Human Molecular Genetics, 2004, Vol. 13, No. 1 **25–33**
DOI: 10.1093/hmg/ddh001
Advance Access published on November 12, 2003

Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations

A. Tenesa[1,*], A.F. Wright[2], S.A. Knott[1], A.D. Carothers[2], C. Hayward[2], A. Angius[3], I. Persico[3], G. Maestrale[3], N.D. Hastie[2], M. Pirastu[3] and P.M. Visscher[1]

[1]Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK, [2]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, UK and
[3]Istituto di Genetica delle Popolazioni, CNR, Alghero, Italy

Received July 16, 2003; Revised October 6, 2003; Accepted October 22, 2003

The extent of linkage disequilibrium (LD) is an important factor when designing experiments for mapping disease or trait loci using LD mapping methods. It depends on the population history and hence is a characteristic of each population. Here, we have assessed the extent of LD in a sub-isolate of the general Sardinian population (775 members of one village) using 22 polymorphic markers on chromosome 19. We found high levels of disequilibrium that extended to 8 cM, when based on D', and 11 cM when based on the significance level of the allelic association. The fact that conclusions based on both methods are similar suggests that the estimates are quite robust. We have also shown, through a simple resampling technique, that small sample sizes can overestimate both the mean value of D' and its variance up to a factor of about 2 and 16, respectively, when the number of diplotypes (the pair of haplotypes that compose the genotype) decreased from 186 to 26. We evaluated the effect on D' of the depth of the pedigree available when using phased founders, and compared the estimates with those obtained when using unphased founders, and also the effect of grouping alleles on the value of D' and the significance level. Owing to the high sampling variance of LD, we recommend the use of at least 200 unrelated individuals when characterizing the extent of LD.

INTRODUCTION

In recent years, human geneticists have advocated the use of linkage disequilibrium, the non-random association of population allele frequencies at two or more loci, to map genes related to common human complex diseases. Linkage disequilibrium (LD) mapping relies on the assumption that there have not yet been enough generations of recombination to break down the association between a causative locus and nearby markers. The association, generated by, for example, mutation, selection, drift or migration is reduced each generation by a function of the recombination fraction between the loci and the population size (1). The closer the marker and trait locus, the larger the number of generations required to break down their association. LD will also be erased faster in large populations than in small ones. However, empirical data has shown conflicting results. Some studies (2,3) found comparable levels of LD in genetic isolates (Sardinia and Finland) and two outbred populations in the USA and UK. They argued that the number of founders in each isolate was large enough to have multiple copies of the common alleles represented in the founder population. Therefore, the recombinational history of common alleles for the four populations would date back to the same origin in the general population. However, sub-isolates of the Sardinian population have been shown to have increased LD levels compared with those of the general Sardinian population (3,4).

The village of Talana was selected as an example of a sub-isolate within the general Sardinian population. Talana is one of the most isolated villages in the Ogliastra region. It was selected because of its documented isolation until 25–40 years ago and the reduced number of founders. It has been estimated that 80% of the ~1300 people that currently live in Talana descend from eight paternal and 11 maternal lineages (5). Talana has experienced a slow population growth from the beginning of the seventeenth century to the present. The estimated population size was 200 in the middle of the seventeenth century, doubling at the end of the nineteenth century and then tripling at the end of the twentieth century.

*To whom correspondence should be addressed. Tel: +44 1316505440; Fax: +44 1316506564; Email: albert.tenesa@ed.ac.uk

Simulation studies (6–8) showed that populations maintained at constant size or showing slow population growth after their founder event followed by rapid expansion are more likely to show high levels of LD than those that experience a rapid growth immediately after their founder event. The Talana population meets these requirements and hence seems more suitable for detecting genes using linkage disequilibrium than other populations, such as the Finnish (9), who have a larger number of founders and experienced rapid growth just after their founder event.

Here, we present results about the extent of LD on chromosome 19 in the Talana population. The study was based on 775 individuals and the region studied spanned 124.4 cM with an average distance between markers of 5.92 cM. We studied the effect of the number of generations available to estimate founder (those individuals without parents recorded in our pedigree) haplotypes and the number of founder haplotypes on the measure of disequilibrium D' (10). We also studied the effect on D' of estimating population haplotype frequencies without using family information.

RESULTS

Hardy–Weinberg equilibrium proportions

A sample of 775 members, distributed in 120 families (Table 1), from the isolated Sardinian village of Talana (total population ~1300) was genotyped at 22 polymorphic loci from chromosome 19, with an average spacing of 5.92 cM, in order to characterize the extent of LD in a large autosomal region spanning 124.4 cM. We first tested for departures from Hardy–Weinberg equilibrium (HWE) proportions using the 381 founder diplotypes available. Here, we define, as founder diplotypes, the pair of haplotypes that compose the genotype of those individuals without parents in the pedigree.

Five out of 22 loci showed departures from HWE proportions after accounting for multiple testing using a Bonferroni correction (10 out of 22 before correction). They showed a deficiency of heterozygotes compared with what we would expect under HWE. Small populations maintained closed during a long period of time are expected to show a reduced amount of genetic variability due to founder and drift effects.

