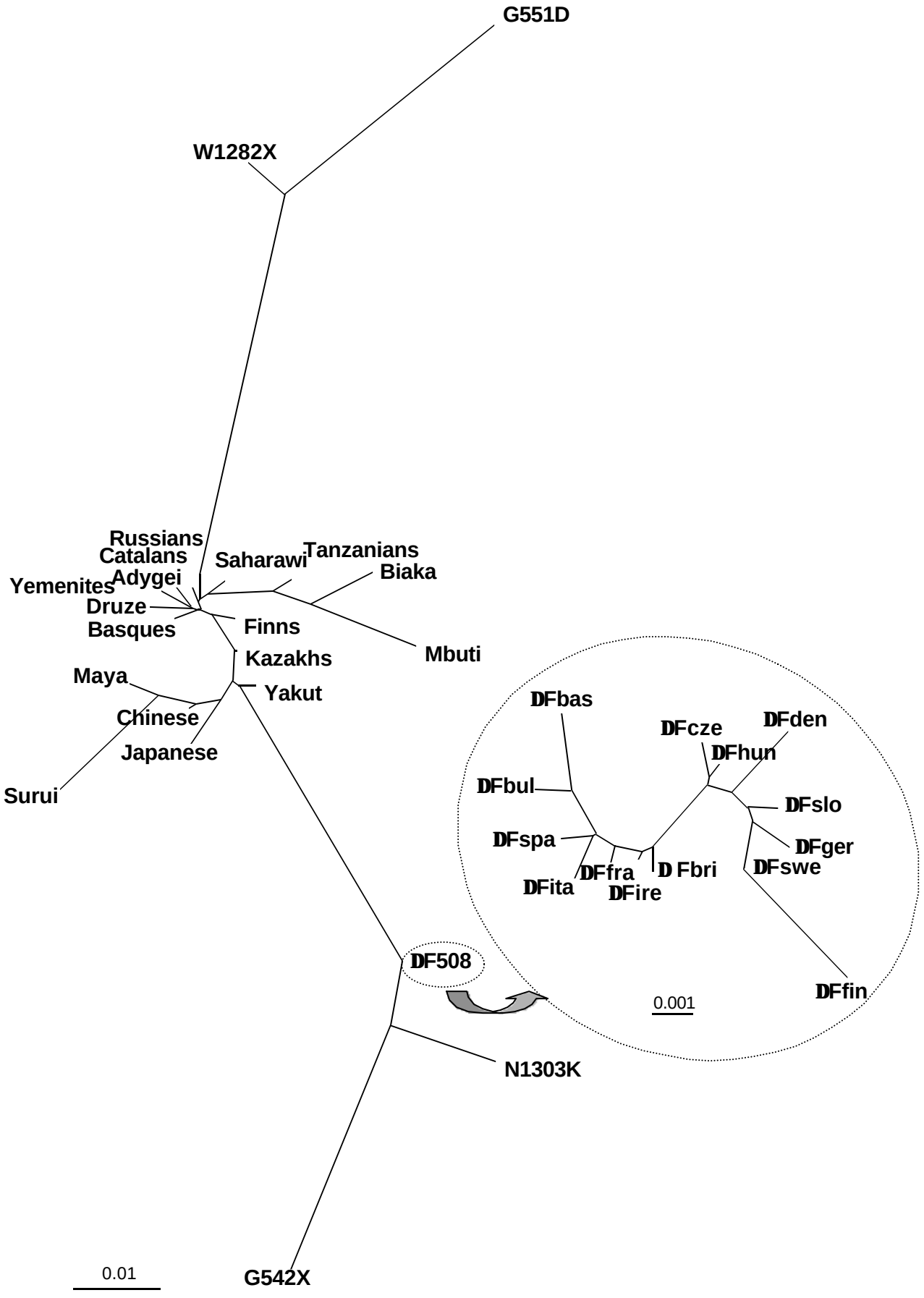
Figure 1. Polymorphisms in the *CFTR* region (*IVS1CA, IVS6aGATT, IVS8CA, T854, IVS17bTA, TUB20*), and location of the five most common CF mutations (ΔF508, G542X, N1303K, G551D and W1282X). *CFTR* exons are numbered 1 to 24.

Figure 2. Distribution of 17 worldwide populations analyzed. Sample sizes (2N) are given next in parentheses after each population name. ADY = Adygei (106); BAS = Basques (222); BIA = Biaka Pygmies (138); CAT = Catalans (176); CHI = Han Chinese (124); DRU = Druze (126); FIN = Finns (70); JPN = Japanese (96); KAZ = Kazakhs (80); MAY = Maya (106); MBU = Mbuti Pygmies (78); RUS = Russians (96); SAH = Saharawi (118); SUR = Surui (94); TAN = Tanzanians (80); YAK = Yakut (102); YEM = Yemenites (86). Further information on these population samples can be found in refs. 3 and 30.

Figure 3. Maximum likelihood tree of allele frequencies of five loci (*IVS6aGATT, IVS8CA, T854, IVS17bTA* and *TUB20*) among normal chromosomes, from worldwide populations, and among CF chromosomes (ΔF508, G542X, N1303K, G551D and W1282X chromosomes). The inset shows an enlarged maximum likelihood tree of allele frequencies of two loci (*IVS8CA* and *IVS17bTA)* among ΔF508 chromosomes in different European populations (Bas, Basque Country; Bul, Bulgaria; Bri, Great Britain; Cze, Czech Republic; Den, Denmark; Fin, Finland; Fra, France; Ger, Germany; Hun, Hungary; Ire, Ireland; Ita, Italy; Slo, Slovakia; Spa, Spain; and Swe, Sweden). Bars show the scale in genetic distance units. Several trees have been built, either using other methods (e.g. Nei and Kimura distance and the neighbor-joining algorithm) or different sets of chromosomes. In all cases results are similar.


MET
INTL1
D7S23
D7S399
D7S8
CFTR
7q31
DF508
G542X/G551D
W1282XN1303K
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
a b
a b
a b
IVS1CA
IVS6aGATT
IVS8CA
T854
IVS17bTA
TUB20


FIN
YAK
RUS
KAZ
BAS
CAT
ADY
DRU
JPN
SAH
CHI
YEM
MAY
MBU
BIA
TAN
SUR


G551D
W1282X
Russians
Catalans
Saharawi
Tanzanians
Biaka
Adygei
Yemenites
Druze
Basques
Finns
Mbuti
Kazakhs
Maya
Yakut
Chinese
Japanese
Surui
DFbas
DFcze
DFden
DFhun
DFbul
DFslo
DFspa
DFger
DFswe
DFfra
DFbri
DFita
DFire
DF508
DFfin
0.001
N1303K
0.01
G542X

CAPÍTOL IV

The PKLR-GBA region shows almost complete linkage disequilibrium over 70 kb in a set of worldwide populations

Eva Mateu, Francesc Calafell, Rosa Martínez-Arias,
Anna Pérez-Lezaun, Aida Andrés, Mònica Vallés i
Jaume Bertranpetit

(manuscrit en preparació)

The PKLR-GBA region shows almost complete linkage disequilibrium over 70 kb in a set of worldwide populations

Eva Mateu, Francesc Calafell, Rosa Martínez-Arias, Anna Pérez-Lezaun, Aida Andrés, Mònica Vallés and Jaume Bertranpetit

Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona.

Correspondence address:

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.
Tel: +34-93-542 28 40
Fax: +34-93-542 28 02
e-mail: jaume.bertranpetit@cexs.upf.es

ABSTRACT

Haplotype diversity in a ~70 kb genomic region in 1q21 comprising genes PKLR and GBA (and at least three others) was characterized by typing one single nucleotide polymorphism (SNP) in PKLR, two SNPs in GBA and one short tandem repeat polymorphism (STRP) in PKLR in 1792 chromosomes from 17 worldwide populations. Additionally, two other SNPs in GBA were typed in three African populations. Both with the basic three SNPs and with the additional two SNPs in GBA, most chromosomes carried one of either two phylogenetically distinct haplotypes that carried different alleles at each site. Allele diversity at the STRP was tightly linked to haplotype background. Linkage disequilibrium (LD) was significant for all SNP pairs in all populations, although it was slightly higher in non-African populations than in sub-Saharan Africans. Both the haplotype structure and an independent analysis of sequence variation at the GBA pseudogene 16 kb downstream from GBA, suggested that this region has experienced one or two selection events that carried the two main haplotypes to high frequencies. Based on variability at the PKLR STRP and on the geographical distribution of LD, these selection events may have predated the “Out of Africa” expansion of anatomically modern humans. Given the slight reduction in LD and in STRP variability in non-Africans vs. Africans, it seems likely that the human group that founded all non-African populations was relatively large.

INTRODUCTION

Two important human diseases, Gaucher disease (MIM #230800, #230900 and #231000, for Gaucher types 1, 2, and 3 respectively) and pyruvate kinase deficiency hemolytic anemia (MIM #266200), are caused by mutations at two genes (GBA and PKLR, respectively), that map closely to each other in 1q21 (Barneveld et al. 1983; Ginns et al. 1985; Satoh et al. 1988). Gaucher disease is the most frequent lysosomal storage disorder, with a high prevalence in the Ashkenazi Jewish population. Pyruvate kinase (PK) deficiency is the most common enzymopathy of anaerobic glycolysis resulting in hereditary hemolytic anemia. Beutler et al. (1992) described twelve polymorphic sites in the introns and flanking regions of glucocerebrosidase (GBA) gene. These polymorphisms were in marked linkage disequilibrium: out of 6,144 different possible haplotypes only four haplotypes were found and only two of these were common, not only in disease chromosomes but also in the normal chromosomes analyzed. Lenzner et al. (1997) described four PKLR polymorphisms defining different haplotype backgrounds of normal and PK genes. Glenn et al. (1994) found that two polymorphisms, one in the PKLR gene and one in the GBA gene, were tightly linked in a sample of 81 normal unrelated subjects. Close linkage of the PKLR and GBA genes was also demonstrated by Rockah et al. (1998) and by Demina et al. (1998), who found linkage disequilibrium between the two common Ashkenazi Jewish mutations in GBA and polymorphisms in the PKLR gene. Demina et al. (1998) determined that the distance between the 5' ends of GBA and PKLR was 71 kb. Within the 1q21 region, several genes and pseudogenes have been described (Winfield et al. 1997). A pseudogene of the GBA gene is located about 16 kb downstream from the functional gene (Horowitz et al. 1989; Zimran et al. 1990). The pseudogene GBA (psGBA) is closely (96%) homologous to the functional gene. Martínez-Arias et al. (2001) described sequence variation for over 5.4 kb in 100 psGBA chromosomes and found only two haplotypes at frequencies >10%, although together they account for more than half of the chromosomes in that sample.

We have typed four GBA single nucleotide substitutions (SNPs) and two PKLR polymorphisms (one short tandem repeat polymorphism (STRP) and one SNP) in a worldwide survey of 1792 chromosomes from 17 populations, in order to understand linkage disequilibrium in depth and to extricate genomic from population factors contributing to it.

MATERIAL AND METHODS

Polymorphic sites

The polymorphisms studied are located within the GBA and PKLR genes as shown in Figure 1. We typed four single nucleotide substitutions(SNPs) in GBA; listed from the 5' to the 3' end of the GBA gene, the polymorphisms analyzed are as follows: GBA2834 g/c/a, GBA3931 a/g, GBA5135 a/c and GBA6144 a/g, located, respectively, in introns four, six, seven and nine of the GBA gene (Beutler et al. 1992). In PKLR, the polymorphisms typed are PKLR$(ATT)_n$, a highly polymorphic trinucleotide short tandem repeat (STRP) located in the eleventh intron of the gene (Lenzner et al. 1997) and PKLR1705 c/a, a SNP located in exon 12 (Lenzner et al. 1997).

Genetic distance between the 5’ ends of both genes is 71 kb (Demina et al. 1998); and between the GBA3931 a/g and GBA6144 a/g sites is 2.2 kb (GenBank accession #J03059). PKLR and GBA genes are respectively 8.4 and 7.6 kb long (Lenzner et al. 1997; Horowitz et al. 1989).

Population samples

We have studied 896 unrelated autochthonous individuals (1792 chromosomes) from 17 populations representing all major world geographic areas (Figure 2). Sample sizes ranged from 46 (Nasioi) to 188 (Basques) chromosomes, with a median of 105.

DNA from four populations (Basques, Catalans, Tanzanians and Saharawi) was extracted from fresh blood of healthy 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. Kidd and K.K.Kidd, at Yale University.

STRP analysis

PCR amplifications were performed using 50 ng of genomic DNA in a final 10 μl volume. The amplifications were carried out in a Perkin Elmer 9600 thermal cycler. PKLR$(ATT)_n$ repeats were amplified with flanking primers JHATT (fluorescently labelled) and JRS1 (Lenzner et al., 1994); with 30 PCR cycles, denaturing at 95° for 30 sec, annealing at 61° for 30 sec and extending at 72° for 50 sec. PCR products 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 377^{TM} sequencer. GeneScan 672^{TM} was used to collect the data, track lanes, and measure fragment sizes. Alleles are named by the estimated number of trinucleotide repeats, found by comparison to two heterozygous control individuals found to carry 13/14 and 14/15 repeat alleles, as determined by sequencing. Control DNAs were kindly provided by E. Beutler, Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla.

Analysis of Single Nucleotide Polymorphisms

Two SNPs from GBA (GBA3931/*PvuII* and GBA6144/*BglI*) and one from PKLR, (PKLR1705/*BspHI*) were analyzed by PCR amplification and digestion with the appropriate restriction enzyme, as described (Beutler et al. 1992, Demina et al. 1998).

Polymorphisms GBA2834g/c/a and GBA5135a/c (Beutler et al. 1992) were analyzed by using a method based on ARMS-PCR and fluorescently labelled primers. Both polymorphisms were amplified separately and then run together on an ABI 377 sequencer (PE Biosystems). Genescan ™ (ver. 3.1) and Genotyper ™ (ver. 2.5) software packages (PE Biosystems) were used to analyze and size the amplified fragments. Primers were chosen so that different genotypes can be visualized as amplified PCR fragments labelled with a different fluorescent dye and slightly different length. We used three forward and one reverse primer to

analyze GBA2834 g/c/a . The forward primer sequences were: 5'-CCTGTGAAATAAGATTTCG-3', 5'-CCCTGTGAAATAAGATTTCC-3', and 5'-GCCCTGTGAAATAAGATTTCA-3'; they were designed to anneal, respectively, to the G, C and A alleles, and were labelled respectively with Hex, Fam and Tet. The reverse, unlabelled primer sequence was 5'-TGCAACTGGAAATCATCAG-3'. GBA5135 a/c was typed with forward primers : 5'-TGCATTCTTCCCGTCACCCAA-3' (specific to allele A and labelled with Hex), 5'-CATTCTTCCCGTCACCCAC-3' (C-specific, Fam labelled), and with reverse primer 5'-GCAAAGGAAGAGCAACTGAT-3'.

Haplotypes and notation

Beutler et al. (1992) typed 12 polymorphisms in GBA and found basically three different haplotypes, which they named + , - or *African*. We have typed four of those GBA polymorphic sites. With these four polymorphisms, haplotype + would be GBA2834c, GBA3931g, GBA5135c and GBA6144g; haplotype - corresponds to GBA2834g, GBA3931a, GBA5135a and GBA6144a, and *African* is GBA2834a, GBA3931a, GBA5135a and GBA6144. Polymorphic site PKLR1705 c/a in the contiguous gene can be added to the former haplotype (Glenn et al. 1994): alleles A and C are tightly linked, respectively, to haplotypes + and -.

In this study, and for the sake of consistency and readability, we will refer to the alleles linked to haplotype + at any site as "1", and to those alleles linked with haplotype - as "2". GBA2834 shows three different alleles; GBA2834a (part of the so-called *African* haplotype) will be designed with 3. In summary, we will refer to haplotype PKLR1705a-GBA2834c-GBA3931g-GBA5135c-GBA6144g as 11111 and PKLR1705c-GBA2834g-GBA3931a-GBA5135a-GBA6144a as 22222. With this nomenclature, recombinants between the two basic haplotypes are easily spotted.

Statistical analysis

Allele frequencies were estimated by direct gene counting. Maximum-likelihood estimates of haplotype frequencies were calculated from the multi-site marker typing data, using the Arlequin program (Schneider et al. 2000), which implements the EM algorithm (Dempster et al. 1977; Slatkin and Excoffier 1996).

Expected heterozygosities for loci and for the haplotypes were estimated as $1-S p_i^2$, where p_i represents the allele or haplotype frequencies for the system.

The time since an expansion event was calculated from the STRP allele size variance by use of equation 5 in Di Rienzo et al. (1998): $V = T m h_2$ where V is the variance in repeat size, T stands for time in generations after a population expansion, and h_2 is the variance in the number of repeats gained or lost at each mutation event. Since neither a direct estimate of the mutation rate of the PKLR trinucleotide nor an average rate for trinucleotide STRPs are known, we used an upper estimate of the PKLR STRP mutation rate based on the average dinucleotide mutation rate. It has been suggested that non-disease-causing trinucleotides may have average mutation rates slower than those of dinucleotides but faster than tetranucleotide (Chakraborty et al. 1997); the average mutation rate (μ) used was 7.8×10^{-4}, which is an average estimated rate over ~2,000 dinucleotide repeats (Gyapay et al. 1994). Mutation-size variance (η_2) was set to 1, and the generation time used was 20 years.

Overall genetic heterogeneity among populations was tested through Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992).

Linkage disequilibrium analysis between pairs of loci was tested for genotype data using a likelihood-ratio test, whose empirical distribution is obtained by a permutation procedure (Slatkin and Excoffier 1996). Linkage disequilibrium analysis extended to psGBA included 94 chromosomes of the worldwide sample that overlapped with the study of Martínez-Arias et al. (2001). The classical linkage disequilibrium coefficient measure D' (Lewontin 1964) was tested for haplotype data by assuming that estimated haplotype frequencies corresponded to actual chromosomes.

RESULTS

Allele frequencies

STRP and SNP allele frequencies by population are shown in Table 1. In a worldwide human population sample we have typed both PKLR polymorphisms (PKLR(ATT)$_n$ and PKLR1705 c/a) and two GBA polymorphisms: GBA3931 a/g and GBA6144 a/g. Since preliminary results showed strong linkage disequilibrium for three SNPs (PKLR1705 c/a, GBA3931 a/g and GBA6144 a/g), and since additional markers in the region could be presumed to be little informative, the two remaining SNPs (GBA2834 g/c/a and GBA5135 a/c) were typed only for sub-Saharan African populations (i.e. Biaka and Mbuti Pygmies, and Tanzanians).

The overall PKLR(ATT)$_n$ allele frequency distribution shows a range of 7-19 repeats. Outside of Sub-Saharan Africa, it is bimodal with peaks at 12 and 14 repeats. In Sub-Saharan Africans, alleles 10, 11 or 13 are also quite frequent, which increases heterozygosity at this locus. As shown in Table 2, expected heterozygosities for PKLR(ATT)$_n$ are highest in Sub-Saharan African populations.

For the SNPs, a general tendency is shown in the allele frequencies: values for the two alleles seem inverted between sub-Saharan Africa and the rest of populations. Within sub-Saharan Africa, there is a clear difference between Biaka pygmies and the other populations.

Haplotype frequencies

SNP haplotype frequencies by population are shown in Table 3. We have found all eight possible haplotypes (111, 112, 121, 122, 211, 212, 221 and 222) with three SNP loci (PKLR1705 c/a-GBA3931 a/g-GBA6144 a/g). The total number of possible haplotypes with five SNP loci (PKLR1705 c/a-GBA2834 g/c/a-GBA3931 a/g-GBA5135 a/c-GBA6144 a/g) is 48, of which we have found 16 in sub-Saharan African populations.

The most common three-locus haplotypes are 111 (equivalent to haplotype + in the nomenclature by Beutler et al., 1992) and 222 (or haplotype –), found in all the populations studied. Sub-Saharan African and Asian populations have higher haplotype 111 frequencies, ranging from 0.513 in Mbuti to 0.685 in Japanese; in contrast with Middle East and Europeans where haplotype 222 is more frequent, ranging from 0.559 in Finns to 0.849 in Druze. The third most common haplotype worldwide is 211 (ranging from 0.008 in Surui to 0.205 in Nasioi and absent in Adygei and Yakut samples), followed by haplotype 122 (absent in a larger number of populations and ranging from 0.012 in Basques to 0.123 in Mbuti). The fact that both haplotypes are complementary in their allele composition as well as their associations with PKLR(ATT) alleles (as discussed below) suggest that these two haplotypes may be the result of a recombination event between the PKLR and GBA genes, which are 71 kb apart (Demina et al. 1998), rather than being in the mutational phylogenetic pathway between the 111 and 222 haplotypes. All other haplotypes (112, 121, 212 and 221) are less frequent worldwide.

The most common five-locus haplotypes for sub-Saharan African populations are 11111 (or haplotype +), 22222 (or haplotype –), 23222, 12222 and 21111, all of which where found in all three populations. The latter two may correspond also to recombination events between PKLR and GBA genes. Haplotype 23222 corresponds to the former *African* haplotype, with frequencies ranging from 0.029 in Biaka pygmies to 0.199 in Tanzanians.

Expected haplotype diversities in all populations are shown in Table 2. Sub-Saharan African populations have higher haplotype diversities than other populations. The lowest value (0.198) was found in the Surui, which may be due to drift and to that some of the chromosomes in the sample may be identical by descent (Calafell et al., 1999).

Figure 3 represents the PKLR(ATT)$_n$ allele distribution in different SNP haplotype backgrounds, for the overall worldwide sample and for three populations. In the worldwide distribution 86.3 % of all 12 alleles were found in a 111 haplotype, while 84.2 % of 14 alleles where found in 222 backgrounds. In

most populations, PKLR(ATT)$_n$ allele distributions are clearly conditioned by SNP haplotype frequencies. As illustrated by Figure 3, the high frequency of 222 determines the high frequency of allele 14 in the Adygei (as in all Europeans), while haplotype 111 and the associated allele 12 are most prevalent in East Asians (see the Japanese in Figure 3). However, the Mbuti Pygmies (as all sub-Saharan African samples examined) show a higher diversity of STRP alleles and slightly less pronounced associations with SNP haplotype backgrounds (for instance, allele 11 is almost equally divided between haplotypes 11111, 22222 and others).

Haplotype frequency estimation from joint genotype PKLR/GBA and psGBA haploid sequence data (Martínez-Arias et al., in preparation) yielded also two most frequent haplotypes; haplotype 222 associated with psGBA haplotype 17 (23.4%) and haplotype 111 associated with psGBA haplotype 3 (21.6%). In that subsample, 84.3% of all 111 PKLR/GBA haplotypes were associated to psGBA haplotype 3; conversely, 70.3% of all 222 PKLR/GBA haplotypes were associated to psGBA haplotype 17. Therefore, if sequence diversity is analyzed at the nucleotide level, the division of this genomic region in two main haplotypes persists, since haplotypes 17 and 3 defined from 5.4 kb of sequence for psGBA constitute the two most frequent haplotypes at psGBA and are phylogenetically distinct (Martínez-Arias et al. 2001)

Dating STRP variability in different SNP haplotype backgrounds

Both STRP variances for SNP haplotypes + and - are very close (1.90 and 1.64 respectively). Therefore, the time necessary for generate the current STRP variability in each haplotype background is also very similar: 49,000 and 42,000 years (see methods). For the overall STRP distribution (with an STRP variance of 2.95) the time turn into 76,000 years. Moreover, as it has been shown in other studies (Bosch et al. 1999; Mateu et al. 2001), STRP allele-size variances are determined to a much greater extent by SNP haplotype than by population; this finding casts doubt on the age estimation by STRP variation without taking into

consideration the genetic background. Population differences would explain only 6.5% of the variation in STRP allele-size variances, as determined by ANOVA.

