Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28

Patricia Taillon-Miller[1], Irma Bauer-Sardiña[1], Nancy L. Saccone[2], Jenna Putzel[1], Tarja Laitinen[4], Antonio Cao[3], Juha Kere[4], Giuseppe Pilia[3], John P. Rice[2] & Pui-Yan Kwok[1]

Linkage disequilibrium (LD), or the non-random association of alleles, is poorly understood in the human genome[1]. Population genetic theory suggests that LD is determined by the age of the markers, population history, recombination rate, selection and genetic drift[2]. Despite the uncertainties in determining the relative contributions of these factors, some groups have argued that LD is a simple function of distance between markers[3,4]. Disease-gene mapping studies and a simulation study gave differing predictions on the degree of LD in isolated and general populations[5,6]. In view of the discrepancies between theory and experimental observations, we constructed a high-density SNP map of the Xq25–Xq28 region[7] and analysed the male genotypes and haplotypes across this region for LD in three populations. The populations included an outbred European sample (CEPH males) and isolated population samples from Finland and Sardinia. We found two extended regions of strong LD bracketed by regions with no evidence for LD in all three samples. Haplotype analysis showed a paucity of haplotypes in regions of strong LD. Our results suggest that, in this region of the X chromosome, LD is not a monotonic function of the distance between markers, but is more a property of the particular location in the human genome.

Table 1 • SNP markers

	SNP	GenBank accession no.[a]	GC content of GenBank sequence (%)	X map distance (Mb)[b]	Distance to next marker (kb)	CEPH heterozygosity	Finnish heterozygosity	Sardinian heterozygosity
1	Xq1226-2			32.55	73,350	0.41	NT	NT
2	Xq1136-1	L13942		105.9	400	0.5	NT	NT
3	Xq544-1	Z73967		106.3	4,500	0.5	0.4	0.46
4	Xq500-1	DXS287		110.8	6,824	0.46	0.48	0.5
5	Xq3562-1	L78810	46	117.624	61	0.48	0.48	NT
6	Xq3655-2	L78810	46	117.685	69	0.42	0.4	NT
7	Xq3656-2	L78810	46	117.754	4,551	0.39	0.27	NT
8	Xq3847-1	Z82207	46	122.305	247	0.29	0.42	NT
9	Xq1063-1	L41185		122.552	2,115	0.43	NT	NT
10	Xq3570-3	Z72519	36	124.667	467	0.5	0.5	NT
11	Xq3855-1	Z73362	36	125.134	1,319	0.46	0.37	NT
12	Xq3862-1	Z84481	35	126.453	110	0.38	0.41	0.33
13	Xq4009-2	Z76735	36	126.563	76	0.49	0.5	NT
14	Xq4007-1	Z76735	36	126.639	44	0.34	0.43	0.14
15	Xq3774-2	Z82209	36	126.683	93	0.44	0.47	0.23
16	Xq3773-1	Z82209	36	126.776	187	0.37	0.45	0.14
17	Xq3846-1	Z75741	35	126.963	193	0.42	0.44	NT
18	Xq3804-1	Z74696	36	127.156	90	0.37	0.35	0.2
19	Xq3812-1	Z74696	36	127.246	281	0.18	0.27	0.16
20	Xq3917-1	Z71187	35	127.527	30	0.32	0.35	0.17
21	Xq3849-1	AC002422	37	127.557	64	0.35	0.46	NT
22	Xq4095-1	AC002422	37	127.621	95	0.37	0.49	NT
23	Xq3699-1	Z82900	39	127.716	113	0.39	0.48	NT
24	Xq3811-1	Z72001	39	127.829	99	0.5	0.5	NT
25	Xq3698-1	Z73418	37	127.928	2,980	0.48	0.46	0.49
26	Xq3879-1	Z82205	40	130.908	140	0.5	0.48	NT
27	Xq3802-1	Z81365	41	131.048	1,395	0.49	0.49	NT
28	Xq3070-1	M26434	40	132.443	8,807	0.37	NT	NT
29	Xq2904-1	AF120094	30	141.25	15,891	0.37	NT	NT
30	Xq3840-1	U82671	51	157.141	50	0.5	0.37	NT
31	Xq3841-1	U82671	51	157.191	732	0.38	0.44	NT
32	Xq3476-1	U52111	57	157.923	117	0.47	0.44	0.47
33	Xq3449-1	X79198	60	158.04	48	0.38	0.38	0.29
34	Xq3471-1	U52112	56	158.088	38	0.38	0.38	0.23
35	Xq4001-1	AF031078	46	158.126	67	0.4	0.39	0.14
36	Xq3413-1	Z47046	46	158.193	44	0.42	0.41	0.2
37	Xq2816-1	Z46936	50	158.237	28	0.41	0.49	0.39
38	Xq3274-1	Z49258	49	158.265	225	0.41	0.47	0.44
39	Xq1452-1	L44140	57	158.49	–	0.49	0.47	NT