Table 2 shows the linkage map, the number of alleles observed in Talana (based on 381 founders) and in the CEPH (Centre d'Etudes du Polymorphisme Humain) families (based on 28 founders), as well as the observed heterozygosity in Talana and CEPH families. We compared the observed levels of heterozygosity and mean number of alleles in the Talana population (considered here to be an inbred human population because of its small effective population size and low immigration during the last two and a half centuries) with an outbred population (with a larger effective population size and relatively large immigration typical of the large cities). We used the CEPH reference families as an example of an outbred population. The mean number of alleles in Talana was 8.8 compared with 7.8 in the CEPH families. The mean difference in the number of alleles in the two populations, tested using a paired t-test, was significant at the 2% level. Also, the mean observed heterozygosity was smaller in Talana (0.66) than in the CEPH families (0.75). The mean difference was highly significant ($P < 0.0001$; paired t-test). The difference observed in the number of alleles could be due to a larger sample size for the Talana population. We tested this by bootstrapping 500 samples of different size from the Talana data. The sample sizes were 28, 50, 100, 134, 183, 150, 200, 250, 300 and 381 diplotypes. For each sample size (28–381) the mean number of alleles across the 500 bootstrapped samples was estimated at each locus and the mean number of alleles across loci for a given sample size estimated. Then we fitted a logarithmic curve to the mean number of alleles across loci obtained from the bootstrapping. Results are shown in Figure 1. Substituting the sample sizes of the CEPH (28 founders) and Talana (381 founders) in the equation shown in Figure 1 gave an expected number of alleles of 7.02 and 8.9, respectively. This showed that the difference in the number of alleles observed in the two populations could be entirely explained by the difference in sample size. These results also showed that the expected number of alleles for a sample of 28 unrelated people from Talana would have an average (over the 22 loci) of ~0.8 alleles less than in the sample from the CEPH families. Talana showed lower allele diversity when corrected for sample size and heterozygosity than an outbred population, which is consistent with the hypothesized drift and founder effects.

Table 1. Distribution of family size for the 120 families available

n	Number of families with n members	Total number of individuals in families of n members
3	18	54
4	22	88
5	18	90
6	9	54
7	13	91
8	13	104
9	8	72
10	9	90
11	2	22
12	2	24
13	2	26
14	2	28
16	2	32
	Total number of families = 120	Total number of individuals = 775

Extent of linkage disequilibrium using phased founders

Figure 2 shows how linkage disequilibrium, measured as D', decayed with genetic map distance. The mean D' was 0.143 (with a maximum of 0.356 and a minimum of 0.055). We fitted an equation of type $y = a + be^{-cx}$ using non-linear regression as implemented by the Genstat FITCURVE directive (11) where y is D' and x is genetic distance in cM. Note that $y \to a$ when $x \to \infty$ (a is the mean background level of LD) and $y \to a + b$ when $x \to 0$ ($a + b$ is the mean level of disequilibrium for loci at the same location). This model is similar to the Malecot model described in the context of the measure of association ρ (12). The fitted curve accounted for 47% of the total variance and the estimated parameters and their standard errors were 0.116 ± 0.004 for a, 0.184 ± 0.016 for b and 0.917 ± 0.012 for e^{-c}. We considered that a useful level of LD

Table 2. Linkage map, number of alleles (NA) and observed heterozygosity (OH) in Talana and the CEPH families

Marker	Linkage map (cM)	NA in Talana	NA in CEPH families	OH in Talana	OH in CEPH families
D19S886	0.0	6	5	0.65	0.66
D19S209	12.4	7	7	0.75	0.77
D19S894	17.7	12	11	0.75	0.77
D19S216	20.8	6	5	0.75	0.75
D19S884	28.3	10	10	0.67	0.86
D19S865	31.4	8	13	0.65	0.88
D19S221	36.7	11	10	0.74	0.86
D19S226	43.1	14	12	0.63	0.84
D19S566	53.0	10	9	0.79	0.86
D19S931	56.1	10	10	0.66	0.77
D19S414	62.5	7	7	0.57	0.77
D19S220	70.0	12	10	0.80	0.84
APOE	74.2	3	3	0.12	0.11
D19S420	74.2	8	7	0.70	0.79
D19S903	76.2	11	7	0.79	0.78
D19S902	83.8	11	9	0.71	0.79
D19S904	92.5	7	4	0.54	0.64
D19S888	106.2	10	7	0.65	0.81
D19S921	107.2	9	8	0.71	0.78
D19S572	109.3	10	7	0.73	0.80
D19S418	115.7	6	6	0.52	0.65
D19S210	124.4	6	6	0.69	0.73
Mean (SD)		8.8 (2.6)	7.8 (2.6)	0.66 (0.14)	0.75 (0.16)

(measured as D') for LD mapping to be effective was half the difference between the fitted maximum (0.300) and minimum (0.116) value (herein, referred to as the half-length). This value was 0.208 and corresponded to a distance of about 8 cM. We compared our definition of useful level of LD with a previous definition of useful level of LD, the swept radius (13). The useful level of disequilibrium estimated through the swept radius for our data set was 11.5 cM.

Figure 3 shows the statistical significance of the locus pairs. The statistical significance was classified as highly significant ($P \leq 0.0002$), significant ($0.0002 < P \leq 0.05$) and not significant ($P > 0.05$). The first category accounts for multiple testing using a Bonferroni correction when 231 independent tests were assumed. The numbers of locus pairs in the three classes were 55, 78 and 98, respectively. At each locus the number of adjacent loci in highly significant LD with this locus (abbreviated to Adlo in Fig. 3) was counted in each direction from the marker locus and the average for all loci obtained. Given a marker locus, the average number of markers adjacent to it that showed highly significant LD was 1.90 (after averaging in both directions) with a variance of 1.60. The average extent of LD from a given marker was estimated by multiplying the average number of markers adjacent to it that showed highly significant LD by the average distance between markers (5.92 cM). This was 11.25 cM (1.90×5.92) with a standard deviation of 7.48 cM (the standard deviation was estimated assuming that 5.92 was constant).

A total of 44 out of the 62 locus pairs (71% of the pairs) that were at a distance $\leq$20 cM were in highly significant ($P < 0.0002$) LD and only 11 locus pairs out of the 169 (6.5% of the pairs) that were at a distance >20 cM were in highly significant LD ($P < 0.0002$). One must keep in mind that choosing a different distance threshold would yield different results; however for this data set the threshold of 30 cM yielded only slightly different results (60.5 and 2% of the loci in highly significant association for distances less or more than 30 cM, respectively). The proportion of highly significant associations for unlinked loci (distances larger than 30 cM) was smaller than expected by chance, showing that the Bonferroni correction is too conservative.