SNP haplotype phylogeny

The nucleotide state in other hominoids at the homologous site can be used to infer the ancestral state for the SNPs (Iyengar et al. 1998). In our case, a chimpanzee GBA gene sequence, including GBA 2834 g/c/a, GBA3931 a/g, GBA5135 a/c and GBA6144 a/g polymorphisms, among others GBA polymorphisms described by Beutler et al. (1992), was available (Martínez-Arias et al., in press). The chimpanzee sequence, which was homozygous for all the sites that are known to be polymorphic in humans, was GBA2834g, GBA3931g, GBA5135a and GBA6144a, that is, 2122 in our notation. This means that neither of the two previously described haplotypes (+ or -) are ancestral. In the African populations, in which all four sites were typed, haplotype 2122 was absent. Haplotype 2222 is one mutational step from the ancestral 2122, while haplotype 1111 is three steps away; thus, haplotype 2222 may have arisen earlier than 1111. Then, both haplotypes may have drifted to high frequency, or, more likely, may have become frequent due to positive selection in that genome region (Martínez-Arias et al., in preparation). PKLR sequences are not available for any non-human primate, and thus, the ancestral state for site PKLR 1705 remains unknown.

Most of our data consist of three-site PKLR/GBA haplotypes; as described, haplotypes 111 and 222 are the most prevalent, and 222 may be closer to the ancestral haplotype. Some aspects of the phylogeny of the remaining haplotypes may be tackled. Haplotypes 122 and 211 are next haplotypes in frequency and may be explained by at least two different mechanisms: i) recent, parallel backmutation at site PKLR1705 from the frequent haplotypes 222 and 111, or ii) recombination involving 111 and 222, between genes PKLR and GBA. These two mechanisms would have different effects on allele associations with PKLR $(ATT)_n$, which is found upstream from PKLR1705. As stated above, the 12-repeat allele is associated with the 111 haplotype, and the 14-repeat allele is associated with the

222 haplotype. Thus, parallel backmutation would associate $(ATT)_{12}$ with 122 and $(ATT)_{14}$ with 211, while the reciprocal associations would be found if those haplotypes had arisen by recombination. In fact, 47.9% of all 211 haplotypes carried 14 repeats at (ATT), while the 12-repeat allele was found at one quarter of that frequency (11.9%). Reciprocally, more than half (53.5%) of the 122 haplotypes carry 12 repeats at ATT, and none carried 14 repeats. Thus, haplotypes 122 and 211 seem to have arisen by recombination rather than by backmutation.

Linkage disequilibrium analysis

Linkage disequilibrium between PKLR/GBA pairs of SNPs from genotype data (Slatkin and Excoffier 1996) was significant in each population ($p<0.005$, except for Nasioi, $p<0.02$). The classical linkage disequilibrium coefficient measure D' (Lewontin 1964), computed assuming that the estimated haplotype frequencies corresponded to actual chromosomes, showed significant linkage disequilibrium between all pairs with most D' = 1 (Table 4), independently of the population origin. A slightly stronger linkage disequilibrium is found between pairs GBA3931-GBA6144 (i.e. the lowest genetic distance, 2.2kb) with a mean population D' value of 0.98, while average D' between PKLR1705 and any of the two GBA polymorphisms was 0.92. Compared with LD studies between SNP pairs in other genome regions, for a similar genetic distance range (Table 5), the PKLR-GBA region presents the highest values of LD found. In the study of Jorde et al. (1994), the highest D' values corresponded to intragenic comparisons, while Kidd et al. (2000) only found significant LD in the American Indian populations.

Eisenbarth et al (2000) related LD to GC content by studying an isochore transition in the NF1 gene region, where they found high values of LD in a region of low (37.2%) GC content, in contrast with low values of LD in a neighboring region of high (51%) GC content. GC content is even higher in the PKLR-GBA region (54.2 % in PKLR; 52.8 % in GBA; 52.3% in the region encompassing from CLK2 to THBS3 (Winfield et al. 1997)), where we found high LD, in contrast with the apparent relation described by Eisenbarth et al (2000).

Linkage disequilibrium was complete for 7 of the 17 populations: Saharawi, Yemenites, all three East Asia samples, the Nasioi and the Surui. Average D' ranged from 0.76 to 0.93 in sub-Saharan Africans, from 0.84 to 1 in Europeans and Middle Easterners, and was 0.89 in the Maya. Thus, the PKLR-GBA region seems to conform to a pattern of lower LD among Africans, which grows in Europeans and East Asians (Kidd et al. 1998, 2000; Tishkoff et al. 1996, 1998, 2000). It should be noted that exceptions to this pattern have been found, such as CFTR (Mateu et al. 2001).

For 94 out of 1792 chromosomes in the worldwide sample, LD analysis could be extended to 16 SNPs on psGBA that had been detected by sequence analysis (Martínez-Arias et al. 2001, in preparation). A likelihood test of linkage disequilibrium between genotypic data from three PKLR/GBA SNPs and 11 variable sites on psGBA (singletons excluded), was significant for 28.6% of all pairs ($p < 0.05$) and when Bonferroni correction was applied 19.8% remained significant ($p < 0.001$). If doubletons in psGBA were excluded as well, LD nominally was significant ($p < 0.05$) for 61.1% of all (PKLR/GBA) - psGBA pairs; when Bonferroni correction was applied 47.2% remained significant ($p < 0.002$).

DISCUSSION

Haplotype phylogeny

Usually, one of at least two criteria are used to decide which allele is ancestral at a SNP: which is the most frequent allele (in fact, the allele frequency itself is the probability that the allele is ancestral under neutral conditions; Watterson and Guess, 1977), or which allele is found in non-human primates (Iyengar et al. 1998). In fact, both criteria tend to coincide: Hacia et al. (1999) showed that the ancestral states in 162 out of 214 SNPs, as determined by typing a number of chimpanzees, bonobos and gorillas, were also the most frequent in a sample of 412 humans from 10 worldwide populations. In GBA, the situation may be slightly more complex. We have found that four SNPs at GBA in chimpanzee determine a haplotype (2122) that is not found in sub-Saharan populations; the ancestral haplotype for the two SNPs for which we had worldwide data (GBA 3931 and GBA 6144), is rare worldwide, with an average 1.6% and a maximum frequency of 10% in any population. When comparing frequency vs. non-human primate states, it should be noted that most human variation is recent, and that for nuclear loci, times to the most recent human ancestors range from ~200,000 years ago (ya) to slightly over one milion ya, while the last human-chimpanzee common ancestor lived ~5 milion ya. Thus, divergence time between humans and our closest relatives is sufficient for some derived states to reach high frequencies. Moreover, selective sweeps can raise the frequency of a derived allele (Fay and Wu, 2000). In conclusion, a sounder phylogenetic analysis is often needed to infer ancestral states rather than relying mechanically on the frequency criterion.

LD in a genomic context

One of the main findings of this study is that LD within PKLR-GBA is higher than in other genomic regions of the same size. And LD might extend further: other genes in the neighbouring region show patterns of linkage disequilibrium and

highly conserved haplotypes. Pratt et al. (1996) described two main haplotypes with three polymorphisms (one SNP, one VNTR and one STRP) in the MUC1 gene, located ~22 kb downstream of psGBA (Vos et al. 1995). Moving away from GBA in 1q21, Volz et al. (1993) described two haplotypes, present in one cell line and established with some restriction enzymes, involving the cluster of epidermal differentiation genes, located ~5Mb upstream of the GBA gene (Human Genome Project). Herrmann et al. (1998) described two highly conserved haplotypes involving a set of polymorphisms in the P-selectin gene (1q21-q24), located ~16Mb downstream of the GBA gene (Human Genome Project). Moreover, Huttley et al. (1999), assessing the distribution of pairwise LD between STRPs throughout the human genome, found high levels of LD (Fisher's exact test probabilities = 0) in a large region of chromosome 1 neighbouring PKLR and GBA genes (between D1S2777-D1S303 and D1S484-D1S2705 loci pairs). Although joint data for all these genes are not available, it is tempting to speculate that LD extends over a broad region in the long arm of chromosome 1. Obviously, low recombination is needed to create and maintain LD, and, in fact, 1q21 is known to have a low recombination rate (~0.25 cM/Mb, or one quarter of the genome average, for roughly 20 Mb in either direction from GBA; Payseur and Nachman, 2000).

LD and selection

Recombination might not be sufficient to maintain two phylogenetically diverse haplotype backgrounds at high frequencies: selection may be acting to preserve those. Martínez-Arias et al. (in preparation) obtained 100 haploid sequences from 10 worldwide populations and described the complete sequence variability in 5.4 kb of the GBA pseudogene, which provided a powerful tool to test for departures from the neutral expectations. They found that the frequency spectra both of alleles at each site and of haplotypes did not fit the expectations of a neutral model and that those frequencies were best explained by selection acting over one or more linked loci, in two consecutive selective sweeps or in a single hitchhiking event plus recombination, which would have yielded similar

patterns. 1q21 is a centromeric region very rich in genes (73 gene entries in the Human Genome Project), which makes it difficult to find which are the actual variants being selected for.

Our results contain an STRP, which is an invaluable resource for following how genetic variability accumulates with time after selection or population expansion. The variability at PKLR(ATT)$_n$ could have accumulated in 49,000 years in the 111 background and in 42,000 years for the 222 background. As stated above, both the haplotype spectrum at PKLR-GBA and sequence variation at closely linked psGBA point to selection acting on both backgrounds. Thus, the times measured by allele variation at PKLR(ATT)$_n$ may mark the end of the selection event(s) that brought 111 and 222 to such high frequencies. It should be noted that those times are very rough estimates, given that they depend linearly on the unknown mutation rate of the STRP. Since both backgrounds are found at unexpectedly high frequencies both in sub-Saharan Africa and in all non-African populations, we may especulate that the selective event(s) predated the “Out of Africa” expansion of anatomically modern humans. However, and independently of the mutation rate, the variability accumulated by PKLR(ATT)$_n$ in either background is very similar, which would imply either balancing selection (i.e., the 111/222 heterozygotes were selectively advantageous), or two independent selective sweeps acting closely in time. In any case, the end product is clear: two very distantly related haplotypes found at inordinately high frequencies in all human populations sampled.

Selection and population history

The PKLR-GBA region seems to conform to a general pattern of lower LD among African populations, which grows outside Africa (Kidd et al. 1998, 2000; Tishkoff et al. 1996, 1998, 2000, among others) though maybe not so marked as in other loci. This pattern is expected under the “out of Africa” or replacement model of human origins, which postulates a recent, African origin of anatomically modern humans. Africans show a wider diversity of PKLR-GBA SNP haplotypes,

and a higher diversity of STRP alleles within each of those haplotypes, as expected also according to the replacement model. The pattern we found at PKLR-GBA contributes to the overall picture; for some loci, such as CFTR, LD in non-Africans in similar to that in Africans (Mateu et al. 2001); in other loci, LD in Africans is moderately less than in non-Africans (e.g., PKLR-GBA), while, in other loci, extreme differences have been found (CD4, Tishkoff et al. 1996; DM, Tishkoff et al., 1998). This overall pattern may be the result of the subsampling of genetic diversity that may have occurred when the first anatomically humans left Africa to found all other human populations; a moderately sized founding population would have sampled at random almost all diversity in some loci, most diversity in others and little in a few, thus generating increasing LD out of Africa for these different sets of loci. However, for some loci in which variability out of Africa is particularly restricted, it can not be ruled out that selection, besides a founder effect, contributed to the reduction in genetic variability observed in non-Africans.

The analysis of variation in and out of Africa for PKLR-GBA may be specially adequate to acquire a rough estimate of the proportion of the pre-existing genetic variation that emerged Out of Africa with the expansion of anatomically modern humans. Selection generated a set of two phylogenetically distinct haplotypes, which account for 81.4% of the chromosomes in Sub-Saharan Africans and for 93% of the chromosomes outside of Africa (excluding the Pacific and American populations, which may have experienced strong subsequent bottlenecks). The proportion between these two frequencies, that is, 0.875, may give a rough estimate of the proportion of the genetic variation in Africa that emerged with the first expansion of anatomically modern humans. An independent estimate can be obtained with heterozygosity at the STRP, which is, under different genetic models, more closely related to the parameter $\theta = 4N_e\mu$, where N_e stands for effective population size and μ for mutation rate. Since mutation rate is presumably the same in all populations, differences in heterozygosity (and thus, in θ), are related to differences in long term effective population size. The average heterozygosity of PKLR$(ATT)_n$ is 0.817 in Sub-Saharan Africans and 0.716 in North Africans, Middle Easterners, Europeans and Asians. This would imply that

the long-term effective population size of non-Africans is 87.6% of that of Africans, a value that is quite close to that obtained from haplotype frequencies. Taken at face value, this estimate would mean that the Out of Africa bottleneck was not narrow. However, this is a rough, single-locus estimate, which should be compared to that obtained from other loci.

We have shown how the use of a haplotype system comprising loci with different evolutionary rates can yield valuable insights for the comprehension of how selection and population history have interacted in the creation of the observed haplotype diversity and phylogeny.

ACKNOWLEDGMENTS

This research was supported by the Dirección General de Investigación Científica y Técnica (Spain) grant PB98-1064, by Direcció General de Recerca, Generalitat de Catalunya (1998SGR00009), and by Institut d'Estudis Catalans. This work was also possible thanks to fellowships to E.M. (Universitat de Barcelona and Universitat Pompeu Fabra). Druze and Yemenite samples were kindly supplied by Batsheva Bonné-Tamir from Sackler Faculty of Medicine (Tel Aviv University, Tel Aviv) and samples from Tanzania by Clara Menéndez from Unitat d'Epidemiologia i Bioestadística (Hospital Clínic, Barcelona). We also especially thank Judith R. Kidd and Kenneth K. Kidd (Department of Genetics, Yale University School of Medicine, New Haven) for supplying Pygmy, Finn, Russian, Adygei, Chinese, Japanese, Yakut, Maya and Surui DNA samples. We are in debt with E. Beutler and the staff of his lab (Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla), for pleasant welcome of E.M. in his lab helping her in the initial work.

REFERENCES

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GC, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WOC. Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 2001; 68:191-197

Barneveld RA, Keijzer W, Tegelaers FPW, Ginns EI, Geurts van Kessel A, Brady RO, Barranger JA, Tager JM, Galjaard H, Westerveld A, Reuser AJJ. Assignment of the gene coding for human β-glucocerebrosidase to the region q21-q31 of chromosome 1 using monoclonal antibodies. Hum Genet 1983; 64:227-231

Beutler E, West C, Gelbart T. Polymorphisms in the human glucocerebrosidase gene. Genomics 1992; 12:795-800

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

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. Mutation rate in human microsatellites: influence of the structure and lenght of the tandem repeat. Am J Hum Genet 1998; 62:1408-1415

Calafell F, Shuster A, Speed WC, Kidd JR, Black FL, Kidd KK. Genealogy reconstruction from short tandem repeat genotypes in an Amazonian population. Am J Phys Anthropol 1999; 108:137-146

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A 1997; 94:1041-6

Cox A, Camp NJ, Nicklin MJH, di Giovine FS, Duff GW. An analysis of linkage disequilibrium in the Interleukin-1 gene cluster, using a novel grouping method for multiallelic markers. Am J Hum Genet 1998; 62:1180-1188

Demina A, Boas E, Beutler E. Structure and linkage relationships of the region containing the human L-type pyruvat kinase (PKLR) and glucocerebrosidase (GBA) genes. Hematopathol Mol Hematol 1998; 11:63-71

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 1977; 39:1-38

Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, Petzl-Erler ML, Haines GK, Barch DH. Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 1998; 148:1269-84

Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics 1998; 148:1667-86

Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu C-F, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BAJ. The extent of linkage disequilibrium in four populations with distinct demographic histories Am J Hum Genet 2000; 67:1544-1554

Eisenbarth I, Vogel G, Krone W, Vogel W, Assum G. An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. Am J Hum Genet 2000; 67:873-880

Excoffier, L., Smouse, P., and Quattro, J. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 1992; 131:479-491

Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 2000; 155:1405-13

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, Bernardi G, Lathrop M, Weissenbach J. The 1993-94 Genethon human genetic linkage map. Nat Genet 1994; 7:246-339

Ginns EI, Prabhakara V, Choudary V, Tsuji S, Martin B, Stubblefield B, Sawyer J, Hozier J, Barranger JA. Gene mapping and leader polypeptide sequence of human glucocerebrosidase: implications for Gaucher disease. Proc Natl Acad Sci USA 1985; 82:7101-7105

Glenn D, Gelbart T, Beutler E. Tight linkage of pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hum Genet 1994; 93:635-638

Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SP, Collins FS. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet 1999; 22:164-7

Herrmann SM, Ricard S, Nicaud V, Mallet C, Evans A, Ruidavets JB, Arveiler D, Luc G, Cambien F. The P-selectin gene is highly polymorphic: reduced frequency of the Pro715 allele carriers in patients with myocardial infarction. Hum Mol Genet 1998; 7:1277-84

Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet 1968; 38:226-231

Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, Beutler E. The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 1989; 4:87-96

Huttley GA, Smith MW, Carrington M, O'Brien SJ. A scan for linkage disequilibrium across the human genome. Genetics 1999; 152:1711-22

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

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

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

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

Lenzner C, Jacobasch G, Reis A, Thiele B, Nürberg P. Trinucleotide repeat polymorphism at the PKLR locus. Hum Molec Genet 1994; 3:523

Lenzner C, Nürberg P, Jacobasch G, Thiele B-J. Complete genomic sequence of the human PK-L/R-Gene includes four intragenic polymorphisms defining different haplotype backgrounds of normal and mutant PK-Genes. DNA seq 1997; 8:45-53

Lewontin RC. The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 1964; 49:49-67

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

Martínez-Arias R, Calafell F, Buchanan A, Weiss KW, Mateu E, Comas D, Andrés A, Bertranpetit J. Sequence variability of a human pseudogene. Genome Res 2001 (in press)

Martínez-Arias R, Calafell F, Mateu E, Bertranpetit J. Profiles of accepted mutation: from neutrality in a pseudogene to disease-causing mutation on its homologous gene. Hum Genet (in press)

Martínez-Arias R, De Lorenzo D, Calafell F, Mateu E, Bertranpetit J. Selection shaped variability on a human pseudogene. (in preparation)

Mateu E, Calafell F, Lao O, Bonné-Tamir B, Kidd JR, Pakstis A, Kidd KK, Bertranpetit J. Worldwide genetic analysis of the CFTR region. Am J Hum Genet 2001; 68:103-117

Payseur BA, Nachman MW. Microsatellite variation and recombination rate in the human genome. Genetics 2000; 156:1285-98

Pratt WS, Islam I, Swallow DM. Two additional polymorphisms within the hypervariable MUC1 gene: association of alleles either side of the VNTR region. Ann Hum Genet. 1996; 60:21-28

Rockah R, Narinsky R, Frydman M, Cohen IJ, Zaizov R; Weizman A, Frisch A. Linkage disequilibrium of common Gaucher disease mutations with a polymorphic site in the pyruvate kinase (PKLR) gene. Am J Med Genet 1998; 78:233-236

Satoh H, Tani K, Yoshida MC, Sasaki M, Miwa S, Fujii H. The human liver-type pyruvate kinase (PKL) gene is on chromosome 1 at band q21. Cytogenet Cell Genet 1988; 47:132-133

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, Excoffier L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 1996; 76:377-383

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

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK. 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 1998; 62:1389-1402

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sajantila A, Lu R -b, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG. Short tandem-repeat polymorphism/*Alu* haplotype variation at the PLAT locus: implications for modern human origins. Am J Hum Genet 2000; 67:901-925

Volz A, Korge BP, Compton JG, Ziegler A, Steinert PM, Mischke D. Physical mapping of a functional cluster of epidermal differentiation genes on chromosome 1q21. Genomics 1993; 18:92-9

Vos HL, Mockensturm-Wilson M, Rood PM, Maas AM, Duhig T, Gendler SJ, Bornstein P. A tightly organized, conserved gene cluster on mouse chromosome 3 (E3-F1). Mamm Genome 1995; 6:820-2

Watkins WS, Zenger R, O'Brien E, Nyman D, Eriksson AW, Renlund M, Jorde LB. Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willenbrand factor region. Am J Hum Genet 1994; 55:348-355

Watterson GA and Guess HA. Is the most frequent allele the oldest? Theor pop biol 1977; 11:141-160

Winfield SL, Tayebi N, Martin BM, Ginns EI, Sidransky E. Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. Genome Research 1997; 7:1020-1026

Zimran A, Sorge J, Gross E, Kubitz M, West C, Beutler E. A glucocerebrosidase fusion gene in Gaucher disease. J Clin Invest 1990; 85:219-222

Table 1. PKLR and GBA polymorphisms allele frequencies by population. BIA=Biaka, MBU=Mbuti, TAN=Tanzanians, SAH=Saharawi, YEM=Yemenites, DRU=Druze, ADY=Adygei, RUS=Russians, FIN=Finns, CAT=Catalans, BAS=Basques, CHI=Chinese, JAP=Japanese, YAK=Yakut, NAS=Nasioi, MAY=Maya and SUR=Surui. Sample sizes in number of chromosomes are shown in brackets.