[a]The GenBank accession number of the sequence from which the STS is developed. [b]Each of the SNP markers was given a Mb distance estimate along the X chromosome.

[1]*Division of Dermatology and* [2]*Department of Psychiatry, Washington University, St. Louis, Missouri, USA.* [3]*IRTAM-CNR, University of Cagliari, Cagliari, Sardinia, Italy.* [4]*Finnish Genome Center, University of Helsinki, Helsinki, Finland. Correspondence should be addressed to P.-Y.K. (e-mail: kwok@genetics.wustl.edu).*

To determine the extent of LD in the Xq25–q28 region, males of the CEPH, Finnish and Sardinian samples were genotyped with a set of SNP markers. Additional SNP markers were typed in Xq25 (a region of low GC content and low gene density) and Xq28 (a region of high GC content and high gene density), where initial results suggested strong linkage disequilibrium.

We genotyped 39 SNPs (Table 1) and identified regions of LD by performing pair-wise comparisons with SAS. LD results for a selected set of pair-wise comparisons in the CEPH, Finnish and Sardinian samples are shown (Table 2), as well as the computed values for D´, r^2, χ^2 and *P* values corresponding to χ^2 for the pair-wise comparisons. Most SNP markers used here were relatively polymorphic, with similar allele frequencies, so that the χ^2 value was suitable for comparisons of LD within a population sample. The χ^2 values corresponded well with the D´ values (Table 2). Comparable allele frequencies (and heterozygosities) for the populations made r^2 appropriate for comparing LD across population samples. Furthermore, the three sample sizes were similar, so that χ^2 was used to compare LD across population samples, as $r^2=\chi^2/N$, where N is the number of observations in the computation of r^2. Although there were more individuals in the Sardinian sample, the average number of observations per two-way comparison was 107 (s.d.=14), giving numbers comparable to those for the CEPH and Finnish samples (average N of 86 (s.d.=4) and 96 (s.d.=4), respectively).

We plotted LD (r^2) versus distance in kilobases within the Xq25 (Xq3862-1 to Xq3698-1) and Xq28 (Xq3476-1 to Xq1452-1) regions (Fig. 1). The distribution of the data points showed that LD is not a simple monotonic function of the distance between markers. We found two extended regions of strong linkage disequilibrium in the CEPH, Finnish and Sardinian samples by pair-wise comparisons in selected regions of Xq25 and Xq28 (Fig. 2). The first region is in Xq25 and extends from marker Xq3862-1 to marker Xq3917-1 (spanning 1 Mb of DNA). The second region of strong LD is in Xq28 and extends from marker Xq3476-1 to marker Xq3274-1 (340 kb). Although there are small quantitative differences in the LD regions among the three samples, the extent and general location of the LD regions were similar across the samples. For example, strong LD was found in almost all pair-wise comparisons in both Xq25 and Xq28 LD regions with abrupt changes in LD at the boundaries. In the CEPH sample, the Xq25 LD region extended slightly more