Effect of the number of generations available to estimate diplotypes. Figure 4 shows how LD decayed as a function of genetic map distance for founders of type G1 (the 187 founders with children only), G2 (the 168 founders with grandchildren only) and G3 (the 26 founders with great-grandchildren). For G1, G2 and G3 the mean values of D' were 0.175, 0.189 and 0.456 respectively. We also fitted a non-linear equation of the type $y = a + b\mathrm{e}^{-cx}$ as described above. The estimated parameter values and their standard errors are shown in Figure 4.

We tested whether there was a difference in the estimates of D' when using founders G1, G2 and G3 by using a likelihood ratio test. We compared the likelihood of the full model, that is fitting three different a, b, e^{-c} and residual variances for G1, G2 and G3, and that of the reduced model in which we fitted, as with the full model, three different residual variances for G1, G2 and G3 but only one a, b, e^{-c}. The full model had 12 parameters estimated and a ln-likelihood equal to 1458 whereas for the reduced model the number of parameters fitted was 6 and it had a ln-likelihood equal to 1395. Twice the difference in ln-likelihood was compared to a chi-square distribution with 6 degrees of freedom. The full model fitted the data significantly ($P < 10^{-8}$) better than the reduced model. This suggests that the extent of LD might differ depending on the number of generations available to infer diplotypes [differences between G1 and G2 were only marginally significant ($P = 0.045$)]. However, since differences in sample size between G1, G2 and G3 might also explain those differences, this was investigated further.

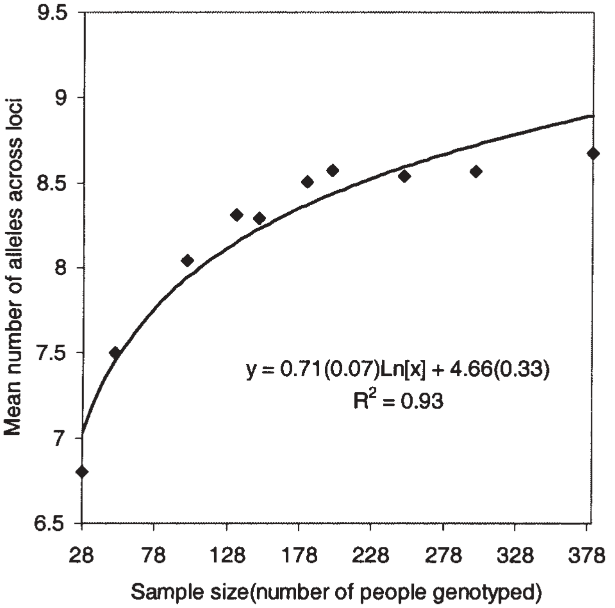

Figure 1. Expected relationship between sample size and the mean number of alleles observed for the 22 loci studied, obtained by bootstrapping. The equation of the fitted line and the standard errors (in brackets) of the estimated parameters are shown.

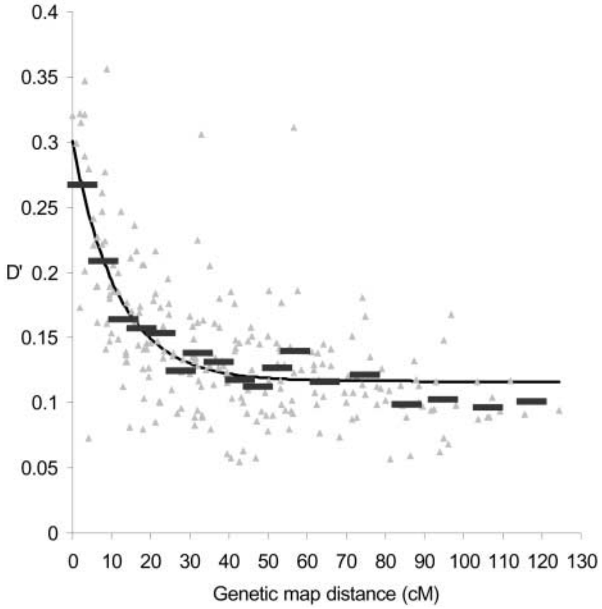

Figure 2. Decay of D′ values observed between marker loci on chromosome 19 as a function of genetic map distance (in cM). Horizontal lines represent the mean of *D*′ values computed at 5 cM intervals. The plotted line represents the fitted line.

Effect of the sample size. In order to assess the sample size effect on the extent of LD, we bootstrapped samples of different sizes of the same data. Figure 5 shows the mean of the means and variances of the 231 locus pairs obtained from bootstrapping samples of 26, 52, 104, 168, 500 and 1000 diplotypes from G2 over 1000 replicates. There was about a 2.1-fold increase in the value of *D*′ when the sample size decreased from 168 to 26. This increase was very similar to that found between G1 or G2 and G3. The estimate of the mean *D*′ was 2.4- to 2.6-fold larger in G3 than in G2 or G1. Figure 5 shows that the variance of *D*′ was about 16-fold larger for a sample size of 26 than for 168 diplotypes. Bootstrapping samples of up to 381 diplotypes from all the founder diplotypes showed that the mean and variance flatten if the sample size exceeded about 200 diplotypes (data not shown). Figure 6 shows how small sample sizes tend to flatten the decay of *D*′ with distance (i.e. it plateaus sooner). Each point is the average value of *D*′ obtained from bootstrapping 1000 samples of size 168 and 26 diplotypes from G2. The maximum and minimum fitted value for the results shown in Figure 6 were 0.44 and 0.192 for a sample size of 168 and 0.623 and 0.432 for a sample size of 26. This makes a difference between maximum and minimum value of 56% ([0.44 – 0.192]/0.44 = 0.56) and 30%, respectively.

Extent of linkage disequilibrium using unphased founders

Two-locus maximum likelihood estimates of haplotype frequencies for the same founder individuals used in the previous section were obtained using the expectation-maximization (EM) algorithm. *D*′ values were computed. Figure 7 shows, for each pair of loci, the estimate of *D*′ obtained from phased individuals (horizontal axis) and unphased individuals (vertical axis). Only nine out of the 231 pairs showed a smaller *D*′ obtained from unphased individuals than from phased individuals. The regression coefficient of *D*′ (unphased) on *D*′ (phased) was not significantly different from 1 and the intercept was significantly different from zero ($P < 10^{-12}$). The continuous line is the fitted line and the estimated parameters and their standard errors (in brackets) are shown in Figure 7.