			BIA (138)	MBU (74)	TAN (80)	SAH (118)	YEM (86)	DRU (126)	ADY (104)	RUS (96)	FIN (68)	CAT (174)	BAS (188)	CHI (112)	JAP (92)	YAK (102)	NAS (46)	MAY (98)	SUR (90)
PKLR		7	0.019	0	0.013	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$(ATT)_n$		8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		9	0	0	0	0	0	0	0	0	0.015	0	0	0.060	0.045	0.225	0	0.044	0.035
		10	0.231	0.141	0.088	0.034	0	0	0	0	0	0.007	0	0	0	0.038	0	0	0
		11	0	0.313	0.025	0	0.013	0	0	0	0	0	0	0	0	0	0	0	0
		12	0.327	0.078	0.325	0.271	0.150	0.200	0.186	0.286	0.324	0.174	0.279	0.476	0.500	0.338	0.225	0.322	0.907
		13	0.212	0.266	0.125	0.042	0.038	0	0.058	0.057	0.044	0.083	0.076	0.048	0.125	0.025	0	0.033	0
		14	0.019	0.047	0.038	0.458	0.438	0.518	0.453	0.429	0.353	0.410	0.436	0.298	0.227	0.238	0.775	0.189	0.023
		15	0.087	0.031	0.138	0.136	0.250	0.218	0.128	0.100	0.132	0.236	0.128	0.083	0.091	0.125	0	0.278	0.023
		16	0.038	0.016	0.113	0.034	0.075	0.064	0.128	0.057	0.059	0.063	0.035	0.036	0.011	0	0	0.089	0
		17	0.067	0.109	0.138	0.025	0.013	0	0.047	0.043	0.074	0.021	0.047	0	0	0.013	0	0.044	0.012
		18	0	0	0	0	0	0	0	0.029	0	0.007	0	0	0	0	0	0	0
		19	0	0	0	0	0.025	0	0	0	0	0	0	0	0	0	0	0	0
PKLR	C	2*	0.312	0.311	0.363	0.619	0.849	0.865	0.779	0.726	0.632	0.759	0.670	0.447	0.315	0.345	0.848	0.594	0.067
1705	A	1	0.688	0.689	0.638	0.381	0.151	0.135	0.221	0.274	0.368	0.241	0.330	0.553	0.685	0.655	0.152	0.406	0.933
GBA	G	2	0.155	0.274	0.188	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2834	C	1	0.818	0.532	0.600	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	A	3	0.027	0.194	0.213	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GBA	A	2	0.130	0.500	0.400	0.552	0.756	0.857	0.798	0.667	0.636	0.718	0.585	0.357	0.272	0.373	0.652	0.598	0.056
3931	G	1	0.870	0.500	0.600	0.448	0.244	0.143	0.202	0.333	0.364	0.282	0.415	0.643	0.728	0.627	0.348	0.402	0.944
GBA	A	2	0.191	0.468	0.425	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5135	C	1	0.809	0.532	0.575	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GBA	A	2	0.179	0.455	0.400	0.569	0.756	0.857	0.808	0.662	0.621	0.709	0.582	0.357	0.272	0.375	0.636	0.691	0.056
6144	G	1	0.821	0.545	0.600	0.431	0.244	0.143	0.192	0.338	0.379	0.291	0.418	0.643	0.728	0.625	0.364	0.309	0.944

* Nomenclature used in the present study for the haplotype reconstruction

Table 2. Expected heterozygosity, by locus and haplotype

Population	PKLR $(ATT)_n$	PKLR 1705	GBA2 834	GBA3 931	GBA5 135	GBA6 144	Haplotype
Biaka	.80	.44	.33	.30	.33	.34	.905 / .914*
Mbuti	.82	.42	.61	.49	.51	.49	.918 / .932*
Tanzanians	.83	.47	.57	.49	.50	.49	.922 / .933*
Saharawi	.71	.48	-	.50	-	.49	.835
Yemenites	.72	.28	-	.37	-	.37	.783
Druze	.65	.26	-	.26	-	.28	.672
Adygei	.73	.37	-	.36	-	.35	.769
Russians	.72	.45	-	.48	-	.48	.778
Finns	.76	.48	-	.48	-	.47	.828
Catalans	.74	.37	-	.40	-	.41	.815
Basques	.72	.45	-	.49	-	.49	.809
Chinese	.68	.47	-	.47	-	.47	.701
Japanese	.68	.42	-	.39	-	.39	.716
Yakut	.77	.46	-	.48	-	.48	.817
Nasioi	.37	.23	-	.48	-	.48	.666
Maya	.78	.50	-	.49	-	.44	.863
Surui	.18	.13	-	.11	-	.11	.198

* 4 loci / 6 loci

Table 3. SNP haplotype frequencies by population. Sub-Saharan African frequencies are listed for three and five loci.

	Three loci:	FREQUENCY OF HAPLOTYPE							
	PKLR1705	1	1	1	1	2	2	2	2
	GBA3931	1	1	2	2	1	1	2	2
Population (2N)	GBA6144	1	2	1	2	1	2	1	2
Biaka (112)		.672	0	0	.042	.149	.027	0	.110
Mbuti (66)		.513	0	.015	.123	.017	0	0	.331
Tanzanians 80)		.520	.067	.014	.037	.013	0	.053	.297
Saharawi (114)		.386	0	0	0	.044	.018	0	.553
Yemenites (86)		.151	0	0	0	.093	0	0	.756
Druze (126)		.127	0	.008	0	.008	.008	0	.849
Adygei (104)		.192	0	0	.029	0	.010	0	.769
Russians (74)		.267	0	.016	.014	.043	.028	.012	.620
Finns (64)		.309	0	0	.066	.050	.016	0	.559
Catalans (156)		.214	0	.009	.021	.062	0	.004	.691
Basques (184)		.319	0	0	.012	.094	.006	.006	.564
Chinese (70)		.629	0	0	0	.014	0	0	.357
Japanese (92)		.685	0	0	0	.044	0	0	.272
Yakut (80)		.625	0	0	.038	0	0	0	.338
Nasioi (44)		.159	0	0	0	.205	0	0	.636
Maya (92)		.282	.086	.011	.045	.022	.012	0	.542
Surui (90)		.933	0	0	0	.011	0	0	.056

	Five loci:	FREQUENCY OF HAPLOTYPE							
	PKLR1705	1	1	1	1	1	1	1	1
	GBA2834	1	1	1	1	2	2	2	3
	GBA3931	1	1	1	2	1	2	2	2
	GBA5135	1	1	2	1	2	2	2	2
Population (2N)	GBA6144	1	2	1	1	1	1	2	2
Biaka (102)		.649	0	0	0	.012	0	.045	0
Mbuti (62)		.498	0	0	.016	0	0	.072	.059
Tanzanians (80)		.495	.067	.025	0	0	.014	.037	0

	Five loci:	FREQUENCY OF HAPLOTYPE							
	PKLR1705	2	2	2	2	2	2	2	2
	GBA2834	1	1	1	2	2	2	3	3
	GBA3931	1	1	1	1	2	2	2	2
	GBA5135	1	1	2	2	2	2	2	2
Population (2N)	GBA6144	1	2	1	1	1	2	1	2
Biaka (102)		.115	.029	.010	.018	0	.093	0	.029
Mbuti (62)		.018	0	0	0	0	.202	0	.134
Tanzanians (80)		.013	0	0	0	.040	.097	.013	.199

Table 4. D' values by pairs of SNP loci and by population. P values for all pairs are 0 except for pairs PKLR1705-GBA3931 and PKLR1705-GBA6144 in Nasioi where p=0.002.

Population	PKLR1705-GBA3931	PKLR1705-GBA6144	GBA3931-GBA6144	Average
Biaka	0.65	0.63	1	0.76
Mbuti	0.89	0.90	1	0.93
Tanzanians	0.94	0.71	0.74	0.80
Saharawi	1	1	1	1
Yemenites	1	1	1	1
Druze	0.93	1	1	0.98
Adygei	0.93	1	1	0.98
Russians	0.92	0.92	0.93	0.92
Finns	0.73	0.79	1	0.84
Catalans	0.87	0.86	1	0.91
Basques	0.94	0.94	0.98	0.95
Chinese	1	1	1	1
Japanese	1	1	1	1
Yakut	1	1	1	1
Nasioi	1	1	1	1
Maya	0.86	0.88	0.94	0.89
Surui	1	1	1	1
Average	0.92	0.92	0.98	

Table 5. LD studies between SNP pairs for a similar genetic distance.

References	Genetic distance	LD coefficient
Abecasis et al. (2001)	~75kb	D'= 0.20-0.30
Cox et al. (1998)	50-55kb	D'= 0.26-0.80
Dunning et al. (2000)	~50kb	D'= 0.15-0.20
Jorde et al. (1994)	0-150kb	D'= 0.55-1.0
Kidd et al. (2000)	65-74kb	D'= 0.24-1.0
Watkins et al. (1994)	78kb	r *= 0.03
Present study	~71kb	D'=0.92

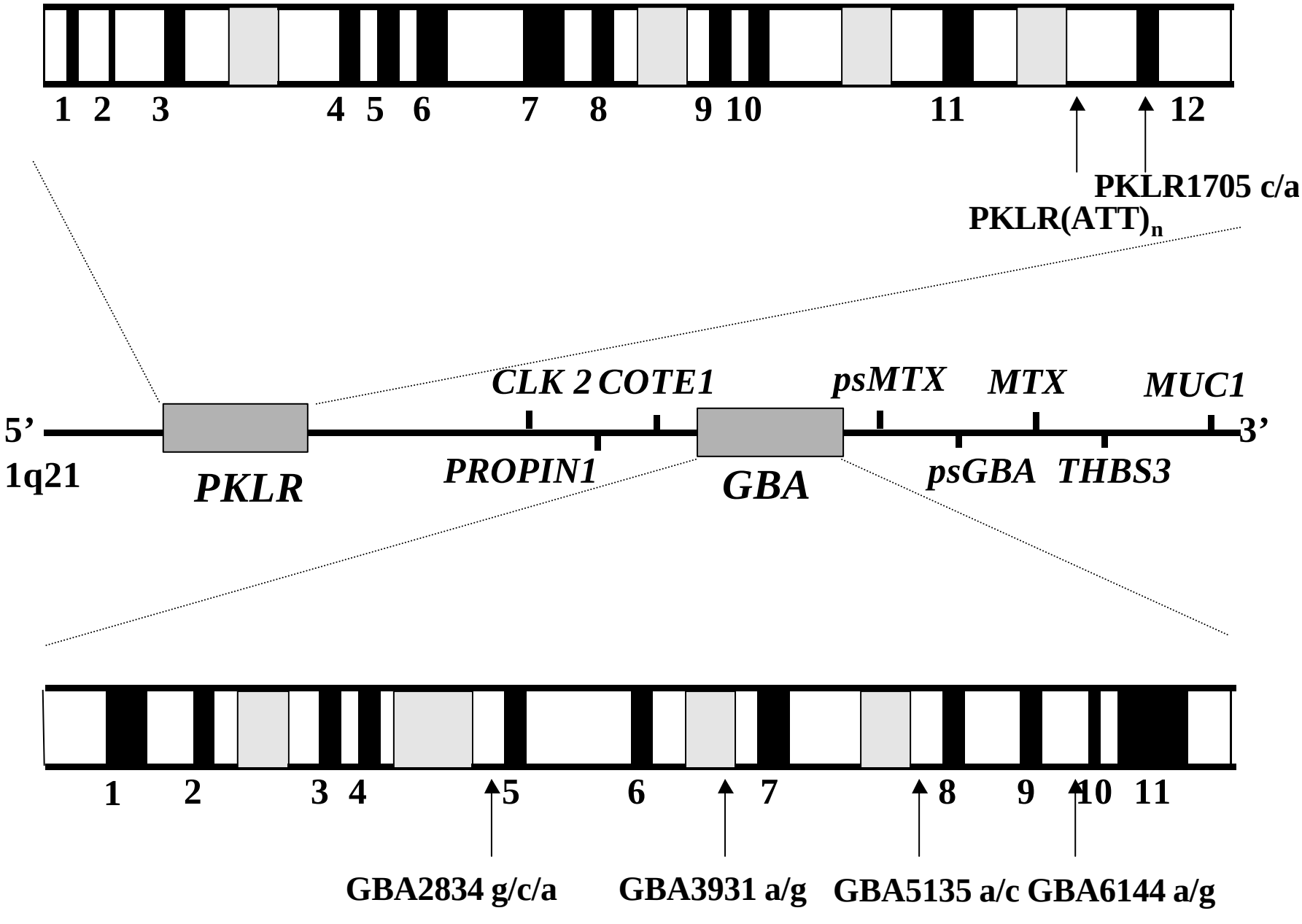
***** Hill and Robertson (1968) LD coefficient measure

FIGURE LEGENDS

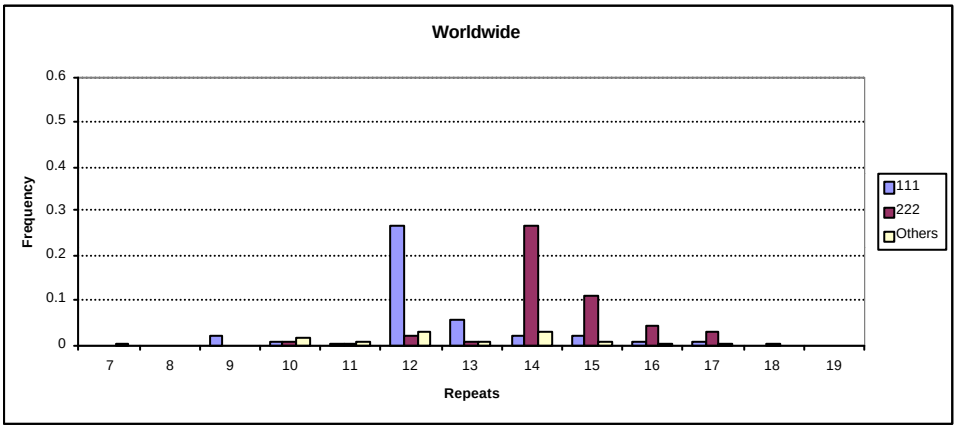
Figure 1. PKLR and GBA genes with all six polymorphic genetic markers studied: PKLR(ATT)$_n$, and PKLR1705 c/a in the PKLR gene and GBA2834 g/c/a, GBA3931 a/g, GBA5135 a/c and GBA6144 a/g in the GBA gene. Gene exons are denoted by numbers (1 to 12 in PKLR and 1 to 11 in GBA gene). Shaded boxes represent Alu insertions in both genes.

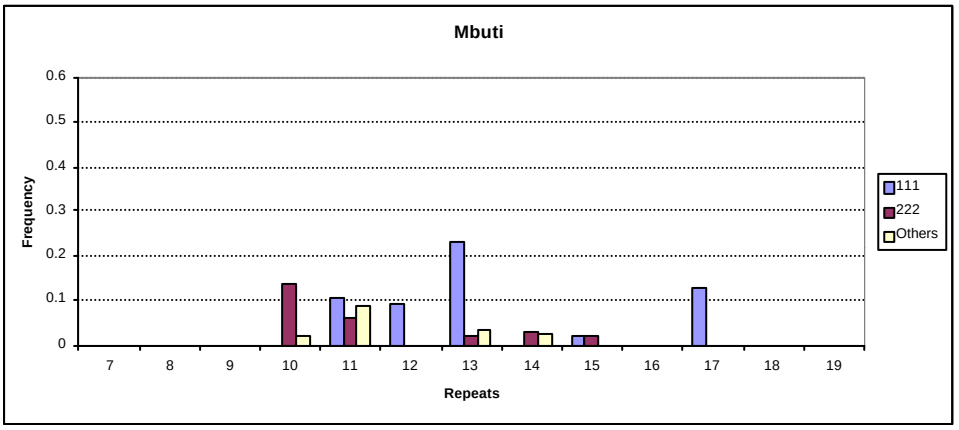
Figure 2. Geographic location of the populations sampled. ADY = Adygei; BAS = Basques; BIA = Biaka Pygmies; CAT = Catalans; CHI = Han Chinese; DRU = Druze; FIN = Finns; JPN = Japanese; MAY = Maya; MBU = Mbuti Pygmies; NAS = Nasioi; RUS = Russians; SAH = Saharawi; SUR = Surui; TAN = Tanzanians; YAK = Yakut; YEM = Yemenites. Additional information for all population samples can be obtained at the ALFRED database (http://info.med.yale.edu/genetics/kkidd).

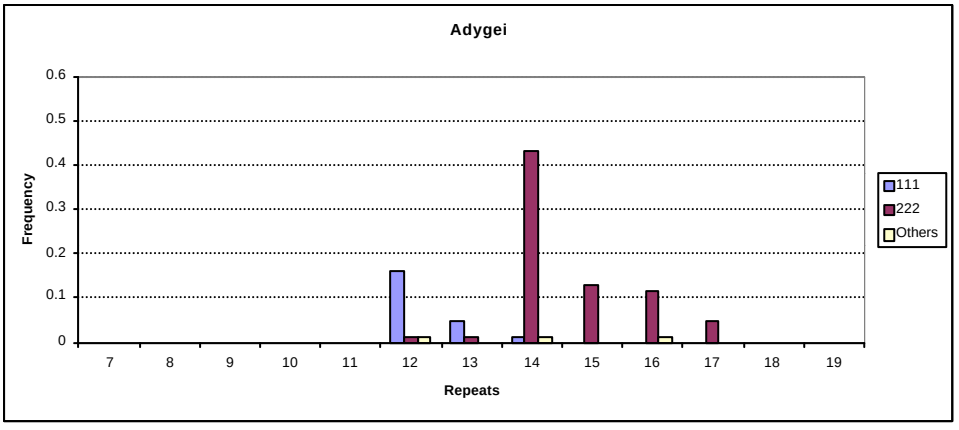
Figure 3. Worldwide and three population specific PKLR(ATT)$_n$ allele distributions in a different SNP haplotype backgrounds.

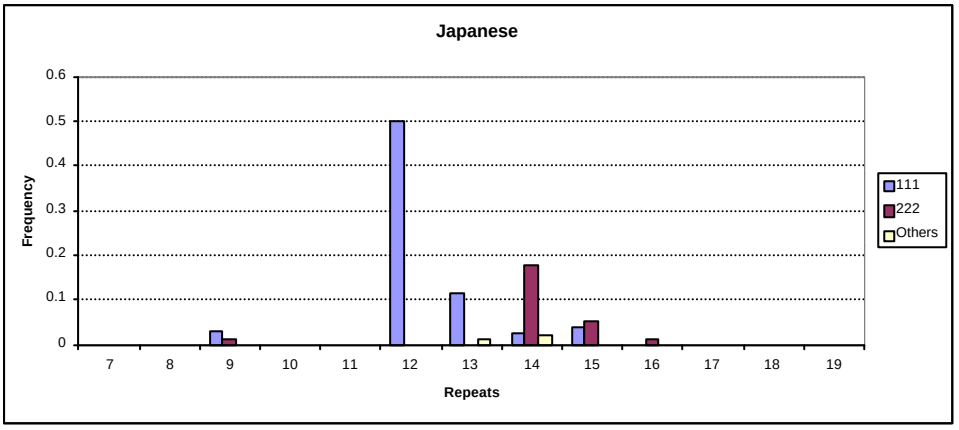

1 2 3 4 5 6 7 8 9 10 11 12
PKLR1705 c/a
PKLR(ATT)n
5'
3'
1q21
PKLR
CLK 2
COTE1
PROPIN1
GBA
psMTX
MTX
MUC1
psGBA
THBS3
1 2 3 4 5 6 7 8 9 10 11
GBA2834 g/c/a
GBA3931 a/g
GBA5135 a/c
GBA6144 a/g


FIN
YAK
RUS
BAS
CAT
ADY
DRU
JPN
CHI
SAH
YEM
MAY
MBU
BIA
NAS
TAN
SUR


Worldwide
Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0
7 8 9 10 11 12 13 14 15 16 17 18 19
Repeats
111
222
Others


Mbuti
Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0
7 8 9 10 11 12 13 14 15 16 17 18 19
Repeats
111
222
Others


Adygei
Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0
7 8 9 10 11 12 13 14 15 16 17 18 19
Repeats
111
222
Others


Japanese
Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0
7 8 9 10 11 12 13 14 15 16 17 18 19
Repeats
111
222
Others

DISCUSSIÓ

En aquest treball hem analitzat la variació genètica existent en dues regions autosòmiques humanes; la primera regió, localitzada a 7q31.2, inclou el gen CFTR, i la segona regió, localitzada a 1q21, inclou els gens GBA i PKLR. Els principals resultats del treball s'han preparat en forma d'articles, alguns dels quals ja han estat publicats, i en cadascun d'ells s'han discutit ja, independentment, els resultats corresponents.

A continuació discutirem els aspectes més rellevants d'aquest estudi i intentarem integrar els principals resultats obtinguts.

TIPATGE DEL MICROSATÈL·LIT IVS1CA DEL GEN CFTR

A partir de l'anàlisi de la variació al·lèlica existent en aquest nou microsatèl·lit del gen CFTR, descrit per primera vegada l'any 1997 per Moulin i col·laboradors, observem que és altament informatiu en totes les poblacions humanes on s'ha tipat, les quals comprenen una representació de poblacions que abasten tots els continents. Essent altament informatiu, aquest microsatèl·lit pot ser utilitzat per al consell genètic en la fibrosi quística, ja que ens pot aportar informació nova en estudis familiars on la mutació CF que afecta a una família en concret és desconeguda i pot confirmar la informació obtinguda amb els tres microsatèl·lits que es tipen rutinàriament en consell genètic (IVS8CA, IVS17bTA i IVS17bCA). En diagnòstic prenatal, com més informació s'obtingui, més fàcil serà detectar possibles contaminacions de teixit matern en el tipatge de teixit fetal, evitant errors de diagnosi.

L'estudi de la variació en un sol *locus* ens reafirma el model de reemplaçament o també anomenat com sortida d'Àfrica (*Out of Africa*) en l'evolució dels humans moderns, en mostrar-nos un major nombre d'al·lels i també els valors d'heterozigositat esperada més alts en les poblacions africanes analitzades que no pas en les poblacions no africanes. Aquests resultats són

semblants al panorama mitjà descrit per conjunts de microsatèl·lits (Pérez-Lezaun et al. 1997; Calafell et al. 1998, entre d'altres).

La inclusió d'un nou microsatèl·lit permet formar haplotips més llargs juntament amb altres marcadors ja coneguts, augmentant la resolució dels haplotips per estudis evolutius, tals com intentar entendre l'origen i dispersió de les mutacions CF més importants. Ens tornarem a referir a aquest punt més endavant en la discussió.

ANÀLISI GENÈTICA DE LA REGIÓ CFTR

Hem dut a terme l'anàlisi de sis polimorfismes (quatre microsatèl·lits i dues substitucions nucleotídiques) localitzats en el gen CFTR, en gairebé dos mil cromosomes d'individus no afectats de CF, en divuit poblacions humanes repartides per tot el món. L'estudi, en aquesta representació global de poblacions humanes, ens mostra importants diferències tant en les freqüències al·lèliques com en les freqüències haplotípiques entre les diferents poblacions. Hem de ressaltar, però, que així com les poblacions sud saharianes tenen una diversitat al·lèlica en els microsatèl·lits més gran comparades amb la resta de poblacions, les diversitats haplotípiques que presenten no són especialment majors.

El tipatge de les dues substitucions nucleotídiques (T854 i TUB20) en mostres de primats no humans ens ha permès inferir l'estat ancestral més probable de l'haplotip format per ambdues substitucions (1-2), que alhora coincideix amb l'haplotip més freqüent en la nostra mostra. Malgrat això, aquest haplotip no és el que porta associada la major diversitat d'haplotips pels quatre microsatèl·lits (IVS1CA-IVS6aGATT-IVS8CA-IVS17bTA) tal com seria d'esperar. Aquest fet recau en l'haplotip 2-2, suggerint que aquest haplotip derivat seria també molt antic.

Observant les variances de les distribucions al·lèliques dels diferents microsatèl·lits per poblacions, veiem que aquestes estan determinades més pel *locus* que per la població, existint molta heterogeneïtat en les diversitats dels

quatre microsatèl·lits, molt probablement degut a diferents taxes de mutació entre ells. El rang de polimorfisme trobat va des de dos al·lels que representen gairebé el 100% dels cromosomes pel *locus* IVS6aGATT fins a 36 al·lels diferents trobats en el *locus* IVS17bTA. Els microsatèl·lits emprats en el nostre estudi es coneixen a partir de la seqüència complerta del gen CFTR i a excepció d'un (IVS17bCA) n'hem tipat tota la resta, és a dir que són una mostra no esbiaixada de la representació de l'heterogeneïtat entre els microsatèl·lits en una determinada zona genòmica.

