DOI: 10.1093/hmg/ddg005

The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories

Teppo Varilo[1,*], Tiina Paunio[1,2], Alex Parker[3], Markus Perola[1,4], Joanne Meyer[3], Joseph D. Terwilliger[5] and Leena Peltonen[1,4,6]

[1]Department of Molecular Medicine, National Public Health Institute, Biomedicum, Helsinki, Finland, [2]Department of Psychiatry, University of Helsinki, Helsinki, Finland, [3]Millennium Pharmaceuticals Inc., Cambridge, Massachusetts, USA, [4]Department of Human Genetics, UCLA, Los Angeles, California, USA, [5]Department of Psychiatry and Columbia Genome Center, Columbia University, New York, USA and [6]Department of Medical Genetics, University of Helsinki, Biomedicum, Helsinki, Finland

Received August 21, 2002; Revised and Accepted November 11, 2002

Linkage disequilibrium (LD) has been an efficient tool for fine mapping of monogenic disease genes in population isolates. Its usefulness for identification of predisposing loci for common, polygenic diseases has been challenged on the basis of anticipated allelic and locus heterogeneity. We compared the extent of LD among marker loci in Finnish subpopulations with divergent but well-characterized histories. One study sample represents the early settlement Finnish population, descended from two immigration events 4000 and 2000 years ago. The second sample represents the geographically large late settlement region, populated 15 generations ago by several small immigrant groups from the early settlement region. The third is a restricted regional subpopulation in northeastern Finland which was founded 12 generations ago by 39 immigrant families from the late settlement region. We genotyped 243 microsatellite markers and 68 single nucleotide polymorphisms (SNPs) on chromosomes 1q and 5q. The genealogy of the families from the early ($n = 16$) and late settlements ($n = 54$) and the isolated settlement ($n = 54$) was studied in detail back to the 1800s. Microsatellite data revealed greater LD in the young, founder subpopulation than was seen in either of the older populations. Observed linkage disequilibrium correlated not only with physical distance between markers but also with the information content of the markers. Using biallelic SNP markers, significant LD could only be detected up to 0.1 cM. Our results demonstrate the complexity of the concept of 'detectable LD' and emphasize the importance of understanding population history when designing a strategy for disease gene mapping.

INTRODUCTION

Genetic drift, founder effects and population bottlenecks can modify gene pools significantly, and may render some populations better suited than others for linkage disequilibrium (LD)-based gene mapping strategies. The Finnish population has been frequently cited as an example of a gene pool which demonstrates the effect of multiple bottlenecks in its population history, and LD has been used successfully in mapping of genes responsible for numerous rare diseases whose frequency is enriched in this population (1).

Finland has been inhabited since it was liberated from the Pleistocene ice-sheet around 10 000 years ago. The first major expansion of the Finnish population began 4000 years ago, when agriculture and animal husbandry were introduced by immigration of eastern Uralic speakers. A second wave of immigration, entering from the south some 2000 years ago, introduced new implements that facilitated permanent settlement in villages (1–3). For centuries, only a narrow strip of land in the coastal areas of the south and southwest was populated, and even as late as the twelfth century the population of Finland was only around 50 000. The inhabitation of the wilderness, referred to as the late settlement, began in the 1500s, originating from the southeastern region of South Savo (Fig. 1). By this time the coastal population had reached 250 000 in number, resulting in pressure to cultivate more land, leading King Gustavus of Vasa to favor inhabitation

*To whom correspondence should be addressed at: Department of Molecular Medicine, National Public Health Institute, Biomedicum, PO Box 104, 00251 Helsinki, Finland. Tel: +358 947447224; Fax: +358 947448480; Email: teppo.varilo@ktl.fi

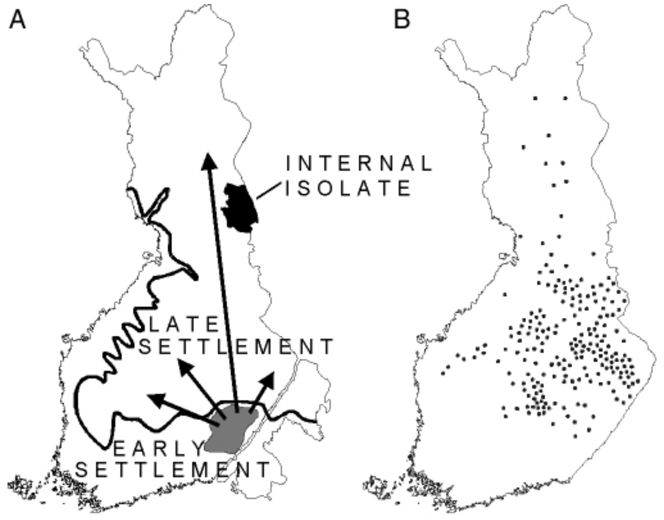

Figure 1. (**A**) The early and late settlement region and Kuusamo subregion (shaded black) on the map of Finland. The inhabitation of the wilderness began in the 1500s in a small southeastern area of South Savo (shaded gray) to the central, western and finally northern parts of the country. Kuusamo was inhabited towards the end of this internal migration movement mainly by families from Ostrobothnia and from South Kainuu; both groups have their ancestral roots in South Savo. (**B**) The birthplaces of the grandparents of the late settlement (LS) family sample used in this study.