Table 2 • Pair-wise comparisons of SNP markers from the Xq25 and Xq28 LD regions

			CEPH				Finland				Sardinia			
Locus 1	Locus 2	Distance (Mb) between loci 1 and 2	D´	r^2	χ^2	*P* values	D´	r^2	χ^2	*P* values	D´	r^2	χ^2	*P* values
Xq25														
Xq3804_1	Xq3917_1	0.000	1.00	0.93	76.25	0.0000	1.00	1.00	94.00	0.0000	1.00	0.91	106.24	0.0000
Xq3774_2	Xq3773_1	0.093	1.00	0.57	50.73	0.0000	1.00	0.81	80.58	0.0000	0.81	0.34	34.11	0.0000
Xq4007_1	Xq3773_1	0.137	1.00	0.94	79.53	0.0000	1.00	0.82	74.81	0.0000	0.88	0.77	94.49	0.0000
Xq4007_1	Xq3774_2	0.044	1.00	0.52	44.10	0.0000	1.00	0.68	62.12	0.0000	0.86	0.34	39.26	0.0000
Xq3774_2	Xq3804_1	0.473	0.71	0.34	28.04	0.0000	0.78	0.28	27.06	0.0000	0.15	0.02	1.55	0.2128
Xq3774_2	Xq3917_1	0.844	0.83	0.36	31.52	0.0000	0.77	0.27	26.12	0.0000	0.11	0.01	0.91	0.3411
Xq3862_1	Xq3804_1	0.703	0.23	0.06	4.29	0.0383	0.61	0.28	25.89	0.0000	0.19	0.02	1.64	0.2001
Xq3804_1	Xq3812_1	0.090	0.44	0.06	4.90	0.0268	0.60	0.24	23.55	0.0000	0.63	0.35	34.08	0.0000
Xq3862_1	Xq3773_1	0.323	0.30	0.09	7.18	0.0074	0.52	0.22	21.20	0.0000	0.68	0.08	7.18	0.0074
Xq3812_1	Xq3917_1	0.281	0.29	0.04	2.76	0.0964	0.57	0.20	19.26	0.0000	0.59	0.35	34.70	0.0000
Xq4007_1	Xq3804_1	0.517	0.17	0.02	1.92	0.1663	0.54	0.21	18.28	0.0000	0.33	0.08	10.10	0.0015
Xq4007_1	Xq3917_1	0.888	0.15	0.02	1.84	0.1755	0.53	0.20	17.48	0.0000	0.34	0.09	12.24	0.0005
Xq3773_1	Xq3804_1	0.380	0.13	0.02	1.40	0.2367	0.52	0.15	15.13	0.0001	0.36	0.08	8.50	0.0036
Xq3773_1	Xq3917_1	0.751	0.12	0.01	1.12	0.2906	0.52	0.15	14.46	0.0001	0.32	0.08	8.99	0.0027
Xq3774_2	Xq3812_1	0.563	0.45	0.05	3.92	0.0478	0.69	0.14	14.29	0.0002	0.44	0.15	12.84	0.0003
Xq3773_1	Xq3812_1	0.470	0.18	0.01	0.93	0.3354	0.53	0.10	10.25	0.0014	0.82	0.57	50.57	0.0000
Xq3774_2	Xq3698_1	1.245	–0.02	0.00	0.02	0.9011	–0.25	0.05	5.29	0.0215	0.61	0.04	3.84	0.0500
Xq4007_1	Xq3812_1	0.607	–0.32	0.00	0.21	0.6449	0.32	0.04	3.81	0.0509	0.69	0.36	36.62	0.0000
Xq28														
Xq3471_1	Xq4001_1	0.038	1.00	0.94	85.93	0.0000	1.00	1.00	99.00	0.0000	0.64	0.14	14.13	0.0002
Xq3449_1	Xq3471_1	0.048	–1.00	0.94	77.95	0.0000	–1.00	0.95	93.91	0.0000	–1.00	1.00	104.00	0.0000
Xq3449_1	Xq4001_1	0.086	–1.00	0.94	77.95	0.0000	–1.00	0.95	92.94	0.0000	–1.00	0.43	44.41	0.0000
Xq2816_1	Xq3274_1	0.028	1.00	0.94	75.20	0.0000	0.95	0.87	75.31	0.0000	0.91	0.79	88.24	0.0000
Xq3471_1	Xq3413_1	0.105	0.94	0.66	59.40	0.0000	0.84	0.60	60.33	0.0000	–0.33	0.08	9.40	0.0022
Xq4001_1	Xq3413_1	0.067	0.88	0.61	55.29	0.0000	0.84	0.60	59.60	0.0000	–0.03	0.00	0.08	0.7817
Xq3449_1	Xq3413_1	0.153	–0.93	0.61	50.37	0.0000	–0.83	0.56	55.65	0.0000	0.25	0.03	3.89	0.0486
Xq3476_1	Xq3413_1	0.270	0.30	0.07	6.20	0.0128	0.53	0.25	23.74	0.0000	0.26	0.01	0.69	0.4073
Xq3476_1	Xq3471_1	0.165	0.59	0.20	18.04	0.0000	0.58	0.24	22.74	0.0000	0.59	0.13	12.68	0.0004
Xq3476_1	Xq4001_1	0.203	0.59	0.20	18.04	0.0000	0.58	0.23	22.30	0.0000	0.83	0.12	13.03	0.0003
Xq3476_1	Xq3449_1	0.117	–0.61	0.20	16.81	0.0000	–0.57	0.22	21.32	0.0000	–0.68	0.22	26.39	0.0000
Xq3413_1	Xq3274_1	0.072	–0.64	0.34	29.39	0.0000	–0.56	0.21	20.53	0.0000	0.16	0.01	0.90	0.3435
Xq3471_1	Xq3274_1	0.177	–0.68	0.42	35.81	0.0000	–0.57	0.18	18.35	0.0000	–0.89	0.36	35.86	0.0000
Xq4001_1	Xq3274_1	0.139	–0.65	0.40	34.64	0.0000	–0.56	0.18	17.94	0.0000	–0.72	0.18	18.69	0.0000
Xq3449_1	Xq3274_1	0.225	0.66	0.37	30.02	0.0000	0.55	0.16	15.98	0.0001	0.80	0.37	43.37	0.0000
Xq3413_1	Xq2816_1	0.044	–0.57	0.30	24.57	0.0000	–0.53	0.16	13.54	0.0002	0.23	0.02	2.53	0.1114
Xq3471_1	Xq2816_1	0.149	–0.71	0.39	31.20	0.0000	–0.55	0.14	12.20	0.0005	–0.85	0.31	31.00	0.0000
Xq4001_1	Xq2816_1	0.111	–0.65	0.35	28.37	0.0000	–0.54	0.14	11.81	0.0006	–0.80	0.17	17.83	0.0000
Xq3449_1	Xq2816_1	0.197	0.69	0.34	26.75	0.0000	0.52	0.12	10.34	0.0013	0.84	0.43	48.40	0.0000
Xq3476_1	Xq2816_1	0.314	–0.30	0.06	4.98	0.0256	–0.27	0.04	3.70	0.0544	–0.42	0.13	14.68	0.0001
Xq3476_1	Xq3274_1	0.342	–0.28	0.05	4.52	0.0335	–0.17	0.02	2.18	0.1400	–0.44	0.15	17.09	0.0000