Si analitzem la variació en els microsatèl·lits depenent del *background* d'SNPs, mesurat amb l'AMOVA, veiem que la variabilitat dels microsatèl·lits és major entre els diferents *backgrounds* haplotípics que entre les poblacions, és a dir, que el *background* genètic predomina sobre el *background* poblacional en l'estructura de la variació genètica dels STRPs en la regió del gen CFTR, suggerint que les substitucions nucleotídiques que van generar el *background* haplotípic van precedir a la diferenciació de les poblacions humanes actuals. El valor d'F_{ST} per *background* d'SNPs més alt correspon al microsatèl·lit IVS6aGATT, confirmant la baixa taxa de mutació en aquest *locus* respecte els altres tres microsatèl·lits i no pas una taxa de mutació major seguida de constriccions en la mida dels al·lels. Aquesta és una visió original en l'estudi de la diversitat genètica humana i té implicacions en la comprensió de la variació des de dues dinàmiques complementàries: la del genoma i la de les poblacions. Moltes vegades s'ha tractat a les poblacions humanes com entitats clarament definides en les que els factors genètics, definits i teoritzant des de la genètica de poblacions, tenen lloc. De fet, però, sorgeix l'evidència de que la genealogia del gens (o regions del genoma) té arrels molt més profundes que la genealogia (o història) de les poblacions i per tant que les inferències a poblacions del passat a partir de les actuals pot estar mancada de sentit. Per gran part de la història genealògica del genoma no pot imbricar-s'hi la gènesi de les poblacions actuals que estarien només implicades en les branques més recents.

ANÀLISI GENÈTICA DE LA REGIÓ PKLR-GBA

Hem analitzat sis polimorfismes (un microsatèl·lit i cinc substitucions nucleotídiques) localitzats en aquesta regió. El microsatèl·lit i un SNP estan situats al gen PKLR i tota la resta de marcadors es troben al gen GBA. L'estudi s'ha dut a terme en ~1.800 cromosomes que corresponen a individus de disset poblacions de tot el món. Ja que un primer anàlisi, en totes les poblacions, de tres SNPs (incloent l'SNP localitzat al gen PKLR) i el microsatèl·lit ens va mostrar un fort desequilibri de lligament entre tots els marcadors en totes les poblacions, els dos SNPs restants van ser analitzats només en les poblacions sud saharianes. L'anàlisi de freqüències al·lèliques i haplotípiques ens mostra l'existència de dos haplotips majoritaris en totes les poblacions analitzades. Aquests corresponen als anteriorment anomenats haplotip + i haplotip – (Beutler et al. 1992b). L'haplotip + es troba a major freqüència a les poblacions africanes, asiàtiques i també en alguna població americana (Surui). L'haplotip – el trobem a major freqüència al nord d'Àfrica, a l'Orient Mitjà, Europa, entre els Nasioi de Melanèsia i els maies americans. Els següents haplotips més freqüents, i que es troben presents en gairebé totes les poblacions, són els que resulten de la recombinació entre el *locus* PKLR i el *locus* GBA.

Les freqüències haplotípiques que obtenim si combinem els genotips GBA/PKLR amb les dades de variació de seqüències de què disposem per al pseudogèn de GBA de Martínez-Arias i col·laboradors (2001, en premsa) ens mostren també la conservació dels dos haplotips majoritaris.

Una seqüència disponible del gen GBA en ximpanzé (Martínez-Arias et al. en premsa), presenta un haplotip que no correspon a cap dels dos haplotips majoritaris humans. Així, cap dels dos haplotips (+ i -) no sembla ser l'ancestral, malgrat que l'haplotip - sembla ser-ne el derivat més proper. A més, l'haplotip (format per quatre SNPs del gen GBA) que trobem al gen del ximpanzé no es troba a cap de les poblacions africanes analitzades. Si tenim en compte, però, els temps de divergència estimat entre els humans i els ximpanzés (~5 milions d'anys), aquest és més que suficient perquè un al·lel derivat pugui assolir altes

freqüències, i especialment si, com discutirem més endavant, la selecció positiva pot haver actuat intensament en aquesta regió.

La distribució al·lèlica del microsatèl·lit localitzat al gen PKLR està clarament condicionada pel *background* d'SNPs. Així, trobem una forta associació entre l'al·lel 12 del microsatèl·lit i el *background* + d'SNPs (o haplotip "111" en el nostre cas) i entre l'al·lel 14 del microsatèl·lit i el *background* - d'SNPs (o haplotip "222"). Per diferenciar haplotips recombinants entre el *locus* PKLR (que conté el microsatèl·lit i el primer SNP) i el *locus* GBA (que conté els dos SNPs restants) de processos mutacionals produïts sobre els haplotips + i – que ens donarien els mateixos haplotips, hem analitzat l'associació al·lèlica d'ambdós haplotips amb els al·lels 12 i 14 del microsatèl·lit de PKLR. Gairebé el 50% dels haplotips "211" porten associat l'al·lel 14 i igualment més de la meitat dels haplotips "122" els trobem associats a l'al·lel 12, mentre que són pocs els haplotips "211" que es troben associats a l'al·lel 12 i cap l'haplotip "122" associat a l'al·lel 14, comprovant que efectivament els processos de recombinació (entre l'SNP del gen PKLR i el primer SNP del gen GBA) són més probables que els de mutació.

L'acumulació al llarg del temps de variació genètica en el microsatèl·lit ens permet estimar el temps que ha transcorregut des d'un esdeveniment d'expansió. La variabilitat que trobem en aquest microsatèl·lit es pot haver acumulat en ~49.000 anys sobre el *background* d'SNPs + i en ~42.000 anys sobre l'haplotip -. Aquesta mesura pot ser interpretada com el temps transcorregut des que un episodi de selecció positiva tingué lloc en els nostres avantpassats. No coneixem de la naturalesa de l'episodi selectiu sobre quin gen o variant(s) concreta(es) pogué actuar però és una troballa indirecta de que quelcom de rellevant tingué lloc. Tot i que és una simple especulació, no se'ns escapa el fet que potser estem davant d'un esdeveniment genètic de gran relleu en la nostra evolució, sobretot perquè la selecció degué ser forta (fins eliminar la resta de variació) i perquè tingué lloc en un moment que pot haver estat crucial en l'expansió dels humans moderns.

LES EINES EMPRADES EN L'ESTUDI HAPLOTÍPIC I LA MESURA DEL DESEQUILIBRI DE LLIGAMENT

Els avantatges de la utilització d'haplotips en els estudis de genètica de poblacions són que permeten reconstruir la filogènia d'una regió genòmica, permeten combinar marcadors de diferent taxa evolutiva, com SNPs i STRPs, i permeten mesurar el desequilibri de lligament. Ara bé, els haplotips tenen com a inconvenient que la seva determinació implicaria tipar cada individu i els seus pares, que no solen estar disponibles. Això s'ha solucionat treballant amb el cromosoma X en homes (Laan i Pääbo 1997), o bé emprant mètodes d'estimació estadística de freqüències haplotípiques a partir dels genotips dels individus (Excoffier i Slatkin 1995; Hawley i Kidd 1995). Malgrat les reticències inicials, els mètodes d'estimació estadística han estat validats experimentalment (Tishkoff et al. 2000a), i els estudis que han emprat freqüències haplotípiques estimades han assolit un gran ressò (Tishkoff et al. 1996).

L'aplicació als nostres casos ha estat satisfactòria, especialment perquè el nombre d'al·lels a cada *locus* no era excessivament elevat, el que permetia que hi haguessin haplotips resolubles directament i que l'algoritme EM podia prendre com a base per a l'estimació. En general, les freqüències haplotípiques absolutes obtingudes s'apropaven a nombres enters. Per tant, aquests mètodes estadístics, que es troben a la base de tota l'anàlisi posterior dels resultats, han estat una eina decisiva per al nostre estudi i fou ja una aposta en el disseny original.

Un cop estimades les freqüències haplotípiques, podem mesurar el desequilibri de lligament entre els diferents *loci*. En l'estudi de la regió CFTR s'ha fet servir la mesura ξ de desequilibri de lligament introduïda per Zhao i col·laboradors (1999) que permet considerar marcadors multial·lèlics. En ser una mesura relativament nova i poc utilitzada en estudis de desequilibri de lligament, vàrem comprovar, satisfactòriament, que realment era una bona mesura del desequilibri de lligament entre marcadors tot comparant-la amb una de les mesures de desequilibri més utilitzades, com és D' (Lewontin, 1964), restringida a *loci* dialèl·lics. De fet ens hem trobat davant d'un problema general de la mesura

del LD: no hi ha una mesura que sigui comparable entre diferents tipus de marcador i aplicable per qualsevol nombre de polimorfismes. Les mesures més àmpliament acceptades, com D', poden emprar-se per dos marcadors dial·lèlics i llavors sí donen resultats comparables. Però si s'inclouen més marcadors o amb més al·lels (com els microsatèl·lits), el problema es complica. Fins el moment, el paràmetre que hem utilitzat , ξ, és l'únic que intenta una mesura general per una regió i per això hem fet diverses anàlisis per comprovar la seva adequació a un cas complex com és la regió CFTR.

Una altra mesura de desequilibri de lligament definida i emprada per primera vegada ha estat una variant de l'estadístic FE (*Fraction of Extra haplotypes*) introduït per Slatkin (2000), que hem anomenat FNF (*Fraction of haplotypes Not Found*), que quantifica la fracció dels possibles haplotips que no s'han trobat en la mostra analitzada, tot relacionant el nombre observat d'haplotips amb el nombre esperat sota equilibri de lligament i donades la grandària mostral i les freqüències al·lèliques. Així, el valor de FNF reflectirà el grau de desequilibri de lligament. Aquesta mesura l'hem feta servir també per l'estudi de la regió CFTR.

Ambdues mesures ens han permès estudiar el desequilibri de lligament en haplotips complexos (aquells que consten tant d'SNPs com STRPs altament polimòrfics).

A la regió PKLR-GBA, on hem analitzat el desequilibri de lligament entre parelles d'SNPs, la mesura emprada ha estat D' (Lewontin, 1964) que és una de les mesures estàndards de desequilibri de lligament per a *loci* dialèl·lics.

FACTORS GENÒMICS EN EL DESEQUILIBRI DE LLIGAMENT

Efectes de la distància física

S'acostuma a assumir que el desequilibri de lligament tendirà a disminuir com més allunyats estiguin físicament dos *loci* en un cromosoma; és a dir, que

existirà una correlació negativa entre desequilibri de lligament i distància física (Jorde, 2000). En el cas del *locus* CFTR, però, el patró de desequilibri de lligament entre els diferents marcadors és més complex, no essent una simple funció decreixent de la distància genètica. Ara bé, hem de tenir en compte, tal com discutirem al següent apartat, que els nostres marcadors inclouen tant STRPs com SNPs i que aquests estan situats alternament al llarg de CFTR.

A la regió PKLR-GBA, en canvi, on hem analitzat el LD entre parelles d'SNPs, s'observa un lleuger augment en el desequilibri de lligament entre una parella de *loci* situada dins del gen GBA respecte les parelles d'SNPs que comprenen marcadors d'ambdós gens i que, per tant, estan separades per una distància física més gran.

Efectes del nivell de polimorfisme

S'ha descrit que la diversitat al·lèlica pot contribuir al patró que observem de desequilibri entre *loci* (Slatkin, 1994; Ott i Rabinowitz, 1997; Sánchez-Mazas et al. 2000), fent que les parelles de *loci* amb major nombre d'al·lels (cas dels STRPs) presentin un desequilibri de lligament més fort, tal com és en el nostre cas, en el gen CFTR. Així, veiem que s'ha d'anar en compte amb segons quines combinacions de marcadors s'utilitzin per estudis de mesures de desequilibri de lligament, especialment si les combinacions inclouen totes les parelles possibles entre SNPs i STRPs, ja que podem veure augmentat el grau de desequilibri de lligament pel sol fet del tipus de marcador utilitzat en l'estudi. De la mateixa manera, la comparació del LD entre poblacions pot no ser adequada si hi ha grans diferències poblacionals en l'heterozigositat dels marcadors emprats.

Efectes de la taxa de recombinació

Tant la regió genòmica on està localitzat el gen CFTR com la regió centromèrica que comprèn els gens GBA i PKLR presenten taxes de recombinació per distància física (en cM/Mb) similars. Pel cas de la regió 7q31.2,

la taxa de recombinació entre els marcadors D7S666 i D7S2509 (regió on es troba CFTR segons dades del *Human Genome Project*) és de l'ordre de 0,23-0,24 cM/Mb (Payseur i Nachman, 2000), i per la regió 1q21, la taxa entre els marcadors D1S2624 i D1S2635 (regió que comprèn els gens GBA i PKLR) és de 0,25 cM/Mb (Payseur i Nachman, 2000). La taxa mitjana de recombinació per distància física al llarg del genoma està establerta en 1,22 cM/Mb (Venter et al. 2001), cinc vegades més gran que les taxes anteriors.

Les dues regions, per tant, posseeixen taxes de recombinació molt baixes, fet que es veu reflectit en el nivells de desequilibri de lligament detectats. Com més baixa és la taxa de recombinació entre *loci*, més lentament anirà disminuint el desequilibri de lligament al llarg de les generacions, i major serà el que detectarem avui en dia.

A la regió PKLR-GBA el desequilibri de lligament és molt fort entre totes les parelles de *loci* analitzades, independentment de la població, i s'observa clarament la conservació de només dos haplotips majoritaris (haplotips formats per un conjunt d'SNPs i un STRP). Malgrat això, la recombinació actua, i hem detectat haplotips que són resultat de la recombinació entre els dos haplotips majoritaris, essent els més freqüents els que provenen de la recombinació intergènica PKLR-GBA, i que per tant inclou els SNPs situats a major distància.

Al gen CFTR, el grau de desequilibri de lligament, que hem trobat entre els *loci* analitzats és més moderat, però no hem d'oblidar el fort desequilibri de lligament existent entre les mutacions causants de fibrosi quística i determinats marcadors o haplotips.

Efectes de la selecció

Per tal de crear i mantenir uns nivells alts de desequilibri de lligament és clar que es necessita una baixa taxa de recombinació a la zona. Això sol, però, pot no ser suficient i caldria invocar l'acció de la selecció per tal d'explicar com s'ha preservat el LD tan intens establert al llarg d'una regió del genoma. De fet, tot i que hi hagués poca recombinació, amb el pas de suficient temps es donarien les

suficients recombinacions per tal d'assolir equilibri i per tant sense LD. La selecció pot produir una gran pèrdua de variabilitat i genera LD si es selecciona positivament un determinat haplotip com a conseqüència d'un arrossegament genètic al seleccionar una variant d'un gen; és com si es posés el rellotge a zero i, amb el pas de les successives generacions, s'anirà reduint el LD.

A la regió PKLR-GBA (i més àmpliament a la regió 1q21), la selecció positiva hauria actuat intensament en algun *locus,* no identificat, de la regió i hauria produït una escombrada selectiva o arrossegament genètic (*hitchhiking, genetic sweep*) de tota una àmplia zona o conjunt de *loci*. La presència de dos haplotips majoritaris s'explicaria per dos processos d'escombrada selectiva consecutius o bé un únic procés d'arrossegament seguit d'una recombinació (Martínez-Arias et al., en preparació; article presentat a l'apartat Apèndix I). Ja hem comentat que l'anàlisi d'un microsatèl·lit en aquesta regió ens ha permès datar el temps que ha transcorregut des d'una expansió de població o bé des d'un procés de selecció. Així, el temps mesurat per la variació al·lèlica en el microsatèl·lit pot estar datant el final de l'esdeveniment(s) de selecció que van portar als dos haplotips majoritaris a les elevades freqüències que els trobem. Com que tots dos haplotips els trobem a elevades freqüències tant a les poblacions africanes com a les no africanes, podem especular que l'esdeveniment(s) de selecció va ser anterior a la sortida africana dels humans moderns. El fet que la variabilitat acumulada tant a l'haplotip + com al - sigui molt similar ens indica que els dos processos d'escombrada selectiva consecutius van ser molt propers en el temps o bé l'existència d'algun avantatge selectiva dels heterozigots +/-.

Cada vegada són més els autors que coincideixen en afirmar la presència d'una pressió selectiva que hauria afavorit als heterozigots CF i concretament als portadors de la mutació ΔF508 (Bertranpetit i Calafell 1996; Slatkin i Bertorelle, en premsa; Wiuf, en premsa), mutació que s'hauria seleccionat conjuntament amb el seu *background* haplotípic. El material del nostre estudi, però, consta de cromosomes no portadors de mutacions CF. I malgrat que ens mostra importants diferències tant en les freqüències al·lèliques com en les freqüències

haplotípiques entre les diferents poblacions, no esperem que aquestes diferències siguin modelades per les pressions selectives que teòricament afavoreixen als heterozigots CF, ja que gairebé el 98% dels cromosomes europeus (i el 100% en la majoria de continents) no són portadors de mutacions CF.

HISTÒRIA DE POBLACIONS I DESEQUILIBRI DE LLIGAMENT: FACTORS POBLACIONALS QUE HI INFLUEIXEN

S'han proposat dos models d'evolució dels humans anatòmicament moderns: el multiregional i l'*Out of Africa* (o model de reemplaçament). En el primer, s'afirma que els humans anatòmicament moderns haurien evolucionat en paral·lel a partir de les poblacions existents a Àfrica, Europa i Àsia, provinents de l'expansió dels *Homo erectus*, ara fa més d'un milió d'anys. En canvi, el model de l'*Out of Africa*, postula un origen africà i una expansió fora d'aquest continent ara fa uns 100.000 anys de l'*Homo sapiens*. Totes les poblacions humanes actuals no africanes serien descendents d'un *H. sapiens* ancestral que hauria evolucionat a l'Àfrica i que després de la sortida s'hauria expandit per tot el món, substituint altres poblacions del gènere *Homo* que encara eren presents fora d'Àfrica.

Són nombrosos els estudis genòmics que donen suport a aquest darrer model, en trobar molt poca variació i un fort desequilibri de lligament, en molts *loci* estudiats, en poblacions no africanes, a diferència de la forta variació i el poc desequilibri de lligament a les poblacions africanes (Kidd et al. 1998, 2000; Tishkoff et al. 1996, 1998, entre molts d'altres). L'efecte fundador hauria establert un patró de desequilibri de lligament, per als *loci* estudiats, que hauria arribat fins als nostres dies.

Tishkoff i col·laboradors (1996), estudiant un STRP i una inserció/deleció *Alu*, en el *locus* CD4, troben un major nombre d'haplotips en les poblacions africanes i un desequilibri de lligament gairebé total a les poblacions no africanes entre la deleció *Alu* i un al·lel concret del microsatèl·lit. Al contrari, a totes les poblacions africanes hi ha nivells molt baixos o absents de LD.

Tishkoff i col·laboradors (1998), analitzaren dos SNPs, un STRP i una inserció/deleció *Alu*, en el *locus* de la distròfia miotònica (DM), i trobaren una menor diversitat haplotípica a les poblacions no africanes, les quals també presenten alts valors de desequilibri de lligament. El desequilibri de lligament per parelles de marcadors dialèl·lics és fortament significatiu a totes les poblacions (a la Taula 4 es presenta una selecció d'aquests valors). Comparat amb les no africanes, les poblacions africanes tenen valors més baixos de D'.

Kidd i col·laboradors (1998) analitzaren tres SNPs i un STRP en el *locus* del receptor D2 de la dopamina (DRD2), i trobaren el mateix patró poblacional de desequilibri de lligament (disminució de la variació genètica tot sortint d'Àfrica) encara que la disminució de la variació no és tan dràstica com l'observada pels marcadors dels *loci* CD4 i DM (Taula 4).

Kidd i col·laboradors (2000), estudiant quatre SNPs en el *locus* de la hidroxilasa de la fenilalanina (PAH), donen suport també al model de l'Out of Africa, malgrat que el patró poblacional de desequilibri de lligament no és tan clar tampoc com als loci CD4 i DM (Taula 4).

A la regió PKLR-GBA, on hem estudiat tres SNPs, el patró de desequilibri de lligament entre marcadors observat per població segueix aquest patró general, encara que no de forma tan marcada: un desequilibri de lligament lleugerament inferior a les poblacions africanes i un increment fora d'Àfrica. Les poblacions africanes són també les que presenten una major diversitat tant al·lèlica com haplotípica, tal com es d'esperar sota el model de reemplaçament.

Ara bé, en els sis marcadors del *locus* CFTR, la quantitat de desequilibri de lligament en poblacions africanes no és molt diferent de la trobada en les no africanes, malgrat tenir en compte només el *background* haplotípic format pels SNPs. Sí que és cert que, considerant només els SNPs, les poblacions africanes mostren un menor desequilibri de lligament, però en afegir-hi els STRPs, i mesurant el LD amb l'estadístic FNF (per tal de corregir l'efecte de la diversitat en els STRPs sobre el LD), veiem que les poblacions africanes presenten valors similars de desequilibri del trobat a poblacions europees o asiàtiques.

Així, podem deduir que el desequilibri de lligament en les poblacions no africanes en relació a les africanes seguiria una àmplia distribució, on alguns *loci,* com CD4 (Tishkoff et al. 1996), mostren forts valors de LD només a les poblacions no africanes; d'altres com PKLR-GBA, mostren valors de desequilibri de lligament en les poblacions africanes només lleugerament menors als de les poblacions no africanes; mentre d'altres, com CFTR, presenten valors similars de LD tant a africans com a no africans.

La sortida d'Àfrica d'un grup prou nombrós dels primers humans anatòmicament moderns hauria portat a aquest patró del desequilibri de lligament en les diferents poblacions actuals. L'atzar hauria fet que la diversitat existent en diferents *loci* del grup fundador fos diferent: prenent tota la diversitat per alguns *loci*, la majoria de la diversitat per altres *loci* o molt poca en d'altres. Això faria que en alguns casos el desequilibri de lligament dins i fora d'Àfrica seria semblant; en altres, seria lleugerament més acusat fora d'Àfrica, i, quan l'efecte fundador hagués estat més intens, l'increment de LD fora d'Àfrica hauria esta dràstic. De totes maneres, en casos com la regió PKLR-GBA, no es pot descartar l'acció de la selecció que també hauria contribuït a la reducció de la variabilitat genètica observada, tot i que l'efecte de la selecció sembla que hauria tingut lloc molt antigament en el llinatge humà.