Effect of the grouping strategy. The effect of different allele grouping strategies is shown in Figure 8. Although we show results only for unphased individuals the results obtained for phased individuals did not differ qualitatively from those shown here. Three different grouping strategies were considered: grouping of alleles with frequencies less than 7%, less than 1%, or not grouping at all. The grouping of rare alleles tended to reduce the value of *D*′, and is therefore a conservative approach. We also compared how the statistical significance of LD changed with grouping for a range of grouping strategies. Table 3 compares the number of times in which the locus pairs were classified as having the same (or different) level of significance (as defined for Fig. 3) when different levels of grouping were compared with no grouping. For example, if a pair of loci was in significant LD ($P = 0.05$) without grouping and in highly significant LD ($P = 0.0002$) when grouping alleles with frequencies less than 7%, then this pair added one value to one displacement in the 7% column. If it was the other way round, namely highly significant without grouping and not significant

Table 3. Number of times the classification of the statistical significance for the 231 locus pairs changed when grouping at different proportions compared with no grouping. The displacement is negative when grouping is less significant than not grouping

Displacement	0%						
	0.1%	1%	5%	7%	10%	25%	50%
−2	0	0	0	0	1	20	27
−1	3	1	13	19	36	70	77
0	227	201	169	170	162	129	124
1	1	29	48	42	32	12	3
2	0	0	1	0	0	0	0

when grouping alleles with frequencies less than 10%, then it added one count on to the displacement of −2 in the relevant column.

The test for LD tended to became more conservative when the threshold frequency for grouping increased (i.e. as we grouped higher proportions, the allelic association frequently became less significant), although there was some variability for the smallest threshold frequencies for grouping that could not be accounted for.

DISCUSSION

We have estimated that LD on chromosome 19 in the Talana population extends to between 8 and 11 cM. In addition, 71% of the locus pairs showed highly significant association when they were less than 20 cM apart, but only 6.5% when they were more than 20 cM apart. Isolated populations with high levels of LD generated by drift pose the problem of how to distinguish between LD due to close linkage and that generated by drift between unlinked loci (the background level of LD) that would lead to false positives. By chance, we would expect about 5% of the unlinked loci to show significant LD at the $P = 0.05$ level. There was only a slight increase (6.5%) on this, suggesting that the problem associated with high levels of background LD might be of little importance in this population. In order to assess this better, one would require microsatellite markers placed on other chromosomes. However, these data were not available and we considered that distances of more than 20 cM were an appropriate threshold (that is, we considered two loci 20 cM apart to be effectively unlinked). If a marker locus, more than 20 cM apart from a trait locus, showed a significant association with the trait, then this information would be of limited utility for mapping the position of the trait locus with any confidence.

We based our conclusions on the extent of LD both on the statistic D' and on the significance level of the allelic association. Each criterion has advantages and disadvantages. The negative aspects of using summary statistics such as D' are 3-fold: (i) small values cannot be interpreted as lack of significant association (6); (ii) they are difficult to interpret; and (iii) their sampling distribution is usually unknown, and changes with parameters such as effective population size, recombination fraction between the two loci, allele frequencies and sample size (14). The positive aspect is that their values are usually standardized, so that their range of possible values is the same regardless of the allelic frequencies, making comparisons easier across pairs, with different numbers and frequencies of alleles. The main advantage of using significance level is that it is easy to interpret. Its main disadvantage is that it depends on the marginals of the contingency table (number and frequency of the alleles and sample size). The fact that conclusions based on both methods are similar suggests that the present estimates are quite robust.

This study also suggests that the average level of LD in this population might be greater than in other isolates, as well as in other larger outbred populations. In a study of 5048 autosomal short tandem repeat polymorphisms (STRP) scatted across the genome, it was found that about 4% of the locus pairs separated by less than 4 cM were in LD for the European Utah and Amish CEPH families (15). Dunning *et al.* (16) studied the extent of LD in three different regions of the genome in four populations of European origin (Afrikaners, Ashkenazim, Finns and East Anglian British) and found that between 75% (Ashkenazim) and 94% (Finns) of the locus pairs showed significant LD for distances <5 kb (~ 0.005 cM). Zavattari *et al.* (3) studied the extent of LD in a sub-isolate of Sardinia (the village of Gavoi) in the same region on the long arm of chromosome X studied by Laan and Paabo (17) in Finns, Estonians, Swedes and Saami. They found similar levels of LD in Gavoi and the Saami; respectively 19/21 and 17/21 of the pairs were in significant LD within a region of ~ 10 cM (9–11.5 Mb). In the present study, we found similar levels of LD to those of the Saami or Gavoi, that is, 25/30 pairs in significant LD ($P < 0.0002$) for distances ≤ 10 cM. In a previous study of Talana spanning a region of 11.1 cM in Xq13, it was found that 6/15 markers pairs were in significant LD (4). The larger proportion of loci pairs in LD in our study compared with that on the same population are probably due to our larger sample size. Nevertheless, one must exercise care when comparing measures of LD across studies with, for example, different sample sizes, genome regions, marker informativeness and density or haplotyping methods. These results should be interpreted as a rough estimate of what one might expect if all these factors were the same.

We compared the effect on LD estimation of two different strategies for inferring population haplotype frequencies by estimating them with and without family information. We compared D' when estimated from phased and unphased individuals and found that estimates from phased individuals yielded lower estimates than from unphased individuals. This could be due to the fact that the program *Simwalk2*, used to obtain haplotypes using family information, assumes that the loci are in linkage equilibrium (LE) and the EM algorithm, used to estimate population haplotype frequencies, does not.