The markers are ordered by the χ^2 values in the Finnish population.

telomeric, with marker Xq3812-1 displaying little LD with other markers in the region. The reduced heterozygosity for a few SNPs in the Sardinian sample lead to small expected cell counts for the χ^2 for a few comparisons. For such comparisons, significance levels from Fisher's exact test led to the same general pattern of LD observed from the χ^2 statistic. For example, for Sardinia (Fig. 2), if the exact test was used, all shaded boxes remained shaded with *P* values of 0.025 or smaller, with the exception of the pairing of Xq3862-1 and Xq3773-1.

Haplotype analyses were performed and the results for regions of extensive LD are shown (Table 3). In regions from Xq28 and Xq25, the number of haplotypes found was reduced, and observed haplotype frequencies differed from those expected under linkage equilibrium. In the Xq25 LD regions, a single haplotype consisting of 7 SNPs was found in 63% of the CEPH, 55% of the Finnish and 82% of the Sardinian samples (markers Xq4007-1 to Xq3917-1), compared with expected frequencies of 21%, 14% and 54%, respectively. In the Xq28 LD regions, the Sardinian sample had its own predominant haplotype (2122111, 58%; 2122211, 5%), which differed by one allele from the most frequent in the CEPH (2122111, 6%; 2122211, 44%) and Finnish (2122111, 2%; 2122211, 37%) samples. In each group, the observed haplotype frequency for the most common haplotype was greater than the frequency expected under linkage equilibrium. It is also noteworthy that individual haplotypes show great frequency variations in each sample, probably reflecting local founder events in the isolated populations.

Reduced haplotype diversity and increased occurrence of select haplotypes may have several explanations. First, it is possible in founder populations that random selection at founding or subsequent genetic drift may have reduced the spectrum of haplotypes. For example, rare disease genes preserved in single founder haplotypes have been enriched in Finland as a result of founder effects[8–10]. Second, it is possible that there is selection for favoured haplotypes (as proposed for the HLA region[11]). Third, non-random sampling based on an unknown confounding factor (for example, a disease gene or population admixture) would cause a high degree of LD. Neither of these sampling artefacts is likely for our sample sets because they have been used for other studies previously without exhibiting these artefacts.

