Demographic history and linkage disequilibrium in human populations

Maris Laan & Svante Pääbo

In the human genome, linkage disequilibrium (LD)—the non-random association of alleles at chromosomal loci[1]—has been studied mainly in regions surrounding disease genes on affected chromosomes[2–6]. Consequently, little information is available on the distribution of LD across anonymous genomic regions in the general population. However, demographic history is expected to influence the extent of overall LD across the genome, so a population that has been of constant size will display higher levels of LD than a population that has expanded[7]. In support of this, the extent of LD between anonymous loci on chromosome 4 in chimpanzees (as a model of a population of constant size) has been compared to that in Finns (as a model of an expanded population; refs 8,9) and found to exhibit more LD than in the latter population. In Europe, studies of mitochondrial (mt) DNA sequences have suggested that most populations have experienced expansion[10], whereas the Saami in northern Fenno-Scandinavia have been of constant size (Table 1). Thus, in northern Europe, populations with radically different demographic histories live in close geographic proximity to each other. We studied the allelic associations between anonymous microsatellite loci on the X chromosome in the Saami and neighbouring populations and found dramatically higher levels of LD in the Saami than in other populations in the region. This indicates that whereas recently expanded populations, such as the Finns, are well suited to map single disease genes affected by recent mutations, populations that have been of constant size, such as the Saami, may be much better suited to map genes for complex traits that are caused by older mutations.

Seven dinucleotide repeat loci, spanning approximately 4.0 cM or 12.5 Mb on Xq13 (refs 11,12), were typed in Saami, Finnish, Estonian and Swedish males. For Saami, two additional loci (*DXS8107* and *DXS1066*) were typed (Fig. 1). Features of the variability of the seven loci typed in all populations are given in Table 2. When haplotypes composed of the seven microsatellite loci were constructed, it was found that among the Estonians and Swedes, every individual studied carried a unique haplotype; among the Finns, the variation was almost as high (75 haplotypes among 80 individuals). In contrast, Saami displayed 32 haplotypes among 54 individuals. It is noteworthy that the reduction in numbers of haplotypes in the Saami is not the result of a lower number of alleles, as the locus diversity for all northern European populations is of similar magnitude (Table 2). Furthermore, mean allele size variance ratios across loci—a measure of genetic diversity of the genomic region[13]—are similar for all populations (0.929–1.075), and the mean number of loci differing between haplotypes in Saami is similar (4.59) to that in the other populations (4.97–5.15). However, the variance of the haplotype mismatch for Saami is twice as high (3.08) as for the neighbouring populations (1.38–1.54). Thus, the Saami are characterized by groups of related haplotypes that differ substantially from one another.

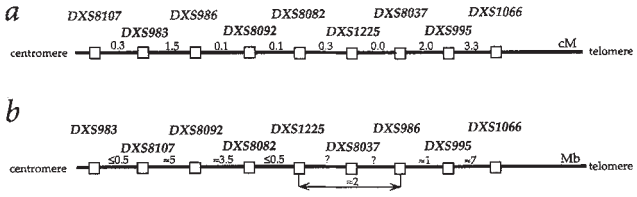

Fig. 1 Genetic (***a***, ref. 11) and physical (***b***; ref. 12) maps of microsatellite loci in the Xq13 region used in this study.

Fisher's exact test was used to detect LD between loci (Table 3). In Saami, allelic associations ($P<0.05$) were detected for 25/36 pairs of loci studied. The largest genetic distances between the loci in LD were 5.6–5.8 cM (three pairs), whereas the largest physical distances were 10.5–14 Mb (five pairs). No locus was preferentially involved in creating LD (Table 4). When the loci were stratified into those located closely on the genetic (≤ 2 cM) or physical (≤ 2 Mb) maps and those that were not (>2 cM, >2 Mb), according to the same arbitrary criteria applied in a previous study of LD on chromosome 4 (ref. 8), LD was found to be inversely correlated with genetic (Fisher's exact test, $P=0.026$) but not with physical ($P=1.00$) distances. In all other populations studied, only 1–3 of the 21 pairs of loci showed LD ($P<0.05$), and an additional 1–3 pairs were suggestive of LD ($0.05 \le P<0.1$). Thus, in Saami, LD exists over a 5.8-cM/14-Mb region; in other populations, few associations were observed even between closely linked loci.