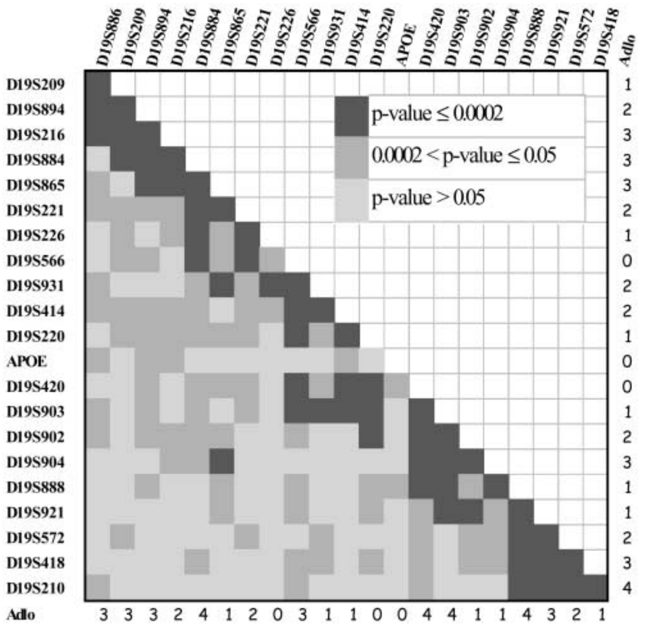

Figure 3. Linkage disequilibrium statistical significance between pairs of loci and number of adjacent loci in highly LD (Adlo).

Moreover, the EM algorithm is expected to work better when the amount of LD increases (18), whereas those programs that assume LE are expected to perform worse as the amount of LD increases (19). In the present case, there was a large difference between the average extent of LD (measured either as D' or as the significance level of the allelic association) when using phased and unphased individuals. We repeated the analysis to find the extent of LD as shown in Figures 2 and 3 but using haplotype frequencies obtained from unphased individuals (data not shown). We found that the extent of LD was between 4 cM (using the half-length of D') and 6.6 cM (1.12×5.92, where 1.12 is the mean number of adjacent markers in highly significant LD estimated from haplotype frequencies obtained without using family information and 5.92 is the average distance between markers; the estimated standard deviation was 4.59 cM). Although values of D' tend to be larger when using unphased individuals, their decay is faster and therefore the estimate of the useful extent of disequilibrium shorter. The differences obtained in the estimate of the extent of LD based on the significance level of the allelic association, when using phased and unphased individuals (that is, 11 versus 6.6 cM), could be due to the methods employed to test it. When using phased individuals one is assuming that one can count the haplotypes and apply a standard chi-squared test. However, what is counted is only an estimate of the haplotypes, not the haplotypes themselves. In other words, one is assuming that the haplotypes are known whereas they are only estimated with some degree of confidence that is not incorporated into the testing procedure. On the other hand, when using unphased individuals one compares the likelihood of the data under the assumption of LE and LD.

There have been suggestions (6,15,20) that the locus heterozygosity might affect the ability to detect LD. We therefore regressed the mean heterozygosity of the locus pairs

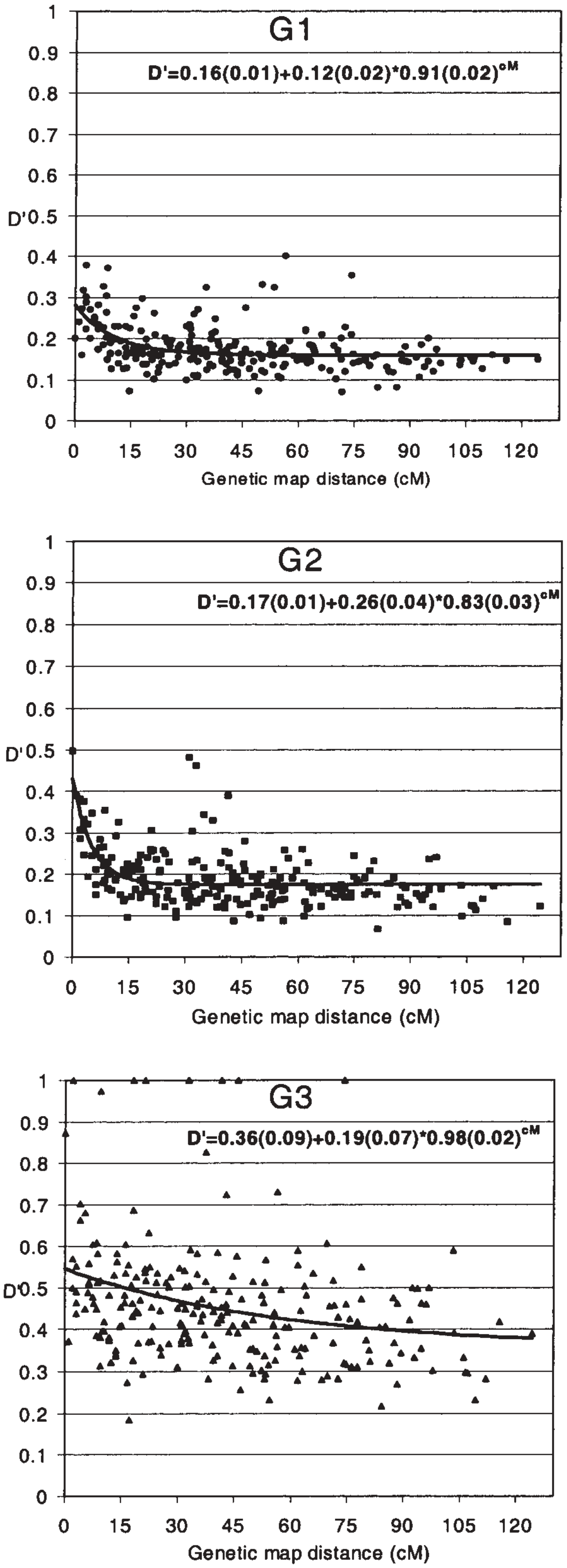

Figure 4. Relationship between genetic distance (cM) and level of linkage disequilibrium (D'). The plotted line represents the fitted line ($y = a + be^{-cx}$). The total number of diplotypes was 187, 168 and 26 for classes of individuals with children only (G1), grandchildren (G2) and great-grandchildren (G3) in the pedigree, respectively. The equation of the fitted line and the standard errors (in brackets) of the estimated parameters are shown.