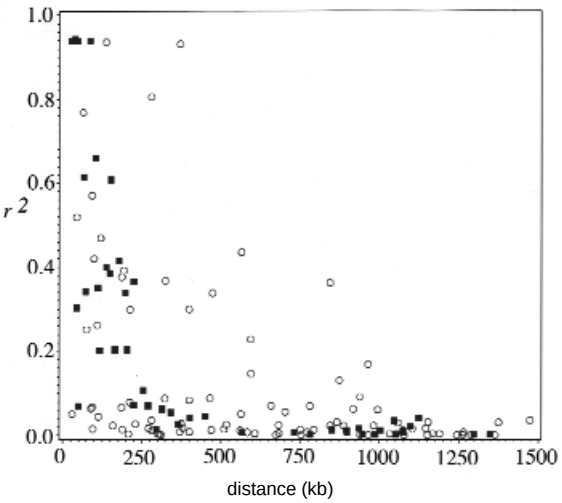

Fig. 1 A plot of pair-wise linkage disequilibrium (r^2) versus distance (kb) within the Xq25 (open circle) and Xq28 regions (filled square), in the CEPH population. Plots for the Finnish and Sardinian populations were similar.

Juxtaposed to the LD regions were regions that exhibited little or no LD (Fig. 2). One region includes markers Xq3562-1, Xq3655-2 and Xq3656-1 in Xq25 (129 kb). In this region, only 1 of 6 possible pair-wise comparisons in the 2 samples showed significant LD, despite the fact that the inter-marker distance between any pair of markers was less than 70 kb. Another region includes four markers (Xq4095-1 to Xq3698-1), also in Xq25 (308 kb). This region contains four markers immediately telomeric to the Xq25 LD region described above.

The heterozygosities of all SNP markers in the LD regions were almost identical in the CEPH and Finnish samples, but a drop in the heterozygosity values was seen in the Sardinian sample (Fig. 3) for markers in the Xq25 and Xq28 LD regions. As this drop in heterozygosity was not seen at the ends of the LD regions or in markers genotyped from other regions, it is not due to a general