	LOCI								
	(kb)								
	Referència								
	DM	**DRD2**	**PAH**			**PKLR-GBA**		**CFTR**	
	(2,5)	(4,7)	(1,8)	(65)	(72)	(2,2)	(71)	(58)	(59)
	Tishkoff et al. 1998	Kidd et al. 1998	Kidd et al. 2000			Mateu et al. (en preparació)		Mateu et al. 2001	
Poblacions:									
Biaka	0,86	*1,00*	0,96	*0,20*	*0,13*	1,00	0,63	*0,01*	*0,12*
Mbuti	1,00	NA	0,92	0,42	*0,08*	1,00	0,90	1,00	*0,06*
Iemenites	0,96	*1,00*	0,82	*0,21*	0,29	1,00	1,00	0,62	0,45
Drusos	1,00	1,00	1,00	*0,06*	*0,01*	1,00	1,00	0,83	*0,06*
Adygei	ND	0,73	0,96	*0,05*	*0,00*	1,00	1,00	0,86	*0,24*
Russos	ND	ND	1,00	*0,13*	0,30	0,93	0,92	1,00	*0,11*
Finlandesos	ND	0,58	1,00	*0,30*	*0,13*	1,00	0,79	1,00	*0,12*
Xinesos	0,94	*0,59*	0,88	*0,69*	*0,61*	1,00	1,00	*1,00*	0,95
Japonesos	1,00	1,00	1,00	0,54	*0,17*	1,00	1,00	NA	0,88
Iacuts	1,00	*1,00*	0,93	0,31	0,34	1,00	1,00	1,00	0,74
Nasioi	0,94	1,00	1,00	1,00	*0,62*	1,00	1,00	*1,00*	1,00
Maia	1,00	*0,44*	0,89	0,85	0,88	0,94	0,88	1,00	0,95
Surui	1,00	*0,49*	0,88	1,00	0,91	1,00	1,00	NA	1,00

Taula 4. Desequilibri de lligament (en valor absolut de la mesura D') per parelles de marcadors en diferents *loci* i per població. Tots els marcadors són dial·lèlics i estant separats per distàncies genètiques relativament semblants a les que hi ha entre els marcadors dels *loci* PKLR-GBA (2,2 i 71 kb) i CFTR (58-59 kb) Si les distàncies genètiques eren diferents d'aquestes, els marcadors no han estat inclosos en la taula. Les distàncies són en kb i estan indicades en parèntesi. Les poblacions són les mateixes pels diferents estudis: només hem inclòs les poblacions en comú als diferents treballs. Els valors de D' no significatius (p>0,05) són en cursiva. NA: no aplicable per falta de variació en un o ambdós marcadors. ND: dades no disponibles.

ORIGEN DE LES MUTACIONS CF MÉS IMPORTANTS

La fibrosi quística és la malaltia autosòmica recessiva més comuna en les poblacions d'origen europeu, afectant un de cada 2.500 individus. Causada per mutacions al gen CFTR, només cinc de les gairebé 1.000 mutacions descrites presenten freqüències globals superiors a l'1% en els cromosomes CF (mutacions ΔF508, G542X, G551D, N1303K i W1282X). La mutació ΔF508 és la majoritària i compren el ~70% dels cromosomes CF.

Per tant, és més versemblant que cada mutació CF hagi sorgit en aquella població on el *background* haplotípic al que es troba associada sigui més freqüent entre els cromosomes normals. Cal tenir en compte, però, que algunes d'aquestes mutacions poden ser molt antigues (Morral et al. 1994; Bertranpetit i Calafell 1996) i, per tant, la composició actual de la població pot no reflectir la de la població on es va originar la mutació.

Al buscar aquests *backgrounds* haplotípics en les poblacions humanes actuals observem que són rars: només es troben esporàdicament en alguna població de l'Orient Mitjà, europea o asiàtica. Hem de ressaltar, però, que són els haplotips majoritaris associats específicament a la mutació ΔF508 (21-6-23-1-31-2 i 21-6-17-1-31-2) els que gairebé no es troben enlloc (només en una població de l'orient mitjà i en una d'europea); i és en incloure altres haplotips filogenèticament relacionats, que n'apareixen traces en més poblacions europees i una asiàtica.

Ja Morral i col·laboradors (1994) varen postular que la mutació ΔF508 hauria sorgit, fa més de 40.000 anys, en una població genèticament diferent a les poblacions europees actuals i que s'hauria estès per Europa durant el Paleolític. Altres autors han donat estimes més recents de l'edat d'aquesta mutació que van des de 3.000 fins a 34.000 anys (Serre et al. 1990; Wiuf, en premsa). Construint un arbre de màxima versemblança incloent cromosomes no portadors de mutacions CF (poblacions de tot el món) i cromosomes CF portadors de les cinc mutacions, veiem clarament que el principal factor estratificador és el *background* genòmic (el fet de portar mutacions CF o no) i no pas la població, fet que confirmaria una edat antiga, pre-Neolítica, de les mutacions.

Així doncs, no tindria sentit el fet de buscar un lloc d'origen d'unes mutacions que probablement són més antigues que les pròpies poblacions actuals i que ja estarien presents als humans abans del procés de formació dels grups humans actuals, és a dir l'etnogènesi i història demogràfica de la humanitat.

Aquest treball, l'estudi de la diversitat genètica en dues regions del genoma en poblacions humanes de tot el món, s'ha mostrat molt fructífer aportant una visió interessant sobre la dinàmica del genoma, la dinàmica de les poblacions i la interrelació entre totes dues. Cal un profund coneixement de la regió concreta del genoma que s'estudia per poder inferir conclusions poblacionals i cal, alhora, tenir en compte les poblacions que s'estudien per entendre la dinàmica d'una regió del nostre genoma.

Aquest estudi ajuda a entreveure la complexitat que té la diversitat del genoma: molt diferents comportaments de les diferents regions segons l'actuació de les diferents forces evolutives (mutació, selecció, recombinació...), fet que anirà mostrant el genoma no només com una seqüència on es troben els gens sinó com un compost heterogeni de regions amb dinàmiques ben diferenciades.

Per altra banda la genètica de poblacions humanes ha de prendre més en consideració la dinàmica de la regió del genoma que estudiï. No és suficient analitzar un conjunt de poblacions genèticament per tal d'inferir-ne la història genètica i demogràfica; els resultats poden ser molt heterogenis i diferents segons regions. És, doncs, en la compenetració dels processos evolutius dels genomes i les poblacions on assolirem una comprensió més acurada del procés únic de la nostra història genètica.

BIBLIOGRAFIA

A

Anderson DH (1938) Cystic fibrosis of the pancreas and its relation to the celiac disease: a clinical and pathological study. Am J Dis Child 56:344

B

Baronciani L, Beutler E (1993) Analysis of pyruvate kinase-deficiency mutations that produce nonspherocytic hemolytic anemia. Proc Natl Acad Sci USA 90:4324-7

Baronciani L, Beutler E (1995) Molecular study of pyruvate kinase deficient patients with hereditary nonspherocytic hemolytic anemia. J Clin Invest 95:1702-9

Baronciani L, Magalhaes IQ, Mahoney DH Jr, Westwood B, Adekile AD, Lappin TR, Beutler E (1995) Study of the molecular defects in pyruvate kinase deficient patients affected by nonspherocytic hemolytic anemia. Blood Cells Mol Dis 21:49-55

Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E (1996) Standardized nomenclature for Alu repeats. J Mol Evol 42:3-6

Bertranpetit J, Calafell F (1996) Genetic and geographical variability in cystic fibrosis: Evolutionary considerations. In: Gail Cardew (ed) Variation in the Human Genome, Ciba Foundation Symposium 197. Chichester, Wiley & Sons, pp 97-118

Beutler E, Gelbart T, Kuhl W, Sorge J, West C (1991) Identification of the second common Jewish Gaucher disease mutation makes possible population-based screening for the heterozygous state. Proc Natl Acad Sci USA 88:10544-7

Beutler E, Gelbart T, Kuhl W, Zimran A, West C (1992a) Mutations in Jewish patients with Gaucher disease. Blood 79:1662-6

Beutler E, West C, Gelbart T (1992b) Polymorphisms in the human glucocerebrosidase gene. Genomics 12:795-800

Beutler E (1993) Gaucher disease as a paradigm of current issues regarding single gene mutations of humans. Proc Natl Acad Sci USA 90:5384-5390

Beutler E, Nguyen NJ, Henneberger MW, Smolec JM, McPherson RA, West C, Gelbart T (1993) Gaucher disease: gene frequencies in the Ashkenazi Jewish population. Am J Hum Genet 52:85-8

Beutler E (1995) Gaucher disease. Adv Genet 32:17-49

Beutler E, Grabowski GA (1995) Gaucher disease. In: Scriver CR (ed) The metabolic and molecular bases of inherited disease 7^{th} ed. Mc Graw-Hill, pp 2641-2670

Beutler E, Baronciani L (1996) Mutations in pyruvate kinase. Hum Mutat 7:1-6

Beutler E, Gelbart T (1998) Hematologically important mutations: Gaucher disease. Blood Cells Mol Dis 24:2-8

Beutler E, Gelbart T (2000) Estimating the prevalence of pyruvate kinase deficiency from the gene frequency in the general white population. Bllod 95:3585-3588

Boas FE (2000) Linkage to gaucher mutations in the ashkenazi population: effect of drift on decay of linkage disequilibrium and evidence for heterozygote selection. Blood Cells Mol Dis 26:348-59

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994a) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455-7

Bowcock AM, Tomfohrde J, Weissenbach J, Bonne-Tamir B, St George-Hyslop P, Giagheddu M, Cavalli-Sforza LL, Farrer LA (1994b) Refining the position of

Wilson disease by linkage disequilibrium with polymorphic microsatellites. Am J Hum Genet 54:79-87

Brady RO, Murray GJ, Barton NW (1994) Modifying exogenous glucocerebrosidase for effective replacement therapy in Gaucher disease. J Inherit Metab Dis 17:510-9

Brinkmann B, Klintschar M, Neuhuber F, Hhne 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

C

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

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

Casals T, Nunes V, Palacio A, Gimenez J, Gaona A, Ibanez N, Morral N, Estivill X (1993) Cystic fibrosis in Spain: high frequency of mutation G542X in the Mediterranean coastal area. Hum Genet 91:66-70

Casals T, Gimenez J, Ramos MD, Nunes V, Estivill X (1996) Prenatal diagnosis of cystic fibrosis in a highly heterogeneous population. Prenat Diagn 16:215-222

Casals T, Ramos MD, Gimenez J, Larriba S, Nunes V, Estivill X (1997) High heterogeneity for cystic fibrosis in Spanish families: 75 mutations account for 90% of chromosomes. Hum Genet 101:365-70

Chabas A, Cormand B, Grinberg D, Burguera JM, Balcells S, Merino JL, Mate I, Sobrino JA, Gonzalez-Duarte R, Vilageliu L (1995) Unusual expression of Gaucher's disease: cardiovascular calcifications in three sibs homozygous for the D409H mutation. J Med Genet 32:740-2

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

Charrow J, Andersson HC, Kaplan P, Kolodny EH, Mistry P, Pastores G, Rosenbloom BE, Scott CR, Wappner RS, Weinreb NJ, Zimran A (2000) The Gaucher registry: demographics and disease characteristics of 1698 patients with Gaucher disease. Arch Intern Med 160:2835-43

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

Cutting GR, Kasch LM, Rosenstein BJ, Zielenski J, Tsui LC, Antonarakis SE, Kazazian HH Jr (1990) A cluster of cystic fibrosis mutations in the first nucleotide-binding fold of the cystic fibrosis conductance regulator protein. Nature 346:366-9

D

De la Chapelle A, Wright FA (1998) Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc Natl Acad Sci USA 95:12416-12423

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

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-322

Dahl N, Lagerstrom M, Erikson A, Pettersson U (1990) Gaucher disease type III (Norrbottnian type) is caused by a single mutation in exon 10 of the glucocerebrosidase gene. Am J Hum Genet 47:275-8

Dahl N, Hillborg PO, Olofsson A (1993) Gaucher disease (Norrbottnian type III): probable founders identified by genealogical and molecular studies. Hum Genet 92:513-5

Demina A, Boas E, Beutler E (1998) Structure and linkage relationships of the region containing the human L-type pyruvat kinase (PKLR) and glucocerebrosidase (GBA) genes. Hematopathol Mol Hematol 11:63-71

Diaz A, Montfort M, Cormand B, Zeng B, Pastores GM, Chabas A, Vilageliu L, Grinberg D (1999) Gaucher disease: the N370S mutation in Ashkenazi Jewish and Spanish patients has a common origin and arose several thousand years ago. Am J Hum Genet 64:1233-8

Diaz A, Montfort M, Cormand B, Zeng B, Pastores GM, Chabas A, Vilageliu L, Grinberg D (2000) On the age of the most prevalent Gaucher disease-causing mutation, N370S. Am J Hum Genet 66:2014-5

E

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

Ellegren H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet 24:400-402

Estivill X, Scambler PJ, Wainwright BJ, Hawley K, Frederik P, Schwartz M, Baiget M, Kere J, Williamson R, Farrall M (1987) Patterns of polymorphism and linkage disequilibrium for cystic fibrosis. Genomics 1:257-263

Estivill X, Morral N, Bertranpetit J (1994) Age of the ΔF508 cystic fibrosis mutation. Nat Genet 8:216-218

Estivill X, Morral N (1996) Evolution of cystic fibrosis alleles. In: Dodge JA (ed) Cystic Fibrosis-Current topics. Wiley & Sons, pp 141-164

Estivill X, Bancells C, Ramos C, the Biomed CF Mutation Analysis Consortium (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. Hum Mutat 10:135-54

Eto Y, Ida H (1999) Clinical and molecular characteristics of Japanese Gaucher disease. Neurochem Res 24:207-211

European Working Group on CF Genetics (EWGGCF) (1990) Gradient of distribution in Europe of the major CF mutation and of its associated haplotype. Hum Genet 85:436-445

Excoffier L, Slatkin M (1995) Maximum-likelihood estimates of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Eyal N, Wilder S, Horowitz M (1990) Prevalent and rare mutations among Gaucher patients. Gene 96:277-283

F

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr, Ellis MC, Fullan A, Hinton LM, Jones NL, Kimmel BE, Kronmal GS, Lauer P, Lee VK, Loeb DB, Mapa FA, McClelland E, Meyer NC, Mintier GA, Moeller N, Moore T, Morikang E, Prass CE, Quintana L, Starnes SM, Schatzman RC, Brunke KJ, Drayna DT, Risch NJ, Bacon BR, Wolff RK (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399-408

G

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 266:107-9

Gavrieli-Rorman E, Scheinker V, Grabowski GA (1992) Structure and evolution of the human prosaposin chromosomal gene. Genomics 13:312-318

Glenn D, Gelbart T, Beutler E (1994) Tight linkage of pyruvate kinase (PKLR) and glucocerebrosidase (GBA) genes. Hum Genet 93:635-8

Gordon D, Simonic I, Ott J (2000) Significant evidence for linkage disequilibrium over a 5-cM region among Africaners. Genomics 66:87-92

Grabowski GA (1997) Gaucher disease: gene frequencies and genotype/ phenotype correlations. Genet Test 1:5-12

Guggino SE (1994) Gates of Janus: cystic fibrosis and diarrhea. Trends Microbiol 2:91-4

Guggino SE (1999) Evolution of the ΔF508 CFTR mutation. Trends Microbiol 7:55-6

Guo S-W (1997) Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum Hered 47:301-314

H

Haardt M, Benharouga M, Lechardeur D, Kartner N, Lukacs GL (1999) C-terminal truncations destabilize the cystic fibrosis transmembrane conductance regulator without impairing its biogenesis. A novel class of mutation. J Biol Chem 274:21873-7

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

Higgins CF (1992) Cystic fibrosis transmembrane conductance regulator (CFTR). Br Med Bull 48:754-65

Högenauer C, Santa Ana CA, Porter JL, Millard M, Gelfand A, Rosenblatt RL, Prestidge CB, Fordtran JS (2000) Active intestinal chloride secretion in human carriers of cystic fibrosis mutations: An evaluation of the hypothesis that heterozygotes have subnormal active intestinal chloride secretion. Am J Hum Genet 67:1422-7

Hollander DH (1982) Etiogenesis of the European cystic fibrosis polymorphism: heterozygote advantage against venereal syphilis? Med Hypotheses 8:191-7

Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, Beutler E (1989) The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 4:87-96

Horowitz M, Zimran A (1994) Mutations causing Gaucher disease. Hum Mutat 3:1-11

Horowitz M, Pasmanik-Chor M, Borochowitz Z, Falik-Zaccai T, Heldmann K, Carmi R, Parvari R, Beit-Or H, Goldman B, Peleg L, Levy-Lahad E, Renbaum P, Legum S, Shomrat R, Yeger H, Benbenisti D, Navon R, Dror V, Shohat M, Magal N, Navot N, Eyal N (1998) Prevalence of glucocerebrosidase mutations in the Israeli Ashkenazi Jewish population. Hum Mutat 12:240-4

Hughes DJ, Hill AJ, Macek M Jr, Redmond AO, Nevin NC, Graham CA (1996) Mutation characterization of CFTR gene in 206 Northern Irish CF families: thirty mutations, including two novel, account for ~94% of CF chromosomes. Hum Mutat 8:340-7

I

Ida H, Rennert OM, Ito T, Maekawa K, Eto Y (1998) Type 1 Gaucher disease: phenotypic expression and natural history in Japanese patients. Blood Cells Mol Dis 24:73-81

Ida H, Rennert OM, Iwasawa K, Kobayashi M, Eto Y (1999) Clinical and genetic studies of Japanese homozygotes for the Gaucher disease L444P mutation. Hum Genet 105:120-126

Iwasawa K, Ida H, Eto Y (1997) Differences in origin of the 1448C mutation in patients with Gaucher disease. Acta Pediatr Jpn 39:451-453

J

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11-14

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435-1444

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

K

Kannai R, Elstein D, Weiler-Razell D, Zimran A (1994) The selective advantage of Gaucher's disease: TB or not TB? Isr J Med Sci. 30:911-2

Kanno H, Fujii H, Hirono A, Miwa S (1991) cDNA cloning of human R-type pyruvate kinase and identification of a single amino acid substitution (Thr384→ Met) affecting enzymatic stability in a pyruvate kinase variant (PK Tokyo) associated with hereditary hemolytic anemia. Proc Natl Acad Sci USA 88:8218-21

Kaplan NL, Lewis PO, Weir BS (1994) Age of the ΔF508 cystic fibrosis mutation. Nat Genet 8:216-8

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

Kerem BS, Zielenski J, Markiewicz D, Bozon D, Gazit E, Yahav J, Kennedy D, Riordan JR, Collins FS, Rommens JM, Tsui L-C (1990) Identification of mutations in regions corresponding to the two putative nucleotide (ATP)-binding folds of the cystic fibrosis gene. Proc Natl Acad Sci USA 87:8447-51

Kidd KK, Kidd JR (1996) A nuclear perspective on human evolution. In: Boyce AJ (ed) Molecular biology and human diversity. Cambridge University press, pp 242-264

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-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, Bonne-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

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139-144

Kugler W, Willaschek C, Holtz C, Ohlenbusch A, Laspe P, Krugener R, Muirhead H, Schroter W, Lakomek M (2000) Eight novel mutations and consequences on mRNA and protein level in pyruvate kinase-deficient patients with nonspherocytic hemolytic anemia. Hum Mutat 15:261-72

L

Laan M, Pääbo S (1997) Demographic history and linkage disequilibrium in human populations. Nat Genet 17:435-438

Latham TE, Grabowski G, Theophilus BDM, Smith FI (1990) Complex alleles of the acid β-glucosidase gene in Gaucher disease. Am J Hum Genet 47: 79-86

Lenzner C, Nürberg P, Jacobasch G, Thiele B-J (1997) Complete genomic sequence of the human PK-L/R-Gene includes four intragenic polymorphisms defining different haplotype backgrounds of normal and mutant PK-Genes. DNA seq 8:45-53

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

Lonjou C, Collins A, Morton NE (1999) Allelic association between marker loci. Proc Natl Acad Sci USA 96:1621-1626

M

MacDonald ME, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, Taylor S, Allitto B, Altherr M, Myers R, Lehrach H, Collins FS, Wasmuth JJ, frontali M, Gusella JF (1992) The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet 1:99-103

Martínez-Arias R, Comas D, Mateu E, Bertranpetit J (2001) Glucocerebrosidase pseudogene variation and Gaucher disease: recognising pseudogene tracts in GBA alleles. Hum Mutat 17:191-198

Martínez-Arias R, Calafell F, Buchanan A, Weiss KW, Mateu E, Comas D, Andrés A, Bertranpetit J (2001) Sequence variability of a human pseudogene. Genome Res (en premsa)

Martínez-Arias R, Calafell D, Mateu E, Bertranpetit J. Profiles of accepted mutation: from neutrality in a pseudogene to disease-causing mutation on its homologous gene. Hum Genet (en premsa)

Martínez-Arias R, De Lorenzo D, Calafell F, Mateu E, Bertranpetit J. Selection shaped variability on a human pseudogene (en preparació)

Meindl RS (1987) Hypothesis: a selective advantage for cystic fibrosis heterozygotes. Am J Phys Anthropol 74:39-45

Mennie M, Gilfillan A, Brock DJ, Liston WA (1995) Heterozygotes for the ΔF508 cystic fibrosis allele are not protected against bronchial asthma. Nat Med 1:978-9

Messaoud T, Verlingue C, Denamur E, Pascaud O, Quéré I, Fattoum S, Elion J, Férec C (1996) Distribution of CFTR mutations in cystic fibrosis patients of Tunisian origin: identification of two novel mutations. Eur J Hum Genet 4:20-4

Mickle JE, Cutting GR (2000) Genotype-phenotype relationships in cystic fibrosis. Med Clin North Am 84:597-607

Morral N, Nunes V, Casals T, Chillón M, Giménez J, Bertranpetit J, Estivill X (1993) Microsatellite haplotypes for cystic fibrosis: mutation frameworks and evolutionary tracers. Hum Mol Genet 2:1015-22

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Giménez J, Reis A Varon-Mateeva R, Macek Jr. M, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo MP, Garnerone S, Restagno G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwarz M, Dallapiccola B, Novelli G, Ferec C, de Arce M, Nemeti M, Kere J, Anvret M, Dahl N, Kadasi L (1994a) The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nat Genet 7:169-175

Morral N, Llevadot R, Casals T, Gasparini P, Macek M Jr, Dörk T, Estivill X (1994b) Independent origins of cystic fibrosis mutations R334W, R347P, R1162X, and 3849 +10kbC→T provide evidence of mutation recurrence in the CFTR gene. Am J Hum Genet 55:890-8

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

Motulsky AG (1995) Jewish diseases and origins. Nat Genet 9:99-101

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

Myles-Worsley M, Coon H, Tiobech J, Collier J, Dale P, Wender P, Reimherr F, Polloi A, Byerley W (1999) Genetic epidemiological study of schizophrenia in Palau, Micronesia: prevalence and familiality. Am J Med Genet 88:4-10

O

Osborne L, Knight R, Santis G, Hodson M (1991) A mutation in the second nucleotide binding fold of the cystic fibrosis gene. Am J Hum Genet 48:608-12

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

P

Payseur BA, Nachman MW (2000) Microsatellite variation and recombination rate in the human genome. Genetics 156:1285-98

Peleg L, Frisch A, Goldman B, Karpaty M, Narinsky R, Bronstein S, Frydman M (1998) Lower frequency of Gaucher disease carriers among Tay-Sachs disease carriers. Eur J Hum Genet 6:185-186

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

Peterson AC, Di Rienzo A, Lehesjoki A-E, de la Chapelle A, Slatkin M, Freimer NB (1995) The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet 4:887-894

Pier GB, Grout M, Zaidi TS, Olsen JC, Johnson LG, Yankaskas JR, Goldberg JB (1996) Role of mutant CFTR in hypersusceptibility of cystic fibrosis patients to lung infections. Science 271:64-7