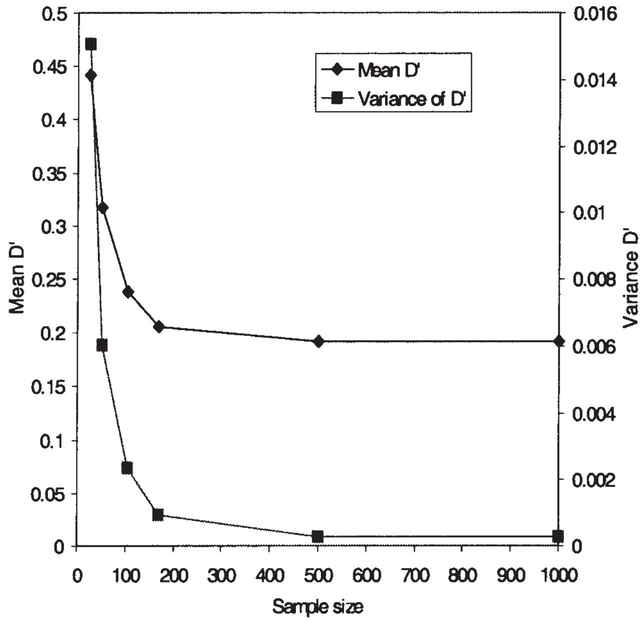

Figure 5. Effect of sample size on the mean of the means and variances of D' for the 231 locus pairs of founders G2. Points represent sample sizes of 26, 54, 104, 168, 500 and 1000 diplotypes.

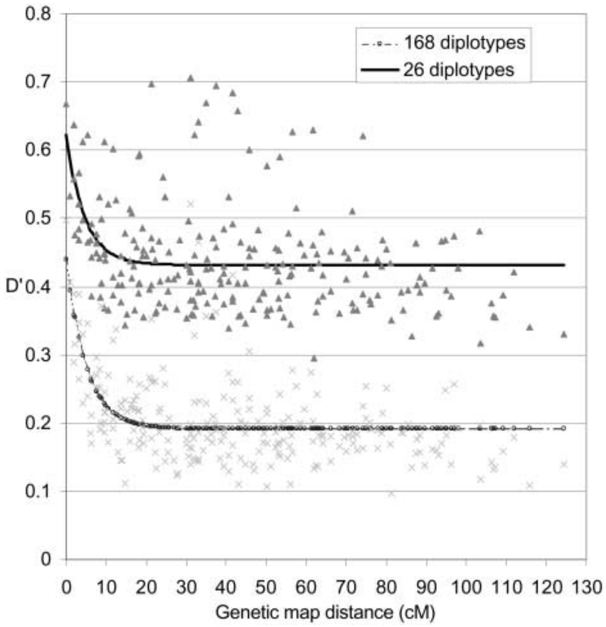

Figure 6. Effect of sample size on the decay of LD. Each dot is the average value obtained from bootstrapping 1000 samples of 168 (crosses) and 26 (triangles) diplotypes.

on the significance level of LD and found it significantly different ($P < 0.001$) from zero. Mean heterozyosity only accounted for about 6% of the total variance. We therefore considered its effect on the significance level to be negligible and so did not correct for it.

In our data set, some loci departed from HWE expectations. HWE is an assumption of the EM algorithm, and departures from HWE might lead to bias in the estimates of haplotype frequencies. It is, nevertheless, unlikely that the departures from HWE expectations we observed (that is, an excess of homozygotes) could produce an important degree of bias. When there is an excess of homozygotes, the number of doubly heterozygous individuals to be resolved is smaller and therefore little or no bias in the estimate of haplotype frequencies is expected (18,21).

Grouping of rare alleles is usually done for statistical reasons and alleles are grouped only with regard to their frequency. A more desirable approach would be grouping microsatellites alleles for biological reasons. For example, it might be more appropriate to group the lower frequency alleles of a microsatellite locus with those alleles that are most similar in length because, in a step-wise mutation model, alleles are assumed to mutate by increasing or decreasing the length of the repeated motif in single steps.

We wanted to determine if there were differences in the estimates of LD depending on the depth of the pedigree available for estimating founder diplotypes. For this, we obtained estimates of D' from the G1, G2 and G3 founders and found that they showed significant differences. However, these differences could be explained by the difference in sample size, rather than by the difference in the number of generations available to estimate diplotypes.

Our results about the sample size requirements for LD estimation are not limited to isolated human populations. Weir and Hill (22) derived the variance of R, the correlation of gene frequencies, for biallelic loci. Their arguments about the two sampling processes involved in estimating LD are also relevant for different measures of disequilibrium, such as D'. Under the null hypothesis of no LD and for closely linked loci the variance of R [$\mathrm{var}(R) = E(R^2)$] is approximately $1/(1 + 4N_ec) + 1/n$ where N_e is the effective population size, c is the recombination fraction between the two loci and n is the sample size. Hence, the variance of R is due to two different sampling processes, one that reflects the finite size of the population [$1/(1 + 4N_ec)$] and another that reflects the fact that a finite sample of the population [$1/n$] was drawn to estimate allele frequencies and disequilibrium. Note that n is either a sample of n identified chromosomes or n unphased individuals from which disequilibrium and allele frequencies were estimated. Our investigations using bootstrapping refer only to the second part of this formula ($1/n$).