of the wilderness during his regime (1523–1560). In the sixteenth century the inhabited land area of Finland doubled, and during the most recent 10–12 generations the Finnish population has expanded rapidly from 250 000 to its present size of 5.1 million. Subsequent to the sixteenth century, migration within Finland has been characterized by the founding of small subpopulations across the sparsely populated parts of the country; the rural isolates so created have remained surprisingly stable over time.

The consequences of multiple founder effects and subsequent isolation of this population are apparent in reduced allelic diversity, and reflected by overrepresentation of 35 rare, mostly autosomal-recessive Mendelian disorders in Finland (1,4). In each of these diseases, one major founder mutation has been identified in Finland, whereas the genetic background of the same diseases elsewhere is more diverse (1). The effect of multiple bottlenecks is demonstrable today by marked frequency differences in these disease mutations among regional subisolates (5). Historical routes of many internal migrations can actually be reconstructed using the distribution of these disease alleles (6).

An example of the regional subpopulation, molded by multiple bottlenecks and isolated by distance, is the north-eastern municipality of Kuusamo. The Finnish inhabitation of Kuusamo is well documented as beginning with the first immigrant in 1676, owing to legal records documenting inevitable conflicts with the native forager population, the Saami. The great famine of 1695–1697 killed about half of the Kuusamo immigrants and resulted in the gradual disappearance of the Saami people from the district. When the parish registers were established in 1718, the population in the 165 houses consisted of 615 individuals belonging to 39 families. During the subsequent century, when diseases repeatedly swept through more densely populated parts of Europe, population growth was rapid in rural parts of Finland; this expansion eventually led Kuusamo to its present population of 18 000.

We sought to address the extent of intermarker LD on chromosomes randomly ascertained from various subpopulations within Finland. Our primary goal was to broadly characterize the effect of different population histories on the distribution of LD. A secondary goal was to compare the extent of LD detected by microsatellite markers with that detected by single-nucleotide polymorphisms (SNPs). We studied intermarker allelic association across two chromosomal regions in chromosomes from the Kuusamo subpopulation, compared with chromosomes from the late and early settlement areas. Our results indicate that LD is detectable over greater genetic distances in the Kuusamo regional subisolate than in the source population, and that the intervals across which LD was detected using microsatellite markers were significantly wider than those detected using SNPs.

RESULTS

Skeleton map and LD among microsatellites on 1q and 5q

Initially 22 microsatellite markers across a 27 cM region of 1q, and 31 markers across a 14 cM region of 5q were tested for intermarker LD. The average intermarker interval was 1.23 cM on 1q and 0.45 cM on 5q. Both the regional isolate of Kuusamo (IS) and late settlement (LS) revealed significant evidence for LD compared to early settlement (ES), with a clear distance-correlated pattern for both chromosomal regions. Although no difference in mean marker heterozygosity was seen between different population samples, significantly more intermarker LD was observed in IS versus the more heterogeneous LS and ES populations (Fig. 2).

Fine map of microsatellite markers and SNPs on 1q

An overview of different chromosomal regions studied on 1q is given in Figure 3 (see also Materials and Methods). To further address the impact of marker heterozygosity on detection of LD, we evaluated LD between SNP markers in our three study samples. Heterozygosity values for the SNPs varied from 0 to 0.5, averaging 0.34 in IS, 0.35 in LS, and 0.35 in ES. On 1q, two small regions 17 cM apart from each other were genotyped with 45 and 23 SNPs, providing on average 0.010 and 0.016 cM intermarker distances, respectively. The results are given in Figure 4. The region containing 45 SNPs also included 16 microsatellite markers distributed over 0.3 cM. Across this entire interval significant LD was observed among 88% (106/120) of microsatellite marker pairs in the IS sample, versus 56% (67/120) in LS, and 21% (25/120) in ES.