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

Pier GB (1999) Evolution of the ΔF508 CFTR mutation: response. Trends Microbiol 7:56-58

Pier GB (2000) Role of the cystic fibrosis transmembrane conductance regulator in innate immunity to Pseudomonas aeruginosa infections. Proc Natl Acad Sci USA 97:8822-8

R

Reiner O, Wigderson M, Horowitz M (1988) Structural analysis of the human glucocerebrosidase genes DNA 7:107-116

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, Breakefield X, Bressman S (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152-159

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 245:1066-73

Rockah R, Narinsky R, Frydman M, Cohen IJ, Zaizov R, Weizman A, Frisch A (1998) Linkage disequilibrium of common Gaucher disease mutations with a polymorphic site in the pyruvate kinase (PKLR) gene. Am J Med Genet 78:233-6

Romeo G, Devoto M, Galietta LJV (1989) Why is the cystic fibrosis gene so frequent? Hum Genet 84:1-5

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 245:1059-65

S

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning, 2nd ed. Cold Spring Harbor Laboratory Press, New York

Sanchez-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

Schroeder SA, Gaughan DM, Swift M (1995) Protection against bronchial asthma by CFTR ΔF508 mutation: a heterozygote advantage in cystic fibrosis. Nat Med 1:703-5

Schwartz M, Johansen HK, Koch C, Brandt NJ (1990) Frequency of the ΔF508 mutation on cystic fibrosis chromosomes in Denmark. Hum Genet 85:427-428

Schwiebert EM, Benos DJ, Fuller CM (1998) Cystic fibrosis: a multiple exocrinopathy caused by dysfunctions in a multifunctional transport protein. Am J Med 104:576-90

Schwiebert EM, Benos DJ, Egan ME, Stutts MJ, Guggino WB (1999) CFTR is a conductance regulator as well as a chloride channel. Physiol Rev 79 (1 Suppl):S145-66

Serre JL, Simon-Bouy B, Mornet E, Jaume-Roig B, Balassopoulou A, Schwartz M, Taillandier A, Boue J, Boue A (1990) Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. Hum Genet 84:449-54

Sheppard DN, Welsh MJ (1999) Structure and function of the CFTR chloride channel. Physiol Rev 79 (1 Suppl):S23-45

Shier WT (1979) Increased resistance to influenza as a possible source of heterozygote advantage in cystic fibrosis. Med Hypotheses 5:661-8

Shoshani T, Augarten A, Gazit E, Bashan N, Yahav Y, Rivlin Y, Tal A, Seret H, Yaar L, Kerem E, Kerem B (1992) Association of a nonsense mutation (W1282X), the most common mutation in the Ashkenazi Jewish cystic fibrosis patients in Israel, with presentation of severe disease. Am J Hum Genet 50:222-8

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

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

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

Slatkin M, Bertorelle G. The use of intra-allelic variability for testing neutrality and estimating population growth rate. Genetics (en premsa)

Sorge, J, West, C, Westwood, B, Beutler, E (1985a) Molecular cloning and nucleotide sequence of human glucocerebrosidase cDNA. Proc Natl Acad Sci USA 82:7289-7293

Sorge J, Gelbart T, West C, Westwood B, Beutler E (1985b) Heterogeneity in type I Gaucher disease demonstrated by restriction mapping of the gene. Proc Natl Acad Sci U S A 82:5442-5

Sorge, J, Gross, E, West, C, Beutler, B (1990) High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. J Clin Invest 86**:** 1137—1141

Southern KW (1997) ΔF508 in cystic fibrosis: willing but not able. Arch Dis Child 76:278-282

Strasberg PM, Triggs-Raine BL, Warren IB, Skomorowski M-A, McInnes B, Becker LE, Callahan JW, Clarke JTR (1994) Genotype-phenotype pitfalls in Gaucher disease. J Clin Lab Anal 8:228-236

T

Tanaka KR, Paglia DE (1995) Pyruvate kinase and other enzymopathies of the erythrocyte. In: Scriver CR (ed) The metabolic and molecular bases of inherited disease 7th ed. Mc Graw-Hill, pp 3485-3511

Tani K, Fujii H, Tsutsumi H, Sukegawa J, Toyoshima K, Yoshida MC, Noguchi T, Tanaka T, Miwa S (1987) Human liver type pyruvate kinase: cDNA cloning and chromosomal assignment. Biochim Biophys Res Commun 143:431-438

Tani K, Fujii H, Nagata S, Miwa S (1988) Human liver type pyruvate kinase: complete amino acid sequence and the expression in mammalian cells. Proc Natl Acad Sci USA 85:1792-1795

Tayebi N, Reissner K, Lau E, Stubblefield B, Klineburgess A, Martin B, Sidransky E (1998) Genotypic heterogeneity and phenotypic variation anomg patients with Type 2 Gaucher's disease. Pediatr Res 43:571-578

The Cystic Fibrosis Genetic Analysis Consortium (TCFGAC) (1994) Population variation of common cystic fibrosis mutations. Hum Mutat 4:167-177

Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 9:578-594

Terwilliger JD, Zöllner S, Laan M, Pääbo S (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. Hum Hered 48:138-154

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Cheung K, Kidd JR, Bonne-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, Bonne-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 (2000a) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67: 518-522

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sanjantila A, Lu R, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG (2000b) Short tandem-repeat Polymorphism/Alu haplotype variation at the PLAT locus: implications for modern human origins. Am J Hum Genet 67:901-25

Tsuji S, Choudary PV, Martin BM, Stubblefield BK, Mayor JA, Barranger JA, Ginns EI (1987) A mutation in the human glucocerebrosidase gene in neuronopathic Gaucher's disease. N Eng J Med 316:570-575

Tsuji S, Martin BM, Barranger JA, Stubblefield BK, LaMarca ME, Ginns EI (1988) Genetic heterogeneity in type 1 Gaucher disease: multiple genotypes in Ashkenazic and non-Ashkenazic individuals. Proc Natl Acad Sci USA 85:2349-2352

V

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO et al. (2001) The sequence of the human genome. Science 291:1304-51

Vidaud M, Fanen P, Martin J, Ghanem N, Nicolas S, Goossens M (1990) Three point mutations in the CFTR gene in French cystic fibrosis patients: identification by denaturing gradient gel electrophoresis. Hum Genet 85:446-9

Vos HL, Mockensturm-Wilson M, Rood PML, Maas AMCE, Duhig T, Gendler SJ, Bornstein P (1995) A tightly organized, conserved gene cluster on mouse chromosome 3 (E3-F1) Mamm Genome 6:820-822

W

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123-1128

Welsh MJ, Tsui L-C, Boat TF, Beaudet AL (1995) Cystic Fibrosis. In: Scriver CR (ed) The metabolic and molecular bases of inherited disease 7th ed. Mc Graw-Hill, pp 3799-3876

Winfield SL, Tayebi N, Martin BM, Ginns EI, Sidransky E (1997) Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. Genome Res 7:1020-1026

Wiuf C. Do ΔF508 heterozygotes have a selective advantage?. Genetical Res (en premsa)

Z

Zamel N, McClean PA, Sandell PR, Siminovitch KA, Slutsky AS (1996) Asthma on Tristan da Cunha: looking for the genetic link. The University of Toronto Genetics of Asthma Research Group. Am J Respir Crit Care Med 153:1902-6

Zanella A, Bianchi P (2000) Red cell pyruvate kinase deficiency: from genetics to clinical manifestations. Baillieres Best Pract Res Clin Haematol 13:57-81

Zapata C, Visedo G (1995) Gametic disequilibrium and physical distance. Am J Hum Genet 57:190-191

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, Rozmahel R, Bozon D, Kerem B, Grzelczak Z, Riordan JR, Rommens J, Tsui LC (1991) Genomic DNA sequence of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Genomics 10:214-28

Zielenski J, Tsui L-C (1995) Cystic fibrosis: genotypic and phenotypic variations. Annu Rev Genet 29:777-807

Zielenski J, Corey M, Rozmahel R, Markiewicz D, Aznarez I, Casals T, Larriba S, Mercier B, Cutting GR, Krebsova A, Macek M Jr, Langfelder-Schwind E, Marshall BC, DeCelie-Germana J, Claustres M, Palacio A, Bal J, Nowakowska A, Ferec C, Estivill X, Durie P, Tsui LC (1999) Detection of a cystic fibrosis modifier locus for meconium ileus on human chromosome 19q13. Nat Genet 22:128-9

Zielenski J (2000) Genotype and phenotype in cystic fibrosis. Respiration 67:117-133

Zimmer KP, le Coutre P, Aerts HM, Harzer K, Fukuda M, O'Brien JS, Naim HY Intracellular transport of acid beta-glucosidase and lysosome-associated membrane proteins is affected in Gaucher's disease (G202R mutation) (1999) J Pathol 188:407-14

Zimran A, Sorge J, Gross E, Kubitz M, West C, Beutler E (1990) A glucocerebrosidase fusion gene in Gaucher disease. J Clin Invest 85:219-222

ADRECES ELECTRÒNIQUES D’INTERÈS

Les adreces electròniques emprades en aquest treball han estat les següents:

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

Expasy, *http://www.expasy.ch/enzyme/*

CFTR (EC 3.6.3.49)
GBA (EC 3.2.1.45)
PKLR (EC 2.7.1.40)

GenBank, *http://www.ncbi.nlm.nih.gov/Genbank*

CFTR (#AC00011, #AC000061)
GBA (#J03059, #M11080)
PKLR (#U47654; #AB015983, #M15465)
psGBA (#J03060)
The Human Genome Project , *http://www.ncbi.nlm.nih.gov/genome/guide/human/*

Online Mendelian Inheritance in Man (OMIM), *http://www.ncbi.nlm.nih.gov/Omim*

CBAVD (MIM 277180)
CFTR (MIM 602421)
Deficiència de PK (MIM 266200)
Fibrosi quística (MIM 219700)
Malaltia de Gaucher tipus 1 (MIM 230800), tipus 2 (MIM 230900), tipus 3 (MIM 231000)

The Human Gene Mutation Database (HGMD), *http://www.uwcm.ac.uk*

GBA (#119262)
PKLR (#120294)

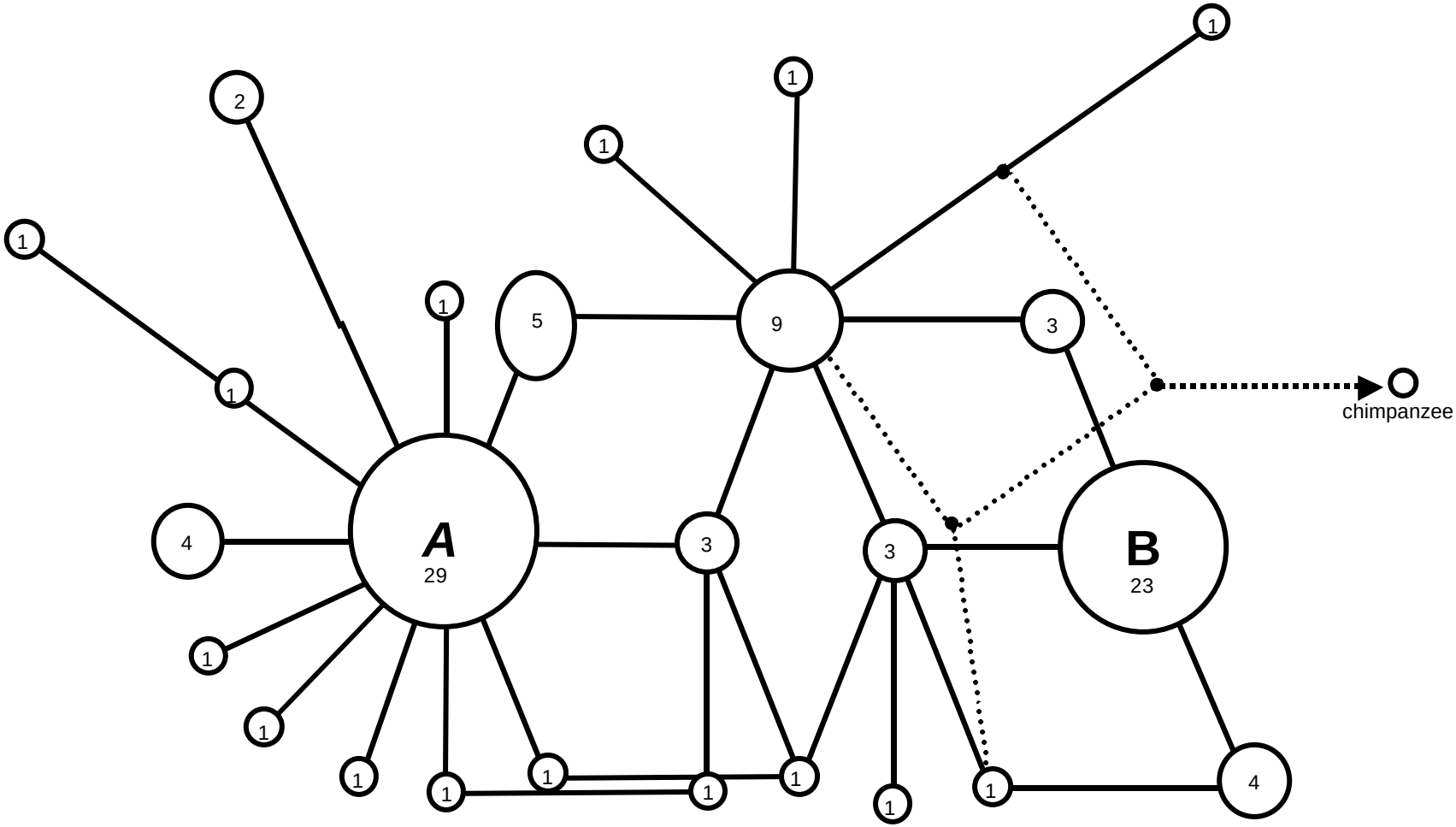
APÈNDIX I

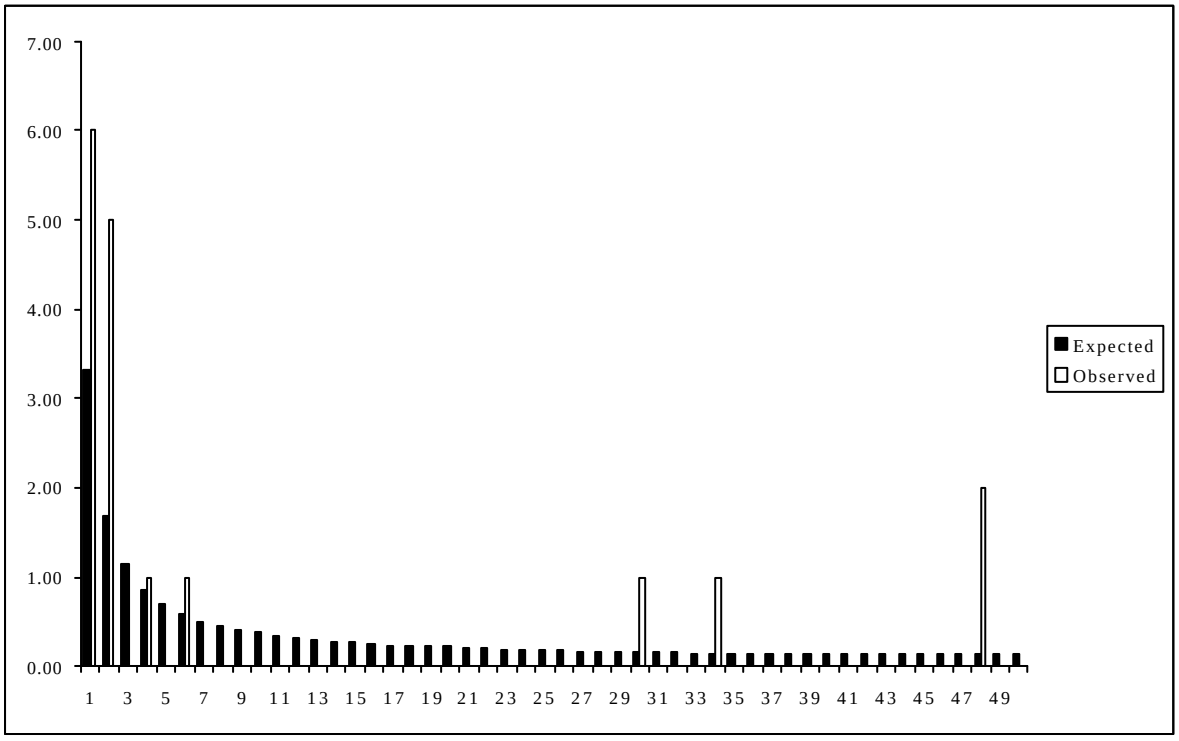
Un treball complementari al de l'estudi del patró de desequilibri de lligament que afecta la regió 1q21, i que comprèn els gens GBA i PKLR, ha estat el realitzat per Rosa Martínez-Arias i col·laboradors, en el seu estudi del pseudogèn de GBA. L'article que presentem a continuació tracta de la possible acció de la selecció en aquesta zona del genoma i inclou les dades que són objecte de la present tesi doctoral.

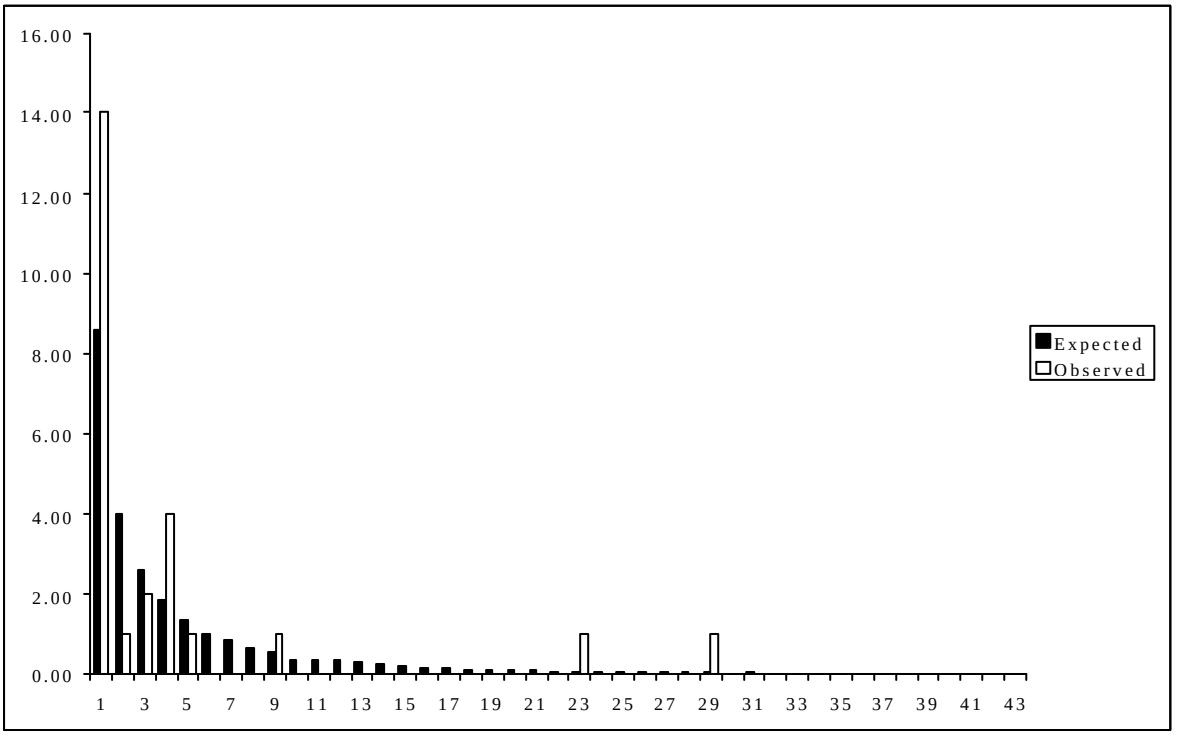
Selection shaping variability on a human pseudogene

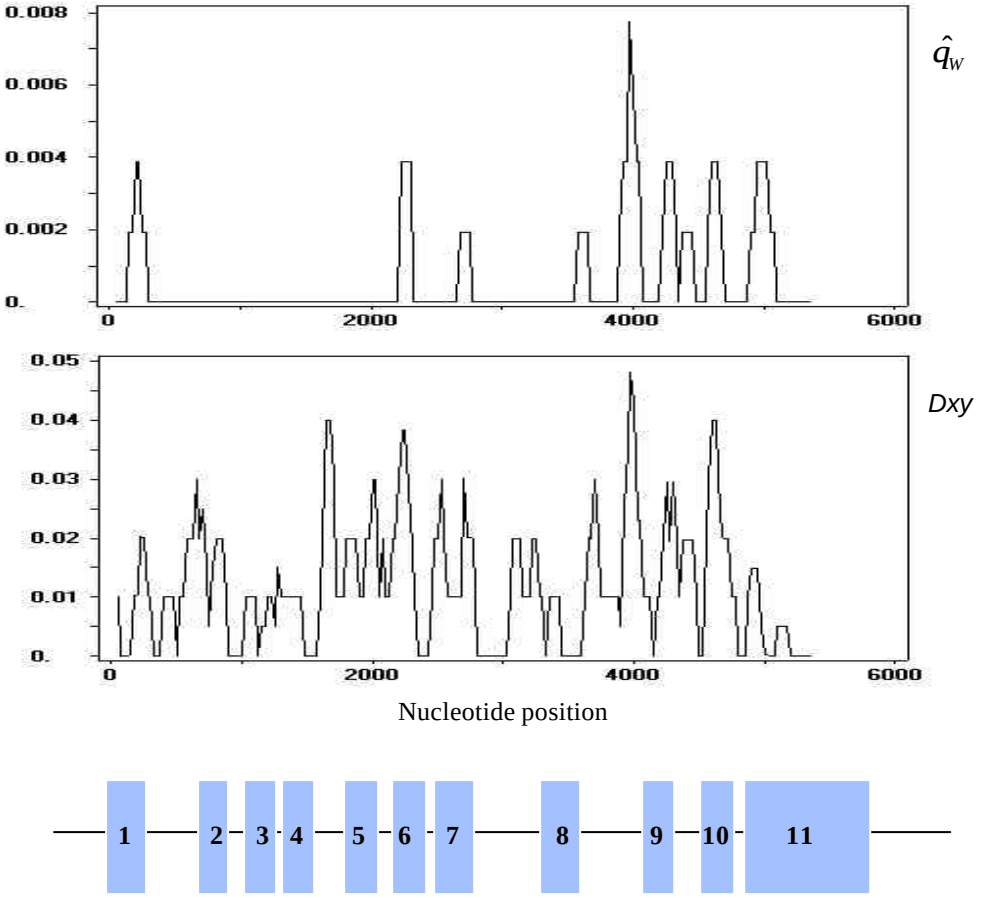
Rosa Martínez-Arias, David De Lorenzo, Eva Mateu,
Francesc Calafell, Jaume Bertranpetit

(manuscrit en preparació)