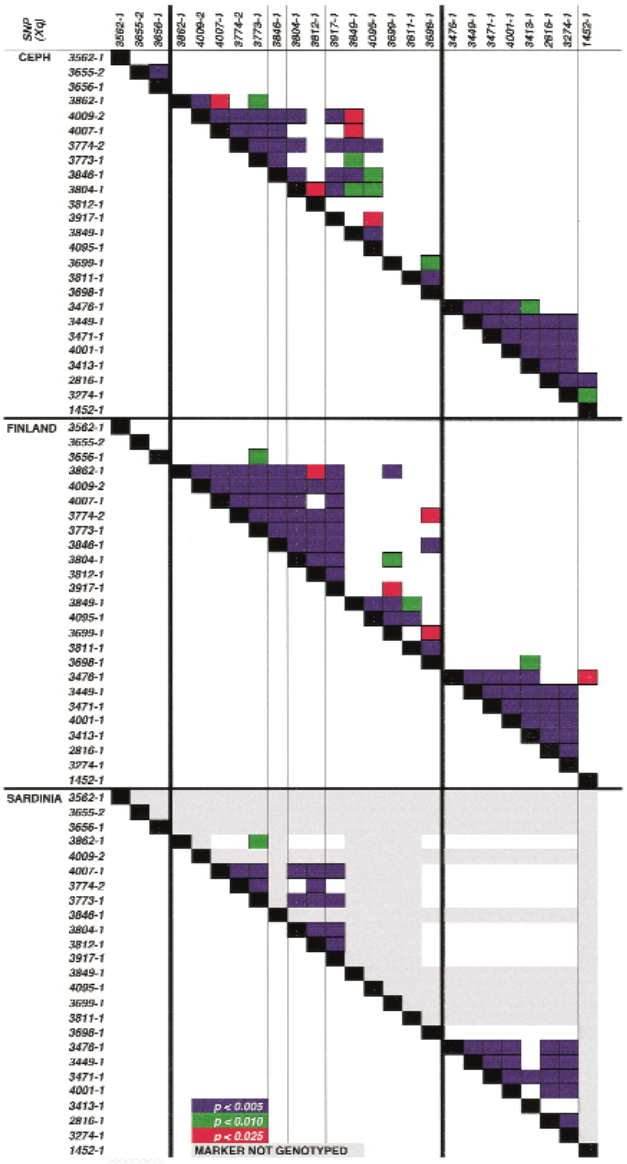

Fig. 2 Linkage disequilibrium map of targeted regions in Xq25 and Xq28. Each SNP marker is listed across the top and down the left side. If significant LD is present between two markers, the box where they intersect is filled. Purple boxes, significant LD at *P*<0.005; green boxes, LD at *P*<0.010; red boxes, LD at *P*<0.025. Heavy black lines indicate that genotyped SNP markers were dropped to simplify the figure. A narrow black line inserted between two markers indicates that they are >100 kb apart. Grey boxes represent markers that were not genotyped in that population.

reduction in genetic variability of the Sardinian sample. This drop in heterozygosity may be due to selection for a specific haplotype in the Sardinian population in these two regions or to a founder effect. Considering the differences in terms of age and history of the Sardinian population compared with other two populations, such divergence is not surprising.

Our work is a comprehensive study of LD with informative SNP markers not associated with any disease phenotype. We have shown that LD is more a property of the chromosomal region than the sole reflection of the physical distance between genetic markers. Juxtaposed to regions of strong LD, we observed regions of equilibrium between markers in close proximity to each other, further suggesting that LD strength and extent may differ greatly by chromosomal region. Extended regions of significant LD were observed, indicating that there is not a simple monotonic relationship between distance and LD. The extended LD regions are independent of the GC content or gene density. Furthermore, the extended regions of LD and equilibrium are generally the same across three samples with distinct population histories (including two isolated populations). This may be due to our choice of informative markers, which are believed to be more ancient. This is the first step in establishing a whole-genome LD map for a detailed look at LD across the human genome. A whole-genome map, indicating the magnitude of background linkage disequilibrium, will determine whether our observations on the X chromosome will apply to the autosomal chromosomes, and will guide the rational design of large-scale association studies.

Table 3 • Major haplotypes

	CEPH		Finland		Sardinia	
Major observed haplotypes[a]	Observed haplotype frequency[b] (%)	Expected haplotype frequency[c] (%)	Observed haplotype frequency[b] (%)	Expected haplotype frequency[c] (%)	Observed haplotype frequency[b] (%)	Expected haplotype frequency[c] (%)
Xq28: markers Xq3476-1, Xq3449-1, Xq3471-1, Xq4001-1, Xq3413-1, Xq2816-1, Xq3274-1						
2122211	44	9	37	7	5	2
2122111	6	4	2	3	58	18
1211122	13	0	12	0	0	0
1122211	15	5	7	3	2	1
2122222	5	1	17	3	0	0
1122111	0	2	0	1	21	11
Xq25: markers Xq4007-1, Xq3774-2, Xq3773-1, Xq3804-1, Xq3812-1, Xq3917-1						
111111	63	21	55	14	82	54
222111	10	1	13	2	0	0
222212	6	0	11	0	0	0

[a]Haplotypes observed in at least 10% of the samples in one or more populations. [b]In the Xq28 LD region, only 12 of 128 possible haplotypes were found in the CEPH sample (16 in the Finnish sample and 9 in the Sardinian sample); in the Xq25 region, only 11 of 64 possible haplotypes were found in the CEPH sample (10 in the Finnish sample and 8 in the Sardinian sample). [c]Frequency expected assuming linkage equilibrium among the listed markers.