LD can be caused by one or more of five factors: (i) recent mutations at a locus, (ii) population founder effects, (iii) admixture between populations, (iv) selection and (v) demographic history[1,4,7,14,15]. Recent mutations in the Saami cannot explain the large extent of observed LD, as several pairs of loci contribute to the creation of non-random allelic associations. Furthermore, founder effects and admixture find no support in historical or

Table 1 • MtDNA sequence variation

Population	Sample size (n)	Variable positions (S_{obs})	Mean pairwise difference	*P* value
Finnish	133	59	3.95	0.0002
Estonian	48	41	4.09	0.0055
Swedish	32	38	4.57	0.0160
Saami	25	17	3.25	0.1943

For a population that has been of constant size, the expected number of variable positions (S_{exp}) can be calculated with coalescent theory as $S_{exp}=\theta \sum_{i=1}^{n-1} (1/i)$, where n is the number of DNA sequences sampled from the population and the mean pairwise sequence difference can be taken as an estimate of θ. Given this estimation of θ, the distributions of the number of variable positions were generated by computer simulations and compared to the observed number of variable positions (S_{obs}; refs 23,24). If the probabilities of S_{obs} given this distribution were<0.05, the hypothesis of constant population size was rejected. The DNA sequences have been described[10,20].

Institute of Zoology, University of Munich, P.O. Box 202136, D-80021 Munich, Germany. Correspondence should be addressed to S.P. e-mail: paabo@zi.biologie.uni-muenchen.de

Table 2 • Features of seven dinucleotide loci used in this study[a]

		Population				
		Saami	Finns	Swedes	Estonians	All
	Sample size	54	80	41	45	220
AFM078za1						
Size range: 166–188 bp	No. of alleles	4	6	6	4	6
Locus: *DXS983*	Mean allele size	179.85	179.18	178.95	179.22	179.32
	Variance of allele size	1.60	3.88	4.10	1.49	2.94
	Diversity	0.579	0.696	0.619	0.482	0.617
AFM116xg1						
Size range: 148–176 bp	No. of alleles	10	11	9	9	16
Locus: *DXS986*	Mean allele size	159.19	160.58	160.85	160.89	160.36
	Variance of allele size	27.46	14.25	12.28	13.69	17.18
	Diversity	0.799	0.800	0.708	0.796	0.810
AFMa040ye1						
Size range: 262–289 bp	No. of alleles	8	12	11	10	13
Locus: *DXS8092*	Mean allele size	271.74	273.51	274.42	273.42	273.23
	Variance of allele size	24.35	19.32	21.40	19.70	21.56
	Diversity	0.825	0.853	0.894	0.878	0.873
AFMc020wb9						
Size range: 211–231 bp	No. of alleles	6	9	7	7	10
Locus: *DXS8082*	Mean allele size	219.70	221.49	222.07	222.68	221.40
	Variance of allele size	43.95	33.16	36.22	32.36	36.9
	Diversity	0.787	0.748	0.759	0.769	0.762
AFM311vg5						
Size range: 192–224 bp	No. of alleles	6	9	8	7	10
Locus: *DXS1225*	Mean allele size	203.19	205.05	203.22	203.87	204.01
	Variance of allele size	33.89	38.38	51.98	50.25	42.32
	Diversity	0.755	0.730	0.816	0.766	0.797
AFM285xg5						
Size range: 238–260	No. of alleles	5	8	7	6	9
Locus: *DXS8037*	Mean allele size	250.04	250.20	249.45	248.93	249.76
	Variance of allele size	5.61	10.69	14.97	16.65	11.56
	Diversity	0.499	0.712	0.778	0.761	0.695
AFM207zg5						
Size range: 191–201	No. of alleles	3	6	5	5	6
Locus: *DXS995*	Mean allele size	194.07	194.68	194.22	194.38	194.38
	Variance of allele size	3.43	5.41	4.18	4.69	4.55
	Diversity	0.435	0.628	0.564	0.564	0.557
Mean variance ratio across loci (R_j)		0.929	0.995	1.075	0.965	
Mean haplotype mismatch[b]		4.59	5.15	5.09	4.97	
Variance of haplotype mismatch		3.08	1.38	1.54	1.43	

[a]For Saami, two additional microsatellites were analysed: AFMa074zd9 (*DXS8107*) and AFM234tf8 (*DXS1066*), with size ranges 169–187 bp and 259–263 bp, respectively. [b]Mean number of loci with different alleles between individuals within a population.