In conclusion, Talana has levels of LD similar to those of the Saami and other Sardinian sub-isolates such as Gavoi. The estimated extent of LD and its variance is highly dependent on the sample size from which it has been estimated. Small samples tend to overestimate the amount of disequilibrium, in our case by a factor of up to two. Caution should therefore be exercised when planning LD mapping studies based on the amount of LD found from a preliminary study with a small sample of individuals. In such preliminary studies aimed at assessing the extent of LD, we recommend the use of about 200 unrelated individuals, and the use of both pairwise significance levels for allelic association and D' to interpret their results. When comparing levels of LD obtained from studies with different sample sizes we recommend accounting for differences in sample size using bootstrapping, as shown in Figure 1.

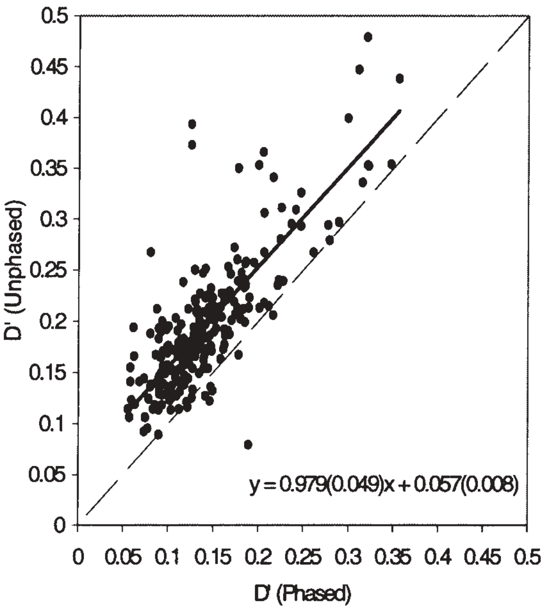

Figure 7. Comparison of D' values obtained from phased and unphased individuals. The fitted line (continuos line) and its equation are shown. In the equation, standard errors (in brackets) follow the parameter estimates.

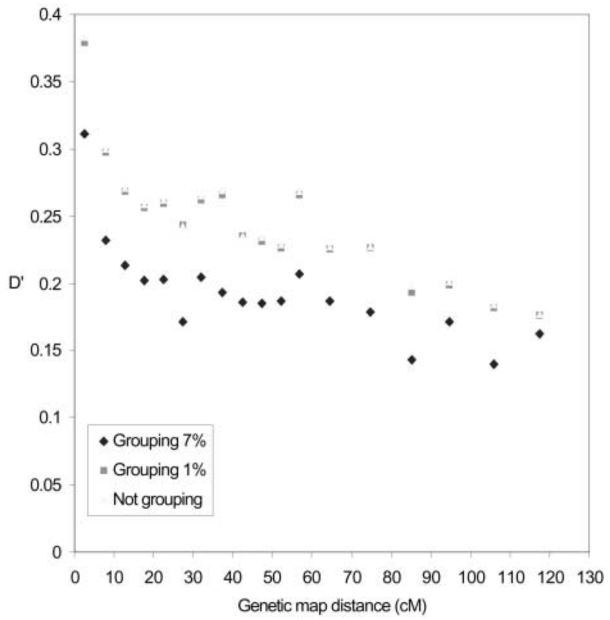

Figure 8. Effect of different grouping strategies on the estimate of D'. Each point represents the average of D' values computed at 5 cM intervals.

MATERIALS AND METHODS

Genetic linkage map

The Applied Biosystems HD5 set of microsatellite markers for chromosome 19 (21 markers) was used to genotype all 775 individuals from the Talana village, together with primers amplifying the three-allele APOE locus. The genetic linkage map for chromosome 19 was constructed using *Cri-Map* (23) (http://linkage.rockefeller.edu/multimap/crimap/) using Haldane's map function. Genotypes were available for 21 microsatellite markers and the APOE gene. The constructed map, which was based on an average of 338 meioses, agreed with the published map (ftp://ftp.genethon.fr/pub/Gmap/Nature-1995/) in the order of all the marker loci. We considered our estimated genetic linkage map more appropriate to this study because it was estimated from the population studied and because it is likely to be more accurate than the published map (24), which is only based on 186 meioses.

Hardy–Weinberg equilibrium proportions

Departures from Hardy–Weinberg equilibrium proportions were tested using Arlequin (25) (http://lgb.unige.ch/arlequin/). The individual test significance level after correction to give a total significance level (γ) of 0.05 was $P = 1 - (1 - \gamma)^{1/n} = 1 - (1 - 0.05)^{1/22} = 0.002$ where n is the total number of tests performed.

The mean number of alleles and observed heterozygosity at the microsatellite loci used in this study for the CEPH families were obtained from the Genethon web page (ftp://ftp.genethon.fr/pub/Gmap/Nature-1995/). These data are based on eight families (134 individuals) and 186 meioses.

Haplotype estimation using family information

All individuals' diplotypes (that is, the pair of haplotypes that compose the genotype) for the 22 available loci were estimated using *Simwalk2* (http://watson.hgen.pitt.edu/docs/simwalk2.html) (26) and founder haplotypes were selected for further analysis. Founder haplotypes were counted and population two-locus haplotype frequencies obtained. Using diplotypes obtained from two different runs of *Simwalk2* with different random seeds gave virtually the same results (only results for one of them are shown).

We could infer a maximum of 381 founder 22-locus diplotypes that were assigned to different categories depending on the number of generations of descent in the pedigree. Those founders that had great-grandchildren in the pedigree were assigned to the category G3, founders with grandchildren only to category G2 and founders with children only to G1. G1, G2 and G3 had a total of 187, 168 and 26 diplotypes, respectively. Unless otherwise stated, the results shown were obtained when using all founders available.

Haplotype frequency estimation without using family information

Maximum likelihood estimates of all 231 [$22 \times (22 - 1)/2$] two-marker locus haplotype frequencies were estimated by employing the EM algorithm (27). No attempt to estimate more than two-locus haplotypes using the EM algorithm was made.

Measuring the amount of linkage disequilibrium

We measured LD using the statistic D' defined for multiallelic loci (10,28). D' was based on estimates of two-locus haplotype frequencies obtained both by counting the 381 22-locus founder diplotypes when using family information (phased founders), and from the two-locus population haplotype frequencies obtained without using family information (unphased founders).