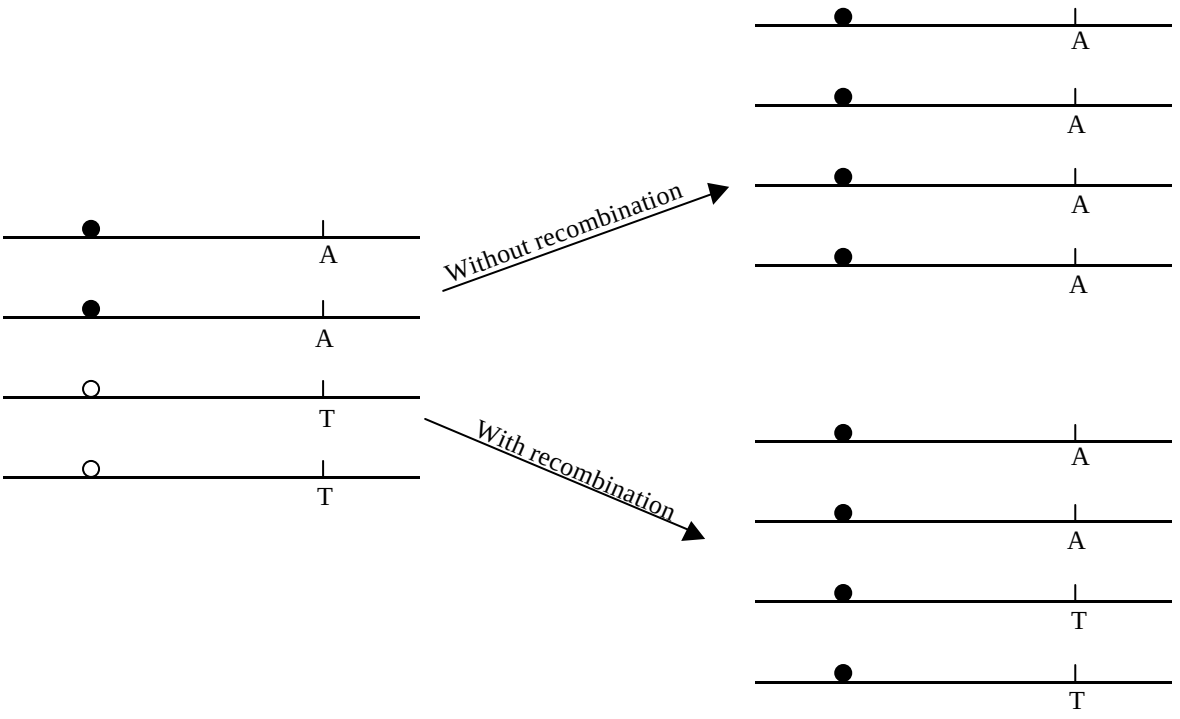

A
29
B
23
chimpanzee

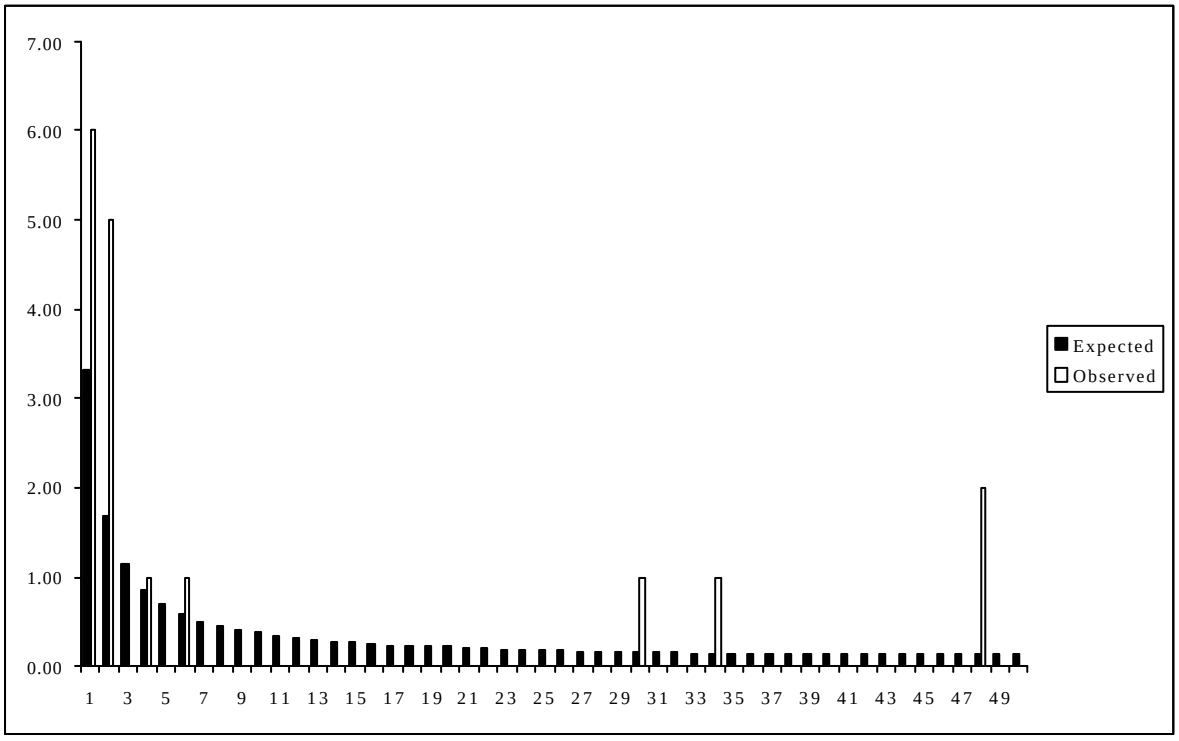

7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Expected
Observed

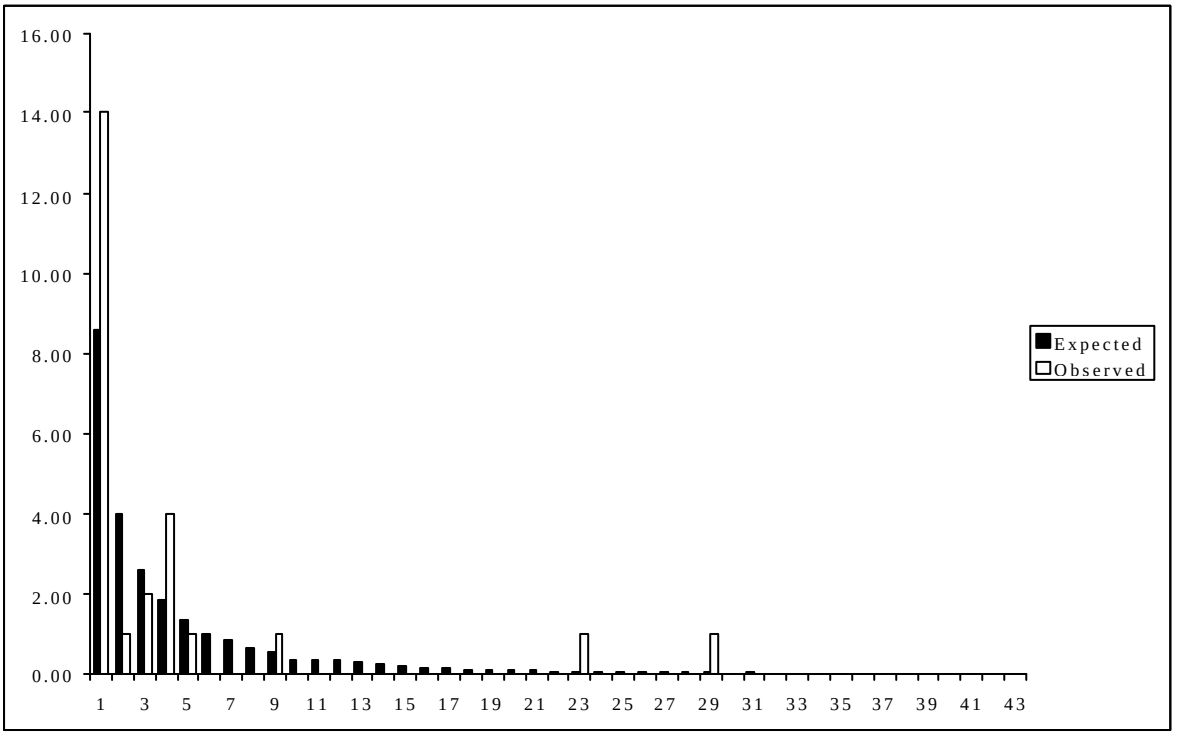

16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43
Expected
Observed

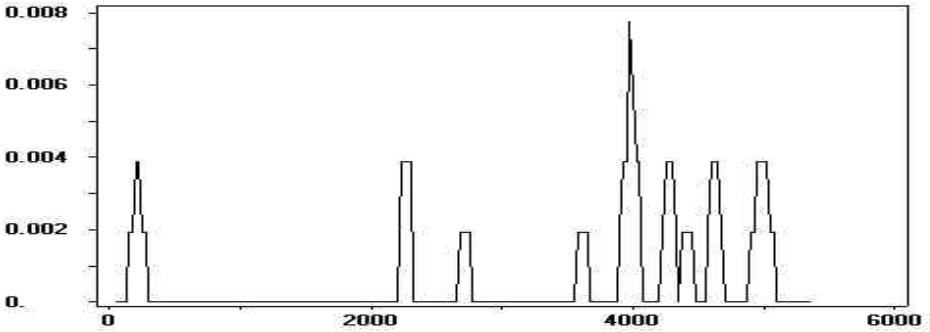

0.008
0.006
0.004
0.002
0.
0
2000
4000
6000
$\hat{q}_w$
0.05
0.04
0.03
0.02
0.01
0.
0
2000
4000
6000
Dxy
Nucleotide position

1
2
3
4
5
6
7
8
9
10
11

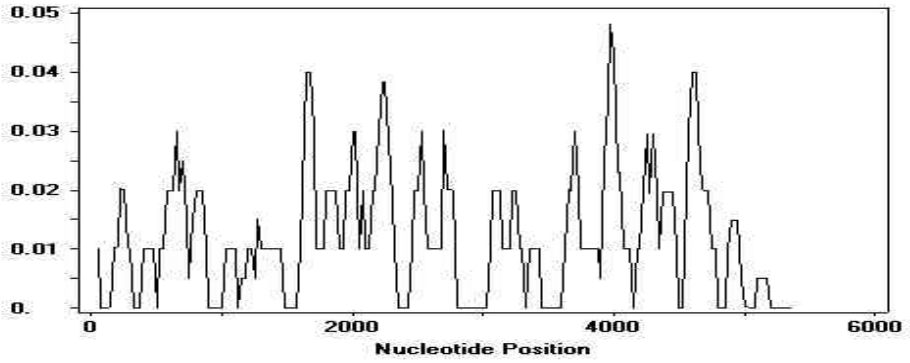

A
A
T
T
Without recombination
With recombination
A
A
A
A
A
A
T
T


7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Expected
Observed

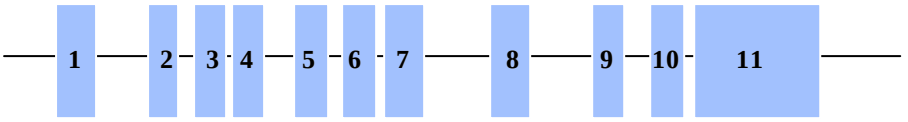

16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43
Expected
Observed

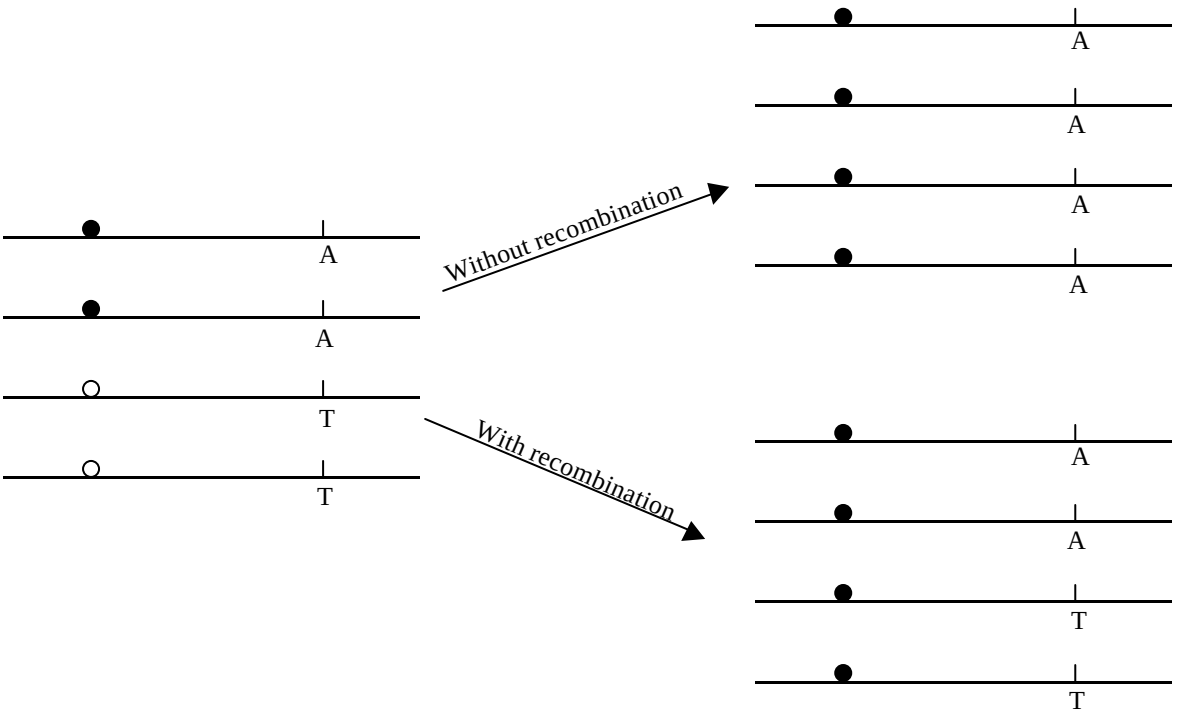

0.008
0.006
0.004
0.002
0.
0
2000
4000
6000

$\hat{q}_w$


0.05
0.04
0.03
0.02
0.01
0.
0
2000
4000
6000
Nucleotide Position

Dxy


1
2
3
4
5
6
7
8
9
10
11


A
A
T
T
Without recombination
With recombination
A
A
A
A
A
A
T
T

APÈNDIX II

Abans de la realització de la tesi doctoral vaig realitzar un altre treball també sobre la diversitat del genoma humà, però en aquest cas l'estudi era sobre DNA mitocondrial. Es tractava de veure com dues històries de població ben diferents (les de les illes de Bioko i de São Tomé, al golf de Guinea), havien influenciat l' estructura genètica de cada població a nivell del segment hipervariable I del DNA mitocondrial.

A continuació presentem la publicació resultant d'aquest treball.

A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea

Eva Mateu, David Comas, Francesc Calafell,
Anna Pérez-Lezaun, Augusto Abade i Jaume Bertranpetit

Annals of Human Genetics (1997) 61: 507-518

A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea.

E. MATEU, D. COMAS, F. CALAFELL[1], A. PÉREZ-LEZAUN, A. ABADE[2], J. BERTRANPETIT

Laboratori d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain
[1] Current address: Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
[2] Departamento de Antropologia, Universidade de Coimbra. Coimbra, Portugal

Author for correspondence:
Jaume Bertranpetit. Laboratori d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08028 Barcelona,Catalonia, Spain.
Tel.: +34-3-402 14 61
Fax: +34-3-411 08 87
e-mail: jaumeb@porthos.bio.ub.es

Running head: mtDNA variation in Bioko and São Tomé

SUMMARY

The hypervariable segment I of the control region of the mtDNA was sequenced in 45 unrelated individuals from Bioko and 50 from S o Tom , two islands in the Gulf of Guinea that have had very different settlement patterns: Bioko was colonised around 10,000 BP, while São Tomé was first settled by the Portuguese, who brought African slaves to the island. Two different patterns of sequence variation are evident and are also clearly consequence of their very different demographic histories. The Bubi present a low genetic diversity and it is likely that the island was colonised by a small number of individuals with small later migration. São Tomeans might be considered a subset of a mainland African population relocated to the island. They present high genetic diversity with a high number of sequences being shared with many continental populations. This study, with knowledge of the population history in island populations, strengths the genetic approach to unravel past demographic events.

INTRODUCTION

Mitochondrial DNA sequence analysis has proved to be a powerful tool in the study of human population history. It has been applied to very different space and time frames, from the origins of anatomically modern humans (Cann et al., 1987; Vigilant et al., 1991) to the colonisation of continents (Ward et al., 1991; Kolman et al., 1996; Richards et al., 1996; Comas et al., 1997), and to specific populations (Santos et al., 1994: Mountain et al., 1995; Bertranpetit et al., 1995; Calafell et al., 1996; Comas et al., 1996). Several properties of mitochondrial DNA (mtDNA) make it particularly suitable for evolution studies: absence of recombination, maternal inheritance, and a high mutation rate (Stoneking, 1993).

The impact of demographic processes in mtDNA sequence variation has been thoroughly modelled. In particular, the distribution of nucleotide pairwise differences (also known as mismatch distribution; Rogers & Harpending, 1992; Harpending et al., 1993) is particularly sensitive to the past demographic history of a population. In absence of selection and with mutation rates uniformly distributed across nucleotides, it has been shown (Rogers & Harpending, 1992; Harpending et al., 1993) that stationary populations would have irregular mismatch distributions, while a population expansion would generate a bell-shaped distribution, with a mode travelling to the right with time. A few African populations (Pygmies and !Kung in particular; Harpending et al., 1993) present irregular mismatch distributions, while other African populations (Graven et al., 1995; Watson et al., 1996) and virtually all other populations present bell-shaped mismatch histograms (Harpending et al., 1993). It has been shown that, in Europe, mismatch modes decrease from SE to NW, which has been interpreted as the footprint of the colonisation of Europe by anatomically modern humans (Calafell et al., 1996; Comas et al., 1996; Francalacci et al., 1996; Comas et al., 1997) recently confirmed by Neandertal mtDNA analysis (Krings et al., 1997). However, some of the implicit assumptions in the Rogers & Harpending (1992) model have been reassessed by Aris-Brosou & Excoffier (1996), who found that mutation rate variation across nucleotides (Wakeley, 1993; Hasegawa et al., 1993) could also produce bell-shaped mismatch distributions. Therefore, alternative hypotheses for the interpretation of mismatch distributions should be considered carefully before inferring a particular population history.

As mentioned above, some African populations present ragged pairwise difference distributions, while others have bell-shaped distributions. Watson et al. (1996) noted that the first are hunter-gatherers while the latter are farmers or pastoralists. Thus, they interpreted that the demographic expansion that produced bell-shaped mismatch distributions in Africa was the spread of the Neolithic. The latest direct counts of mutation events in the mtDNA (Howell et al., 1996; Parsons et al., 1997) resulted in estimates of the mutation rate that were

higher than previously thought; this adds support to a role for a recent demographic event, such as the Neolithic expansion, in generating bell-shaped mismatch distributions in Africa (Watson et al., 1996) and may be elsewhere (Sajantila et al., 1995; P bo, 1996; von Haeseler et al., 1996).

Sub-Saharan African populations share a pattern of mitochondrial DNA variation, with high levels of genetic diversity both within and between populations (Vigilant et al., 1991). This has also been found in nuclear genes (Tishkoff et al., 1996; Armour et al., 1996), and is compatible with the "Out of Africa" hypothesis, according to which non-African humans have a common, recent origin in Africa. This genetic pattern could increase the power of genetic population history analysis in African populations: a population bottleneck in an African population might be detected by comparison to the high levels of genetic diversity in other populations, or possible migrations can be traced back due to interpopulation variability.

We have analysed two samples from two African islands: the Bubi from Bioko and individuals from S o Tom . Bioko is a 2,000 Km^2 island in the Gulf of Guinea, 30 Km off the Cameroon coast, which, together with four smaller islands and the mainland territory of Rio Muni constitute the Republic of Equatorial Guinea. The island was first colonised 10,000 years BP (Vara & Bolekia, 1993), at the end of the last glacial period. Around 2,000 years BP, farming, and possibly a Bantu language, were introduced to the island (Mart n del Molino, 1993). The Bantu-speaking Bubi are the only population native to Bioko, and thought to be the descendants of the original colonisers of the island; the contact with Europeans decimated them to a few thousand at the turn of the 20th century. Nowadays, they number 35,000, and share the island with mainland Fang and Fernandinos, the later being descendants of former slaves liberated by the English in the 19th century. S o Tom e Pr ncipe, a former Portuguese colony, is located on the Equator in the Gulf of Guinea. It consists of two main islands (S o Tom and Pr ncipe) and a number of islets. Their total area is 964 Km^2, of which S o Tom comprises 865 Km^2. S o Tom island was probably uninhabited when first visited by European navigators in the 1470s. Thereafter, the Portuguese began to settle convicts and exiled Jews from Portugal on the island and established sugar plantations, using slave labour from the African mainland; for some years S o Tom was important in the trade and transshipment of slaves. A recent (1995) population size estimate for S o Tom is around 100,000 inhabitants and is mostly of African descent.

Bioko and S o Tom are, thus, two African islands in close proximity, with similar areas and population sizes, but with very different settlement patterns. The Bubi of Bioko are

the descendants of one or a few ancient waves of migration from the continent, whereas the S o Tomeans represent an admixed population with a recent origin. It is likely that the original settlers of Bioko were a small number of individuals whose descendants in around 500 generations increased to the actual number of Bubis. However, S o Tom may have been peopled by a larger number of imported slaves. These different patterns of settlement may have had genetic consequences. If the number of settlers of Bioko was small enough, we may be able to observe a reduction of genetic diversity in the Bubi when compared to the mainland populations and to S o Tom . Under isolation, gene diversity should be reduced, whereas the complexity of the genetic variation, as measured by coalescence patterns, mean pairwise differences, and nucleotide diversity, should remain comparable to that of mainland populations. In order to discover how two very different population histories have influenced the genetic structure of the population, we have sequenced a fragment of 360 base pairs in the hypervariable segment I of the mtDNA in 45 Bubi and 50 individuals from S o Tom , and we have compared the sequences to a set of fifteen African and two European populations.

MATERIAL AND METHODS

Population sampling

A 360-nucleotide sequence in hypervariable segment I (HVSI) of the mtDNA control region was analysed in 45 individuals from Bioko island, Equatorial Guinea. The sample comprised self-described unrelated Bubi individuals from the villages of Moka-Bioko, Moka-Malabo, in southern Bioko, and Rebola, in northern Bioko.The sample from S o Tom island comprises 50 individuals from different places covering the whole island.