other genetic data, whereas selection is highly unliklely as a cause for the high levels of LD in Saami because their locus diversity is only moderately reduced, and selective factors affecting the Saami but not geographically adjacent populations are difficult to envision. We therefore conclude that differences in demographic history are best able to explain the larger extent of LD in Saami than in the other populations. In this respect, it is highly relevant that mtDNA sequence analyses (Table 1) indicate that the Saami have been of constant population size, whereas the other populations studied have not. Thus, the predicted effect of demographic history on the extent of LD[7] is dramatically borne out not only when humans are compared to chimpanzees[9] but also when geographically adjacent human populations are compared.

Several studies[2,3,16,17] have used LD in Finns to map genes for rare monogenic diseases, which have been found to be in LD with polymorphic loci across 2–11 cM. At first, this might seem to be in contradiction to the low level of LD at Xq13 in Finns. However, historical[18,19] as well as genetic[20] data indicate that Finns have experienced a population expansion that is more extreme than that of other populations in the region. We note that alleles carrying disease mutations that have occurred during a population expansion will be surrounded by large regions of LD; in contrast, random loci will display low levels of LD[7]. In recently expanded populations, the background level of LD is low, while

Table 3 • Allelic association between loci

Locus pair	Mb[a]	cM[b]	*P* values Saami	Finns	Estonians	Swedish
DXS1066–DXS983	19.5	7.3	0.166	N.D.	N.D.	N.D.
DXS1066–DXS8107	19	7.6	0.460	N.D.	N.D.	N.D.
DXS1066–DXS8092	14	5.7	**0.001**	N.D.	N.D.	N.D.
DXS995–DXS983	12.5	4	**0.012**	0.508	0.314	0.593
DXS995–DXS8107	12	4:3	0.242	N.D.	N.D.	N.D.
DXS986–DXS983	11.5	1.5	**0.000**	0.829	0.468	0.400
DXS986–DXS8107	11	1.8	**0.017**	N.D.	N.D.	N.D.
DXS1066–DXS8082	10.5	5.6	**0.001**	N.D.	N.D.	N.D.
DXS1066–DXS1225	10	5.3	0.252	N.D.	N.D.	N.D.
DXS8037–DXS8107	9–11	2.3	0.586	N.D.	N.D.	N.D.
DXS8037–DXS983	9.5–11.5	2.0	0.300	0.683	0.104	0.924
DXS1225–DXS983	9.5	2.0	**0.000**	0.630	0.520	0.480
DXS8082–DXS983	9	1.7	**0.000**	0.565	0.730	0.082
DXS1225–DXS8107	9	2.3	**0.035**	N.D.	N.D.	N.D.
DXS1066–DXS8037	8–10	5.3	0.060	N.D.	N.D.	N.D.
DXS8082–DXS8107	8.5	2.0	**0.002**	N.D.	N.D.	N.D.
DXS1066–DXS986	8	5.8	**0.031**	N.D.	N.D.	N.D.
DXS1066–DXS995	7	3.3	0.113	N.D.	N.D.	N.D.
DXS995–DXS8092	7	2.4	0.104	0.115	0.732	0.425
DXS8092–DXS986	6	0.1	**0.000**	0.331	0.100	0.332
DXS8092–DXS983	5.5	1.6	**0.000**	0.314	0.153	0.746
DXS8092–DXS8107	5	1.9	**0.018**	N.D.	N.D.	N.D.
DXS8037–DXS8092	4–6	0.4	**0.000**	0.180	0.072	0.028
DXS1225–DXS8092	4	0.4	**0.000**	0.283	0.120	0.676
DXS995–DXS8082	3.5	2.3	**0.000**	0.128	0.237	0.490
DXS8082–DXS8092	3.5	0.1	**0.000**	**0.044**	0.065	0.102
DXS995–DXS1225	3	2	**0.001**	0.154	0.591	0.563
DXS8082–DXS986	2.5	0.2	**0.000**	0.092	0.143	0.618
DXS1225–DXS986	2	0.5	**0.000**	0.393	0.688	0.448
DXS995–DXS8037	1–3	2.0	0.124	0.874	0.615	0.225
DXS8037–DXS8082	0.5–2.5	0.3	**0.012**	0.238	0.625	**0.033**
DXS995–DXS986	1	2.5	**0.000**	0.729	0.829	0.429
DXS1225–DXS8082	≤ 0.5	0.3	**0.000**	**0.000**	**0.000**	**0.000**
DXS983–DXS8107	≤ 0.5	0.3	**0.000**	N.D.	N.D.	N.D.
DXS8037–DXS986	?	0.5	**0.000**	0.61956	0.739	0.256
DXS8037–DXS1225	?	0.0	0.091	0.836	0.488	0.242