Test for association when using phased individuals

We used *Gold* (29) to test the statistical significance of the allelic association between all pairs of loci when using inferred haplotypes. Association was tested by means of a standard chi-square test (based on the observed and expected haplotype frequencies) with $(k-1)\times(l-1)$ degrees of freedom where k and l are the number of alleles at the two loci. *Gold* groups low-frequency alleles in order to avoid spurious results due to small sample sizes and sparse contingency tables. We used two different grouping strategies, grouping at 1 and 7%. Unless otherwise stated the results presented are those corresponding to the 7% grouping. Not all loci from the 381 22-locus diplotypes were scored, therefore the number of two-locus haplotypes on which D' and the significance of the allelic association were based varied from 312 to 606 haplotypes.

Test for association when using unphased individuals

The statistical significance of allelic association was tested comparing the likelihood ratio statistic $S = 2\ln(L_{\mathrm{LD}}/L_{\mathrm{LE}})$ with a χ^2 distribution with $(k-1)\times(l-1)$ degrees of freedom (30). Assuming random mating, L_{LD} is the likelihood computed using the haplotype frequencies found by the EM algorithm and L_{LE} is the likelihood under the assumption of linkage equilibrium. We used *Gold* at different grouping frequencies of the alleles (1 and 7%) and our own implementation of the EM algorithm, which does not do any grouping. Comparing the statistic S with a χ^2 distribution is, strictly speaking, only valid under asymptotic assumptions, which are likely to hold here due to the large sample size of the data set.

ACKNOWLEDGEMENTS

We thank Susan Campbell and Isla Campbell for the genotyping. P.M.V. and S.A.K. acknowledge support from the Biotechnology and Biological Sciences Research Council and the Royal Society, respectively. We thank two anonymous referees for helpful comments on the original manuscript.

REFERENCES

1. Hill, W.G. and Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res.*, **8**, 269–294.
2. Eaves, I.A., Merriman, T.R., Barber, R.A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F. and Todd, J.A. (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.
3. Zavattari, P., Deidda, E., Whalen, M., Lampis, R., Mulargia, A., Loddo, M., Eaves, I., Mastio, G., Todd, J.A. and Cucca, F. (2000) Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography chromosome recombination frequency and selection. *Hum. Mol. Gen.*, **9**, 2947–2957.
4. Angius, A., Bebbere, D., Petretto, E., Falchi, M., Forabosco, P., Maestrale, B., Casu, G., Persico, I., Melis, P.M. and Pirastu, M. (2002) Not all isolates are equal: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian sub-populations. *Hum. Genet.*, **111**, 9–15.
5. Angius, A., Melis, P.M., Morelli, L., Petretto, E., Casu, G., Maestrale, G.B., Fraumene, C., Bebbere, D., Forabosco, P. and Pirastu, M. (2001) Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits *Hum. Genet.*, **109**, 198–209.
6. Slatkin, M. (1994) Linkage disequilibrium in growing and stable populations. *Genetics*, **137**, 331–336.
7. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
8. Wright, A.F., Carothers, A.D. and Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.*, **23**, 397–404.
9. Peltonen, L., Jalanko, A. and Varilo, T. (1999) Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.*, **8**, 1913–1923.
10. Hedrick, P.W. (1987) Gametic disequilibrium measures—proceed with caution. *Genetics*, **117**, 331–341.
11. Genstat 5 Committee (1993) Genstat 5 Reference Manual. Clarendon Press, Oxford.
12. Collins, A. and Morton, N.E. (1998) Mapping a disease locus by allelic association. *Proc. Natl Acad. Sci. USA*, **95**, 1741–1745.
13. Collins, A., Lonjou, C. and Morton, N.E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA*, **96**, 15173–15177.
14. Hudson, R.R. (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, **109**, 611–631.
15. Huttley, G.A., Smith, M.W., Carrington, M. and O'Brien, S.J. (1999) A scan for linkage disequilibrium across the human genome. *Genetics*, **152**, 1711–1722.
16. Dunning, A.M., Durocher, F., Healey, C.S., Teare, M.D., McBride, S.E., Carlomagno, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S. *et al.* (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.*, **67**, 1544–1554.
17. Laan, M. and Paabo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat. Genet.*, **17**, 435–438.
18. Fallin, D. and Schork, N.J. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.*, **67**, 947–959.
19. Schaid, D.J., Mcdonnell, S.K., Wang, L., Cunningham, J.M. and Thibodeau, S.N. (2002) Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am. J. Hum. Genet.*, **71**, 992–995.
20. Varilo, T., Paunio, T., Parker, A., Perola, M., Meyer, J., Terwilliger, J.D. and Peltonen, L. (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Hum. Mol. Genet.*, **12**, 51–59.
21. Osier, M., Pakstis, A.J., Kidd, J.R., Lee, J.F., Yin, S.J., Ko, H.C., Edenberg, H.J., Lu, R.B. and Kidd, K.K. (1999) Linkage disequilibrium at the Adh2 and Adh3 loci and risk of alcoholism. *Am. J. Hum. Genet.*, **64**, 1147–1157.
22. Weir, B.S. and Hill, W.G. (1980) Effect of mating structure on variation in linkage disequilibrium. *Genetics*, **95**, 477–488.
23. Green, P., Falls, K. and Crooks, S. (1990) *Cri-Map Version 2.4.* Washington University School of Medicine, St Louis, MO.
24. Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.
25. Schneider, S., Roessli, D. and Excoffier, L. (2000) *Arlequin: a software for Population Genetics Data Analysis. Version 2.000*. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva.
26. Sobel, E. and Lange, K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.*, **58**, 1323–1337.
27. Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood-estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
28. Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**, 49–67.
29. Abecasis, G.R. and Cookson, W.O.C. (2000) Gold—graphical overview of linkage disequilibrium. *Bioinformatics*, **16**, 182–183.
30. Slatkin, M. and Excoffier, L. (1996) Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. *Heredity*, **76**, 377–383.