Sample collection and DNA extraction

DNA was extracted from hair roots for the Bubi individuals. Hairs with their roots were plucked and stored in a vial with 95% ethanol. One root of each sample was introduced in a 1.5 ml sterile microfuge tube containing 0.5 ml of extraction buffer (10 mM Tris pH 8.0, 10 mM EDTA pH 8.0, 100 mM NaCl, 2% SDS, 39 mM DTT, and 20 mg/ml proteinase K), then incubated at 37 C and shaken at 180-200 rpm for at least 3 hours. After a phenol-chloroform extraction (Sambrook,1989), DNA was concentrated in Centricon-30 tubes and stored at -20 C. For the S o Tom sample DNA (supplied by A. Abade, Coimbra) was extracted from fresh blood using standard protocols.

mtDNA amplification

Amplification was performed using approximately 150 to 250 ng of the DNA sample in a 25 l reaction volume; the temperature profile for 30 cycles of amplification was 94 C for 1 min, 58 C for 1 min and 72 C for 1 min. The primers used in this reaction, L15996 (5'-CTCCACCATTAGCACCCAAAGC-3'), and H16401 (5'-TGATTTCACGGAGGATGGTG-3'; Vigilant et al., 1989) amplified a 446-base pair (bp) segment containing the 360-bp region that was subsequently sequenced.

mtDNA sequencing

Out of the 45 Bubi samples, 13 were sequenced with an automatic DNA sequencer, while the remaining 32 were sequenced manually; the choice of method depended only on sequencer time availability. All the S o Tom samples were sequenced with an automatic DNA sequencer. Automated sequencing was performed according to manufacturer's specifications. The sequencing reaction was performed separately on each strand with the DNA Sequencing Kit^{TM} (Perkin Elmer), Dye Terminator Cycle Sequencing with AmpliTaqR DNA Polymerase. The product of the sequence reaction was run in an ABI PRISM 377 (Perkin Elmer).

When sequencing manually, the amplification product was purified with GeneClean (BIO 101). Seven l of the purified amplified product were sequenced with Sequenase

Version 2.0 (USB) following supplier's recommendations, except for the annealing step, which was performed by boiling the annealing reaction mixture for 3 min in presence of nonidet P-40, followed by a short time in a dry-ice ethanol bath. Both strands were sequenced using the amplification primers. Reaction products were separated by electrophoresis, dried, fixed, and subjected to autoradiography.

Sequences were aligned with the ESEE program (Cabot, 1988), and the segment from positions 16024 to 16383 (Anderson et al., 1981) was used for analysis.

Statistical analysis

A set of fifteen African population samples, comprising a total of 645 individuals and including the Fang in mainland Equatorial Guinea (Figure 1; Table 1) and two European (a British sample, N=100, Piercy et al., 1993 and a Portuguese sample, N= 54, C rte Real et al., 1996) were used as reference.

Nucleotide diversity (Nei & Tajima, 1981) was estimated as $(n/n-1) \sum_{i=1}^{l} (1-x_i^2)$, where n is sample size, l is sequence length, and x_i is the frequency of a nucleotide (A, C, G or T) at position i. Similarly, sequence diversity was estimated as $(n/n-1) \sum_{i=1}^{k} (1-p_i^2)$, where p_i is the frequency of each of the k different sequences in the sample. The significance of the difference in sequence diversity between two populations was tested through a permutation procedure: the individuals in both populations were randomly assigned to one of two samples of the same size as the original ones; sequence diversities were computed for the new random samples, and the difference was recorded. This procedure was repeated 10,000 times, and the probability of the difference not being significantly smaller than zero was estimated as the fraction of permuted differences that were less extreme than the observed value. Tajima's (1989) D statistic, which is the standardised difference between two different estimates of $=2N_e$, was computed. Under a number of assumptions, Tajima's D measures the deviation from mutation-drift equilibrium. Bertorelle & Slatkin (1995), and Aris-Brosou & Excoffier (1996) have examined the effects of population expansion and mutation rate variability on D. We also estimated from the number of segregating sites (Watterson, 1975) through equation $=P_n/(1+1/2+...+(n-1)^{-1})$ in Nei (1987), where P_n is the number of segregating sites divided by 360 and n is the number of the observed sequences.

The phylogeny of mtDNA sequences in the Bubi and in the S o Tomeans was approached through a neighbour-joining tree (Saitou and Nei, 1987) based on KimuraÆs two-parameter model with a transition to transversion ratio set to 15:1. We used PHYLIP (Felsenstein, 1989) to produce these sequence phylogenies. The distribution of branch lengths in a phylogeny can be used to quantify its degree of *starness*, that is, whether most of the sequences attach with long branches to a central point, as opposed to a pattern in which branching events are more regular. The two patterns may reflect different population histories: a star-like tree may correspond to an expanding population, whereas a regular tree

could reflect a stationary population (von Haeseler et al., 1996). In order to characterise the branch length distribution, we computed its third- and fourth- degree moment (i.e., its skewness and kurtosis).

Genetic distances among populations were estimated as $d=d_{ij}-(d_{ii}+d_{jj})/2$ (Nei, 1987), where d_{ij} is the mean nucleotide pairwise difference between populations *i* and *j*, and d_{ii} and d_{jj} are the mean internal pairwise differences of populations *i* and *j*, respectively. The standard error of the distances was estimated by resampling nucleotide positions in 1,000 bootstrap iterations. A neighbour-joining tree (Saitou & Nei, 1987) was built from the genetic distance matrix, and its robustness was assessed from 1,000 boostrap iterations (Felsenstein, 1985).

RESULTS

Nucleotide diversity

The complete sequence of a 360-bp segment of the mtDNA control region (HVS1, positions 16024 to 16383 according to the numeration by Anderson et al., 1981) was determined in a sample of 45 Bubi from Bioko and 50 individuals from S o Tom (Table 2). For the Bubi sample, 18 different sequences were found, with 32 variable nucleotide positions. The S o Tomeans presented 32 different sequences, with 53 variable nucleotides. Overall, 48 different sequences and 61 segregating sites were found. Four of the segregating sites presented transversional changes, 36 were transitions between pyrimidines (C to T or vice versa), 19 were purine transitions, and in two positions (16114 and 16265), we observed both transitions and transversions. Whereas two thirds of the transitions observed involve pyrimidines, these represent 55.8% of nucleotides in this segment. This might represent part of the uneven distribution of mutation rates across nucleotides (Bertorelle & Slatkin, 1995; Aris-Brosou & Excoffier, 1996).

Nucleotide diversity (Nei & Tajima, 1981) was 0.0210 0.0006 in the Bubi and 0.0231 0.0006 in the S o Tomeans. Both values are similar to those found in other West African populations (Table 3). Tajima's statistic was D=0.114 in the Bubi and D=-1.034 in the S o Tomeans, both not significantly different from zero according to Table 3 in Tajima (1989). All African populations sequenced so far present non-significant TajimaÆs statistics, whereas European populations tend to present TajimaÆs statistics significatively smaller than zero.

Sequence diversity

Three sequences (MAL 6, MAL 10, and MAL 19; Table 2) accounted for 42% of the individuals in the Bubi sample, while 9 sequences were found once. In contrast, we found 32 different sequences in 50 S o Tomeans, and the three most frequent sequences represented 26% of the individuals in the sample. Sequence diversity was lower in the Bubi (0.928) than in the S o Tomeans (0.973; Table 3). The difference in sequence diversity was significant (p=0.0001) according to the permutation test. The Bubi have also significantly smaller sequence diversities than the Mandenka (p=0.0001) and the Yoruba (p=0.0001), while sequence diversity in the S o Tomeans was similar to that in the Mandenka (p=0.5601) but lower than in the Yoruba (p=0.0060). We also estimated $=2N_e$ as suggested by Watterson (1975). Differences in among populations should be a function of effective population size, since mutation rate is presumably constant across populations. After the !Kung, the Bubi present the second lowest value among sub-Saharan African populations (Table 3), whereas the São Tomeans have a in the mid to upper African range.

The Bubi shared five sequences with other African populations, including two with the S o Tomeans. Two of them (MAL 12 and SAO 122; Table 2) were found in several West African populations and in the Kenyan Kikuyu. The other three sequences were found in one sample each (i.e., the S o Tomeans, the Pygmies, and the Fulbe). Eleven of the sequences found in the S o Tomeans had previously been described in other African populations. The S o Tomeans shared sequences with a variety of West and East African populations; they shared five sequences with the Fulbe, four with the Mandenka, three with the Kikuyu and the Hausa, two with the Bubi, Yoruba, Turkana, and Tuareg and one with Kanuri and Mozabite. None of the sequences we found in both samples had previously been described in individuals of European ancestry.

A group of four sequences (SAO 118, SAO 198, SAO 160, and MAL 46; Table 2) bears a characteristic group of mutations that were found to be associated with a 9-bp deletion in the CoII/tRNA^{Lys} intergenic region in African populations (Soodyall et al., 1996). If the association described by these authors holds in the Bubi and S o Tomeans, then the frequency of the 9-bp deletion would be 8.9% in the Bubi and 8% in the S o Tomeans. These frequencies are intermediate between those found in West Africans (0-2%) and in Central Africans (0-30%).

Pairwise difference distribution

The mean number of pairwise nucleotide differences was 7.56 in the Bubi and 8.30 in the S o Tomeans; both values fall in the range observed for other West African populations (Table 3). The nucleotide pairwise difference distributions, also called mismatch distribution (Rogers & Harpending, 1992; Harpending et al., 1993) are shown in Figure 2. The Bubi distribution presents several modes, at zero, seven, ten, and thirteen differences, and its raggedness coefficient (Harpending et al., 1993) is 0.0223. The S o Tomean distribution is smoother (raggedness coefficient, 0.0045), bell-shaped, and presents a single mode at nine differences. As shown by the standard errors (Figure 2), the different shapes of the two distributions cannot be attributed to a random effect. The multiple peaks and troughs in the ragged Bubi distribution are significatly different from the values in the smoother S o Tom distribution.

Phylogenetic tree of sequences

We estimated the phylogeny of Bubi and S o Tomean sequences by means of genetic distances and neighbour-joining trees (Figure 3). While the Bubi phylogeny (Figure 3a) presented a few, deep-rooting branches, the S o Tom tree reflects a star-like phylogeny (Figure 3b). It has been shown that stable demographic histories result in deep-rooting phylogenies, while population expansion tend to generate star-like phylogenies (Slatkin & Hudson, 1991; Haeseler et al., 1996). The difference between the two populations may be

measured by parameters of the distribution of branch length in the two trees. Skewness and kurtosis are significative higher in Bubis ($_3$=2.142 0.414; $_4$=5.360 0.809) than in S o Tomeans ($_3$=1.250 0.304; $_4$=0.597 0.599), being these parameters an indirect estimator of the degree of *starness* of a phylogeny.

Genetic distances among African populations

Genetic distances were computed between several African populations, plus two European reference populations. The Bubi presented relatively long distances to other African populations. S o Tom presented the shortest genetic distances to the Yoruba (0.11), and to other West African and Sahelian populations (0.2-0.3). A neighbour-joining tree (Figure 4) based on those distances presented two particularly long, robust branches, which divide the populations in three main groups: i) The Pygmy and !Kung populations; the latter join the tree with a long branch; ii) the Western African and Sahelian populations which are connected to each other by short branches, and iii) the North African and European populations. The Bubi, as well as the mainland Equatorial Guinea Fang, appear at the base of the Pygmy-!Kung branch, whereas the S o Tomeans are found in the tight West African-Sahelian cluster.

DISCUSSION

We have sequenced the mtDNA HVSI region in individuals from two islands in the Gulf of Guinea that have very different settlement patterns: Bioko was first colonised around 10,000 BP (Vara & Bolekia, 1993) and might have received few immigrants since then, while São Tomé was first settled by the Portuguese, who brought African slaves to the island.

We have found some similarities in the mtDNA sequences of both populations. The sequence pool of both samples is clearly African, as shown by the tree of populations (Figure 4) and by the fact that the Bubi and the São Tomeans share identical sequences only with Africans.Thus not a single sequence of the present population has a female European ancestry. As most African populations, and unlike populations from other continents, both Bubi and São Tomeans present a high mean number of pairwise differences. However, two different patterns of sequence variation between the two populations are evident and are also clearly consequences of their very different demographic histories.

The Bubi present a low sequence diversity and estimate; both may be the result of a long-term female effective population size that is lower than that of other African populations. From what is known from historical and archaeological sources (Vara & Bolekia, 1993), it is likely that the island was colonised by a small number of individuals and that later contributions from the continent were not significant in population size terms. Thus, the population growth might have been slow, even during the Neolithic transition, and until recent times. This is reflected in the sequence phylogeny, with deep-rooting branches, and in the ragged distribution of pairwise differences. The ancient origin of the Bubi is also underscored by their location in the population tree, out of the tight West African cluster and at the base of the branch leading to such ancient African populations as the Pygmies and the !Kung.

Given the origin of the São Tomeans as the descendants of slaves, they might be considered a subset of a mainland African population relocated to the island. This origin is clearly reflected in their mtDNA sequence variation pattern, which is typical of an African population. They present high nucleotide and sequence diversities, and high estimates, which are indications of a relatively high female effective population size. Graven et al. (1995) suggested that the Mandenka from Senegal had an effective population size that was noticeably larger than the observed female census size, due to past or present migration from other populations. The same can be applied to the São Tomeans and their immediate ancestors in the continent. Those were most likely farming Bantu-speakers. The pairwise difference distribution we observed in the São Tomeans conforms to the pattern described by Watson et al. (1996) for the mainland farmers: it is smooth and bell-shaped, which corresponds to a starlike sequence phylogeny. The very recent mainland African origin of the São Tomeans resulted in a higher number of sequences being shared with the mainland populations, and, in the population tree, they clearly belong to the tight West African cluster.

We have seen how two very different population histories generated different mtDNA sequence variation patterns in two populations. The accuracy with which sequence variation patterns could be predicted from known population history provides reliability to the opposite endeavour: inferring unknown or partially known population histories from observed genetic variation. This approach, thus, has a robust genetic foundation with widening possibilities thanks to the growing theoretical framework that helps to understand and explain the extent of human genome variation.

ACKNOWLEDGEMENTS

This research was supported by DGICYT (Spain) grant PB92-0722 and PB95-0267-C02-01 and by Generalitat de Catalunya (Catalonia), Grup de Recerca Consolidat 1995SGR00205 and 1996SGR00041 to JB. EM was awarded a PhD fellowship from the University of Barcelona; FC was supported by a post-doctoral fellowship from CIRIT (Catalonia), who also granted a PhD fellowship to DC. AP-L was awarded a PhD fellowship by DGICYT (Spain). Samples from Bioko were kindly supplied by Carmen Mat and Montserrat Colell during her field primate behaviour studies. Samples from S o Tom were collected with funding from STRDA/P/CEN/532/92 portuguese project. We also thank the Servei de Seq enciaci , Serveis Cient fico T cnics, University of Barcelona, for their invaluable technical support.

REFERENCES

ANDERSON, S., BANKIER, A. T., BARRELL, B.G., DE BRUIJN, M.H., COULSON, A.R., SANGER, F., SCHREIER, P.H., SMITH, A.J.H., STADEN, R. & YOUNG, G. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465.

ARIS-BROSOU, S. & EXCOFFIER, L. (1996).The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol*. **13**, 494-504.

ARMOUR, J.A.L., ANTTINEN, T., MAY, C.A., VEGA, E.E., SAJANTILA, A., KIDD, J.R., KIDD, K.K., BERTRANPETIT, J., P BO, S. & JEFFREYS, A.J. (1996). Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics*, **13,** 154-160.

BERTORELLE, G., & SLATKIN, M. (1995). The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. *Mol. Biol. Evol.* **12**, 887-892.

BERTRANPETIT, J., SALA, J., CALAFELL, F., UNDERHILL, P., MORAL, P. & COMAS, D. (1995). Human mitochondrial DNA variation and the origin of the Basques. *Ann. Hum. Genet.* **59**, 63-81.

CABOT, E.L. (1988). *ESEE, the eyeball sequence editor, version 1.06.* Burnaby, B.C., Canada, V5C 2YZ.

CALAFELL, F., UNDERHILL, P., TOLUN, A., ANGELICHEVA, D. & KALAYDJIEVA, L. (1996). From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann. Hum. Genet*. **60**, 35-49.

CANN, R.L., STONEKING, M. & WILSON, A.C. (1987). Mitochondrial DNA and human evolution. *Nature* **325**, 31-36.

COMAS, D., CALAFELL, F., MATEU, E., P REZ-LEZAUN, A. & BERTRANPETIT, J. (1996). Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol. Biol. Evol.* **13**, 1067-1077.

COMAS, D., CALAFELL, F., MATEU, E., P REZ-LEZAUN, A., BOSCH, E. & BERTRANPETIT, J. (1997). Mitochondrial DNA variation and the origin of the Europeans. *Hum. Genet.* **99**, 443-449.

C RTE-REAL, H.B.S.M., MACAULAY, V.A., RICHARDS, M., HARITI, G., ISSAD, M.S., CAMBON-THOMSEN, A., PAPIHA, S., BERTRANPETIT, J. & SYKES, B.C. (1996). Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet*. **60**, 331-350.

FELSENSTEIN, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **35**, 785-791.

FELSENSTEIN, J. (1989). PHYLIP -- Phylogeny Inference Package. *Cladistics* **5**, 164-166.

FRANCALACCI, P., BERTRANPETIT, J., CALAFELL, F. & UNDERHILL, P. (1996). Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am. J. Phys. Anthropol*. **100**, 443-460.

GRAVEN, L., PASSARINO, G., SEMINO, O., BOURSOT, P., SANTACHIARA-BENERECETTI, S., LANGANEY, A. & EXCOFFIER, L. (1995). Evolutionary correlation between control region sequence and restriction polymorphism in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol. Biol. Evol*. **12,** 334-345.

HARPENDING, H.C., SHERRY, S.T., ROGERS, A.R. & STONEKING, M. (1993). The genetic structure of ancient human populations. *Curr. Anthrop.* **34,** 483-496.

VON HAESELER, A., SAJANTILA, A. & P BO, S. (1996). The genetical archaeology of the human genome. *Nature Genetics* **14**, 135-140.

HASEGAWA, M., DI RIENZO, A., KOCHER, T.D. & WILSON, A.C. (1993). Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.* **37**, 347-354.

HOWELL, N., KUBACKA, I. & MACKEY, D.A. (1996). How rapidly does the human mitochondrial genome evolve? *Am. J. Hum. Genet*. **59,** 501-509.

KOLMAN, C.J., SAMBUUGHIN, N. & BERMINGHAM, E. (1996). Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics* **142,** 1321-1334.

KRINGS, M., STONE, A., SCHMITZ, R.W,. KRAINITZKI, H., STONEKING, M. & P BO, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell* **90,** 1-20.

MART N DEL MOLINO, A. (1993). *Los Bubis. Ritos y creencias*. Madrid, Labrys 54.

MOUNTAIN, J.L., HEBERT, J.M., BHATTACHARYYA, S.S., UNDERHILL, P., OTTOLENGHI, C., GADGIL, M. & CAVALLI-SFORZA, L.L. (1995). Demographic history of India and mitochondrial DNA sequence diversity. *Am. J. Hum. Genet.* **56,** 979-992.

NEI, M. & TAJIMA, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics* **97,** 145-163.

NEI, M. (1987). *Molecular evolutionary genetics.* New York: Columbia University Press.

P BO, S.. (1996).Mutational hot spots in the mitochondrial microcosm. *Am. J. Hum. Genet*. **59**, 493-496.

PARSONS, T.J., MUNIEC, D.S., SULLIVAN, K., WOODDYATT, N., ALLISTON-GREINER, R., WILSON, M.R., BERRY, D.L., HOLLAND, K.A., WEEDN, V.W.,GILL, P. & HOLLAND, M.M.. (1997). A high observed substitution rate in the human mitochondrial DNA control region. *Nature Genetics* **15**, 363-368.

PIERCY, R., SULLIVAN, K.M., BENSON, N. & GILL, P. (1993). The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int. J. Leg. Med.* **106,** 85-90.

PINTO, F., GONZßLEZ, A.M., HERNßNDEZ, M., LARRUGA, J.M. & CABRERA, V. (1996). Genetic relationship between the Canary islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann. Hum. Genet.* **60,** 321-330.

RICHARDS, M., C RTE-REAL, H., FORSTER, P., MACAULAY, V., WILKINSON-HERBOTS, H., DEMAINE, A., PAPIHA, S., HEDGES, R., BANDELT, H-J. & SYKES, B. (1996). Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59,** 185-203.

ROGERS, A.R. & HARPENDING, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9,** 552-569.

SAITOU, N. & NEI, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406-425.

SAJANTILA, A., LAHERMO, P., ANTINNEN, T., LUKKA, M., SISTONEN, P., SAVONTAUS, M-L., AULA, P., BECKMAN, L., TRANEBJAERG, L., GEDDE-DAHL, T.,IISSEL-TARVER, L., DI RIENZO, A. & P BO, S. (1995). Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Research* **5**, 42-52.

SAMBROOK, J., FRITSCH, E.F. & MANIATIS, T. (1989). *Molecular cloning. A laboratory manual.* Second Edition. New York: Cold Spring Harbor Laboratory Press.

SANTOS, M., WARD, R.H. & BARRANTES, R. (1994). mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. *Hum. Biol.* **66,** 963-977.

SLATKIN, M. & HUDSON, D. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555-562.
SOODYALL, H., VIGILANT, L., HILL, A.V., STONEKING, M. & JENKINS, T. (1996). mtDNA Control-Region sequence variation suggests multiple independent origins of an ôAsian-Specificö 9-bp deletion in Sub-Saharan Africans. *Am. J. Hum. Genet*. **58**, 595-608.

STONEKING, M. (1993). DNA and recent human evolution. *Evol. Anthropol.* **2,** 60-73.

TAJIMA, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585-595.

TISHKOFF, S.A., DIETZSCH, E., SPEED, W., PAKSTIS, A.J., KIDD, J.R., CHEUNG, K., BONN - TAMIR, B., SANTACHIARA-BENERECETTI, A.S., MORAL, P., KRINGS, M., P BO, S., WATSON, E., RISCH, N., JENKINS, T.& KIDD, K.K. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271,** 1380-1387.

VARA. S. & BOLEKIA, J. (1993). Bioko, tierra de los Bubis. *Africa 2000* **20,** 16-21.

VIGILANT, L., PENNINGTON, R., HARPENDING, H., KOCHER, T.D. & WILSON, A.C. (1989). Mitochondrial DNA sequences in single hairs from a southern African populations. *Proc. Natl. Acad. Sci. USA* **86,** 9350-9354.

VIGILANT, L., STONEKING, M., HARPENDING, H., HAWKES, K. & WILSON, A.C. (1991). African populations and the evolution of mitochondrial DNA. *Science* **253,** 1503-1507.

WARD, R.H., FRAZIER, B.L., DEW-JAGER, K. & P BO, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA* **88,** 8720-8724.

WAKELEY, J. (1993). Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. *J. Mol. Evol.* **37,** 613-623.

WATTERSON, G.A. (1975). On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**, 256-276.

WATSON, E., BAUER, K., AMAN, R., WEISS, G., VON HAESELER, A. & PÄÄBO, S. (1996). mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.* **59,** 437-444.

Figure 1. African populations included in our study.

Figure 2. Nucleotide pairwise difference (mismatch) distribution in the Bubi (black line) and São Tomeans (grey line). Standard errors were estimated by resampling with replacement 1,000 times over pairs of individuals.

Figure 3. Phylogenetic trees of Bubi (Figure 3a) and São Tomeans (Figure 3b) sequences.

Figure 4. Hypervariable segment I neighbour-joining tree among seventeen African and two European population. Figures in the branches represent the percent fraction of times they were found in 1,000 bootstrap iterations.