[a]Approximate distances in Mb calculated from a YAC/STS map[14]. [b]Distances in cM according to the Généthon genetic map[11]. Significant allelic associations ($P<0.05$) are given in bold, and associations suggestive of LD ($P<0.1$) are underlined; N.D., not determined.

the extent of LD around disease alleles that were created during the expansion is large. Thus, populations that have expanded (for example, the Finns) are much better suited for the mapping of monogenic traits caused by recent mutations than are populations that have been of constant size (for example, the Saami). Of interest is that the situation is probably the reverse for alleles involved in complex traits. Such alleles, which are generally common in the population, are likely to be old. They will therefore be surrounded by large areas of LD in constant populations, whereas they will display very little LD in a recently expanded population. Thus, while populations such as the Finns are well suited for the mapping of rare diseases that stem from recent mutations, populations such as the Saami may be much better suited for the mapping of genes involved in complex traits such as susceptibility to common diseases[25].

Methods

Locus typing. Finnish (n=80), Swedish (n=41) and Estonian (n=45) males were genotyped for seven, Saami males from Sweden (n=54) for nine, dinucleotide microsatellites on Xq13 (Fig. 1, Table 1). The loci and primers were chosen from the Généthon genetic map[11]. Distances between loci, and in two cases their order, differ from those on a physical map[12]. However, when *DXS986*, which differs drastically in its location on the two maps, was eliminated from the analyses, results were not affected. PCR was performed with a hot-start procedure in a total volume of 20 µl, containing 40–80 ng of genomic DNA, 50 nM of each primer (one of which was

Table 4 • Loci displaying linkage disequilibrium

		Finns, Estonians and Swedes						
		DXS983	*DXS986*	*DXS8092*	*DXS8082*	*DXS1225*	*DXS8037*	*DXS995*
	DXS983		—	—	—	—	—	—
	DXS986	Sa		—	—	—	—	—
	DXS8092	Sa	Sa		F	—	S	—
Saami	*DXS8082*	Sa	Sa	Sa		F,E,S	S	—
	DXS1225	Sa	Sa	Sa	Sa		—	—
	DXS8037	—	Sa	Sa	Sa	—		—
	DXS995	Sa	Sa	—	Sa	Sa	—	

Pairs of loci displaying LD in Saami (Sa), Finns (F), Estonians (E) and Swedes (S).

fluorescently labelled), 0.125 mM dNTP, 1 U of *Taq* polymerase, 10 mM Tris (pH=9.0), 50 mM KCl, 1.5 M $MgCl_2$, 0.1% Triton X-100, 0.01% gelatin and 1 mg/ml BSA. An initial denaturation at 96 °C for 5 min was followed by 35 cycles of denaturation (94 °C for 40 s) and annealing (55 °C for 30 s). Subsequently, an elongation step (72 °C for 2 min) was performed. For AFMc020xb9 and AFM311vg5, the annealing temperature was 52 °C. Amplicons were electrophoresed on an automated laser fluorescence DNA sequencer and molecular sizes determined with Fragment Manager software (Pharmacia). A total of eleven alleles could not be determined. Haplotypes were constructed manually.

Data analysis. Allele frequencies for each microsatellite locus were estimated by gene counting. Locus diversity was calculated as $n/(n-1)\times(1-\sum_i p_i^2)$, where p_i is the estimated frequency of the i^{th} allele at the locus[1]. Alternatively, the variance of the allele size distribution (V_{ij}, for locus i in population j) was used. To analyse variance across loci, the average ratio $R_j=\frac{1}{7}\sum_{i=1}^{7} V_{ij}/V_i$ was calculated[13], where V_i is the allele size variance for locus i across all populations.