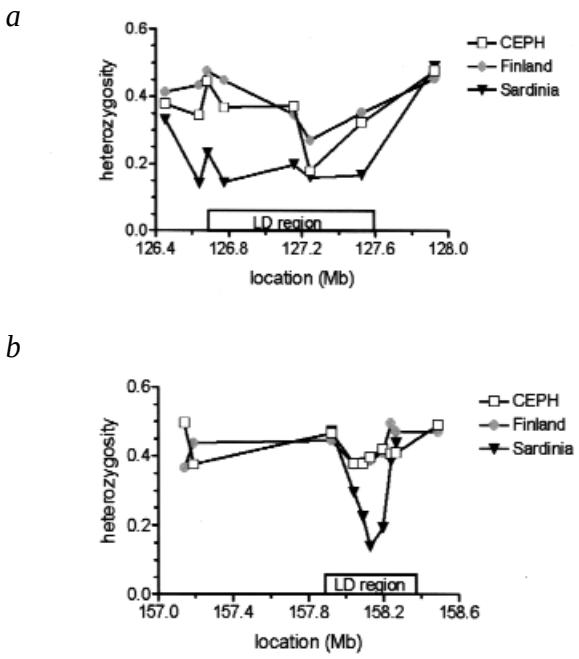

Fig. 3 A plot of the heterozygosity values for each marker in the three populations across the Xq25 (***a***) and Xq28 (***b***) LD regions. The Xq25 plot includes the seven markers that were genotyped in all three populations (3862, 4007, 3774, 3773, 3804, 3812, 3917). The Xq28 plot includes all markers from Xq28 (3840, 3841, 3476, 3449, 3471, 4001, 3413, 2816, 3274, 1452).

Methods

SNP discovery and genotyping. Methods for SNP discovery and genotyping have been described[12–14]. Only informative SNPs with estimated minor allele frequencies greater than 20% in the CEPH parents were used here, except one case (marker Xq3812-1, 15% estimated frequency in CEPH parents) for which no suitable markers were found in the region. Each SNP was placed on the X map and assigned a Mb distance[7].

We developed a short genotyping PCR assay for all SNPs used here. Most SNPs were genotyped using homogeneous, single-base extension assay with fluorescence resonance energy transfer detection[15] (FRET–TDI assay). In the early days of the project, a number of markers were genotyped with the oligonucleotide ligation assay[16] (OLA).

DNA samples. We genotyped 92 CEPH male DNAs, 100 Finnish male DNAs and 150 Sardinian male DNAs. The Finnish samples represented healthy males from a late-inhabited region of Finland, Kainuu, settled by hundreds rather than thousands of founders about 400 years ago. This population underwent an approximately tenfold growth during the past 250 years[17]. We collected Sardinian DNA samples from 150 healthy males living in Sardinian towns scattered all over the island. The first inhabitants of Sardinia were pre-Neolithic and we have some estimates that in the late Paleolithic era, the Sardinian population may have consisted of up to 700–1,800 individuals[18]. The relative isolation of Sardinia, which is 200 km from both the Italian coast and the North African coast, generated an endogenous population that experienced little admixture of genes from immigrants or invading forces. Many genes deviate grossly from average gene frequencies in any other population in Europe or Africa[18]. On the basis of these data, Sardinians consists of a major 'outlier' group in Europe, second, in order of divergence, only to the Lapps. Also, founder effects have been demonstrated for monogenic disorders like thalassaemia and Wilsons disease. In these cases, a single mutation accounts for most of the defects detected in Sardinia, whereas the same defect is rare in most other places[19].

Our CEPH panel included a mixture of European male grandparents and parents; 72 from Utah, 14 French, 3 Venezuelan and 3 Amish. Thirty-two families were represented in our sample. The average number of individuals per family was 3 and the range was 1–4 individuals. Only one family was represented by four individuals.

Analysis for linkage disequilibrium. We computed pair-wise linkage disequilibrium for all possible two-way comparisons of SNPs within each sample using SAS (SAS Institute). Significance levels were determined by the χ^2 statistic for the corresponding 2×2 table (1 degree of freedom).

Several widely used measures of linkage disequilibrium were also com-

puted: D, D´ and r^2. For two loci L_1 and L_2, each with two alleles 1 and 2, let p_i be the frequency of allele 1 and $q_i=1–p_i$ be the frequency of allele 2, at locus i (i=1, 2). Let p_{11} be the frequency of the 11 haplotype and in general let p_{jk} be the frequency of the jk haplotype. The coefficient of disequilibrium, D, is the difference between the observed haplotype frequency and the frequency expected under statistical independence: $D=p_{11}–p_1\,p_2$. The labelling of the alleles may affect the sign of D, but not its absolute value. The normalized disequilibrium coefficient is obtained by dividing D by its maximum possible (absolute) value: $D´=D/|D|_{max}$, where $|D|_{max}=\max(p_1\,p_2, q_1q_2)$ if D<0 and $|D|_{max}=\min(q_1p_2, p_1q_2)$ if D>0. The correlation coefficient is $r^2=D^2/(p_1\,p_2\,q_1\,q_2)$.

The choice of disequilibrium measure has been debated as interest in disequilibrium mapping has grown. Because D and r^2 are frequency-dependent, they can lead to false inferences[20]. Although there are no generally frequency-independent measures of disequilibrium[21], the standardized D´ is often used, as its range ($–1\le D´\le 1$) is independent of allele frequencies. Our study reports values for D´, r^2 and χ^2, and *P* values corresponding to χ^2.

Once regions of high and of low disequilibrium were identified, haplotypes were constructed for further comparison. Of the total number of haplotypes possible for a given marker set, the percentage observed in each sample was recorded. Haplotype frequencies were computed and compared with frequencies expected under linkage equilibrium, computed from SNP allele frequencies. We compared SNP heterozygosities across populations. Heterozygosity for a biallelic marker with allele frequencies p and q=1–p is given by $H=1–p^2–q^2=2p(1–p)$.

Acknowledgements

We thank D. Schlessinger for encouragement and support; R.D. Miller for discussions; and P. Buzby and S. Spurgeon for reagents. This work was supported in part by grants from the National Human Genome Research Institute (RO1HG1439, P50HG00201, P50HG00835) and by MH37685 and MH17104.

Received 14 January; accepted 23 May 2000.

1. Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genet*. **21** (suppl.), 56–60 (1999).
2. Hartl, D.L. & Clark, A.G. *Principles of Population Genetics* (Sinauer Associates, Sunderland, 1989).
3. Watkins W.S. *et al*. Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. *Am. J. Hum. Genet*. **55**, 348–355 (1994).
4. Jorde, L.B. *et al*. Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am. J. Hum. Genet*. **54**, 884–898 (1994).
5. Kruglyak, L. Prospects for whole-genome genome linkage disequilibrium mapping of common disease genes. *Nature Genet*. **22**, 139–144 (1999).
6. Jorde, L.B., Watkins, W.S., Viskochil, D., O'Connell, P. & Ward, K. Linkage disequilibrium in the neurofibromatosis 1 (NF1) region: implications for gene mapping. *Am. J. Hum. Genet*. **53**, 1038–1050 (1993).
7. Taillon-Miller, P. & Kwok, P.-Y. A high density single nucleotide polymorphism map of Xq25–Xq28. *Genomics* **65**, 195–202 (2000).
8. Kere, J. *et al*. Cystic fibrosis in a low-incidence population: two major mutations in Finland. *Hum. Genet*. **93**, 162–166 (1994).
9. Tahvanainen, E. *et al*. The gene for a recessively inherited human childhood progressive epilepsy with mental retardation maps to the distal short arm of chromosome 8. *Proc. Natl Acad. Sci. USA* **91**, 7267–7270 (1994).
10. Höglund, P. *et al*. Fine mapping of the congenital chloride diarrhea gene by linkage disequilibrium. *Am. J. Hum. Genet*. **57**, 95–102 (1995).
11. Huttley, G.A., Smith, M.W., Carrington, M. & O'Brien, S.J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).
12. Taillon-Miller, P., Piernot, E.E. & Kwok, P.-Y. Efficient approach to unique single nucleotide polymorphism discovery. *Genome Res*. **9**, 499–505 (1999).
13. Taillon-Miller, P. *et al*. The homozygous complete hydatidiform mole: a unique resource for genome studies. *Genomics* **46**, 307–310 (1997).
14. Kwok, P.-Y., Carlson, C., Yager, T., Ankener, W. & Nickerson, D.A. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23**, 138–144 (1994).
15. Chen, X., Zehnbauer, B., Gnirke, A. & Kwok, P.-Y. Fluorescence energy transfer detection as a homogeneous DNA diagnostic method. *Proc. Natl Acad. Sci. USA* **94**, 10756–10761 (1997).
16. Kwok, P.-Y., Gremaud, M.F., Nickerson, D.A., Hood, L. & Olson, M.V. Automatable screening of yeast artificial-chromosome libraries based on the oligonucleotide-ligation assay. *Genomics* **13**, 935–941 (1992).
17. Laitinen, T. *et al*. Genetic control of serum IgE levels and asthma: linkage and linkage disequilibrium studies in an isolated population. *Hum. Mol. Genet*. **6**, 2069–2076 (1997).
18. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. Demic expansions and human evolution. *Science* **259**, 639–646 (1993).
19. Cao, A., Galanello, R. & Rosatelli, M.C. Genotype-phenotype correlations in β-thalassemias. *Blood Rev*. **8**, 1–12 (1994).
20. Hedrick, P.W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341 (1987).
21. Lewontin, R.C. On measures of gametic disequilibrium. *Genetics* **120**, 849–852 (1988).