To describe the extent of non-random allelic association between pairs of loci, the tail probability (*P* value) of Fisher's exact test was computed with GenePop software[21]. The advantage of using this approach for detecting LD when loci with different numbers of alleles are compared has been discussed[7,8]. Briefly, for each pair of loci, an r × c contingency table of gametes was formed and 1,000 tables with the same marginal totals were generated on the basis of a Markov chain algorithm[22]. The *P* value, which should not be interpreted as indicative of statistical significance in a formal sense, is the fraction of such tables that were equal to or less likely than the observed table.

Acknowledgements

We thank J. Terwilliger for both alerting us to and helping us think about the issues concerning complex traits. We are also indebted to L. Beckman, A. Kindmark, L. Peltonen and the Tallinn Blood Centre for DNA and blood samples; to C. Kilger, M. Markmann, A. M'charek, A.-H. Salem, M. Stoneking, R. Ward and especially G. Weiss and S. Zöllner for discussion and help; to the Alexander-von-Humboldt Stiftung for a fellowship (to M.L.); and to the Deutsche Forschungsgemeinschaft for financial support (Pa 452/3-1).

Received 30 May; accepted 15 September 1997.

1. Weir, B.S. *Genetic Data Analysis* (Sinauer Associates, Sunderland, Massachusetts, 1996).
2. Hästbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
3. Peltonen, L., Pekkarinen, P. & Aaltonen J. Messages from an isolate: lessons from the Finnish gene pool. *Biol. Chem. Hoppe-Seyler* **376**, 697–704 (1995).
4. Jorde, L.B. Linkage disequilibrium as a gene-mapping tool. *Am. J. Hum. Genet.* **56**, 11–14 (1995).
5. Kaplan, N.L., Hill, W.G. & Weir, B.S. Likelihood methods for locating disease genes in nonequilibrium populations. *Am. J. Hum. Genet.* **56**, 18–32 (1995).
6. Terwilliger, J.D. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* **56**, 777–787 (1995).
7. Slatkin, M. Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336 (1994).
8. Peterson, A.C. *et al.* The distribution of linkage disequilibrium over anonymous genome regions. *Hum. Mol. Genet.* **4**, 887–894 (1995).
9. Crouau-Roy, B., Service, S., Slatkin, M. & Freimer, N. A fine-scale comparison of the human and chimpanzee genomes: linkage, linkage disequilibrium and sequence analysis. *Hum. Mol. Genet.* **5**, 1131–1137 (1996).
10. Sajantila, A. *et al.* Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res.* **5**, 42–52 (1995).
11. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
12. Nagaraja, R. *et al.* X chromosome map at 75-kb STS resolution, revealing extremes of recombination and CG content. *Genome Res.* **7**, 210–222 (1997).
13. Jorde, L.B. *et al.* Microsatellite diversity and the demographic history of modern humans. *Proc. Natl. Acad. Sci. USA* **94**, 3100–3103 (1997).
14. Thompson, E.A. & Neel, J.V. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am. J. Hum. Genet.* **60**, 197–204 (1997).
15. Stephens, J.C., Briscoe, D. & O'Brien, S.J. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.* **55**, 809–824 (1994).
16. Mitchison, H.M. *et al.* Batten disease gene, *CLN3*: linkage disequilibrium mapping in the Finnish population, and analysis of European haplotypes. *Am. J. Hum. Genet.* **56**, 654–662 (1994).
17. Varilo, T. *et al.* The age of human mutation: genealogical and linkage disequilibrium analysis of the CLN5 mutation in the Finnish population. *Am. J. Hum. Genet.* **58**, 506–512 (1996).
18. Norio, R., Nevanlinna, H.R. & Perheentupa, J. Hereditary diseases in Finland. *Ann. Clin. Res.* **5**, 109–141 (1973).
19. Solantie, R. Climatic conditions for the cultivation of rye with reference to the history of settlement in Finland. *Fennoscandia Archaeologica* **V**, 3–20 (1988).
20. Sajantila, A. *et al.* Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc. Natl. Acad. Sci. USA* **93**, 12035–12039 (1996).
21. Raymond, M. & Rousset, F. GenePop., version 1.0. Institute des Sciences de l'Evolution, Université de Montpellier, France, 1994.
22. Guo, S.W. & Thompson, E.A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372 (1992).
23. Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601 (1989).
24. Watson, E. *et al.* mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.* **59**, 437–444 (1996).
25. Terwilliger, J.D., Zöllner, S., Laan, M. & Pääbo, S. Mapping genes through the use of linkeage disequilibrium generated by genetic drift *Hum. Hered.* (in the press).