Next we proceeded to analyze the study samples using a larger number of microsatellite markers ($n = 212$) covering an extended region of chromosome 1q with an average ‘fine mapping’ intermarker interval of 0.29 cM. We first compared the distribution of marker polymorphism in the study samples. In IS, heterozygosity varied from 0.09 to 0.91; average heterozygosity was 0.63 in IS, 0.64 in LS and 0.63 in ES, suggesting that this characteristic had no impact on the population differences

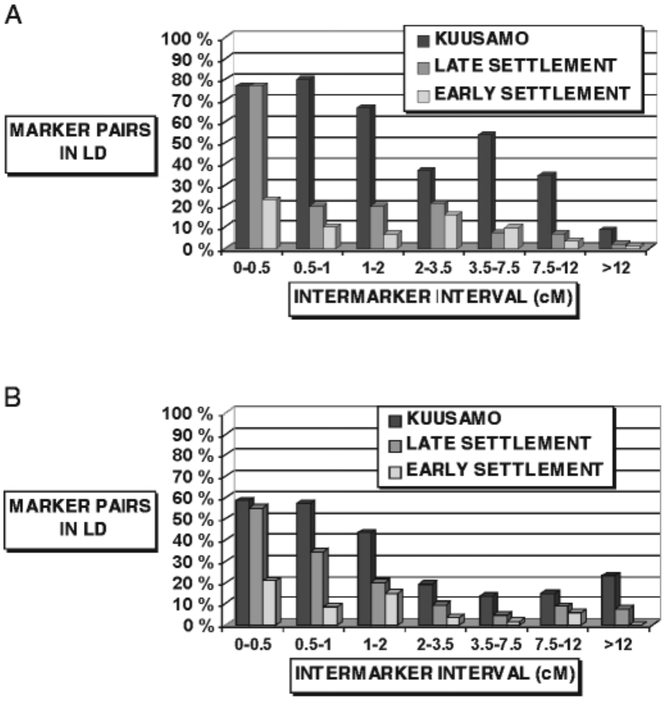

Figure 2. The percentage of microsatellite pairs which rejected the null hypothesis of linkage equilibrium ($P < 0.05$) over different intervals. LD observed in Kuusamo, LS and ES populations (**A**) between 231 marker pairs on 1q, and (**B**) between 465 marker pairs on 5q. The skeleton map of 22 markers for 1q and 31 markers for 5q was used here.

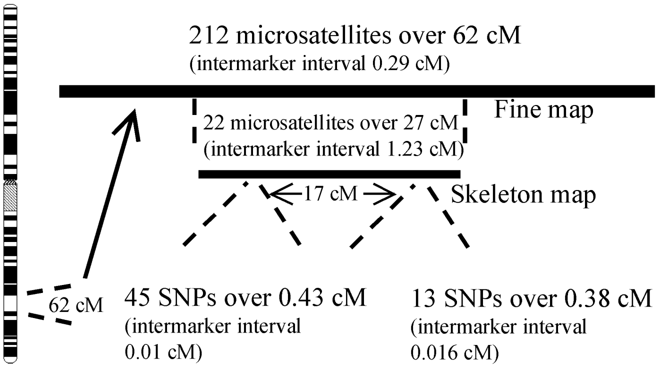
212 microsatellites over 62 cM
(intermarker interval 0.29 cM)
Fine map
22 microsatellites over 27 cM
(intermarker interval 1.23 cM)
17 cM
Skeleton map
62 cM
45 SNPs over 0.43 cM
(intermarker interval 0.01 cM)
13 SNPs over 0.38 cM
(intermarker interval 0.016 cM)

Figure 3. An overview of different chromosomal regions studied on 1q.

discovered. The average number of alleles per marker was 6.8. Distinct differences in allele frequency distribution were observed between Kuusamo and the late settlement population, with 48% of the markers (101/212) differing significantly ($P < 0.01$) between the two. A total of 87 alleles at 67 markers were observed only in IS, while 167 alleles at 104 markers were observed only in the late settlement sample.

Using this dense map of microsatellite markers, we performed analyses of LD within several intervals defined by increasing genetic distance (Fig. 5). In IS, the null hypothesis of linkage equilibrium was rejected ($P < 0.05$) for 76% of microsatellite pairs separated by <0.5 cM, 30% of pairs spaced $0.5 < 7.5$ cM, 12% of pairs spaced $7.5 < 12$ cM, and 7% at greater than 12 cM intervals. For LS, the corresponding figures were 52, 8, 4 and 5%, and for ES they were 19, 5, 4 and 5%. The expectation is that 5% of pairs would show LD under the null hypothesis; we can therefore conclude that at distances >0.5 cM there was essentially no evidence of LD in the ES sample. The absolute number of marker pairs observed to be in LD were, for IS, 587/770 at <0.5 cM, 2431/7983 at $0.5 < 7.5$ cM, 442/3674 at $7.5 < 12$ cM, and 724/9939 at >12 cM intervals; the corresponding figures in LS were 399/770, 651/7983, 143/3674 and 467/9939, and in ES 144/770, 423/7983, 141/3674 and 467/9939. These main results are presented in Figure 5, utilizing denser intermarker interval subdivision to detail the decline of LD with distance in different populations; a summary of results is given in Table 1. A likelihood ratio test rejected the hypothesis of equivalence of these distributions across the three study samples ($P \ll 10^{-10}$).

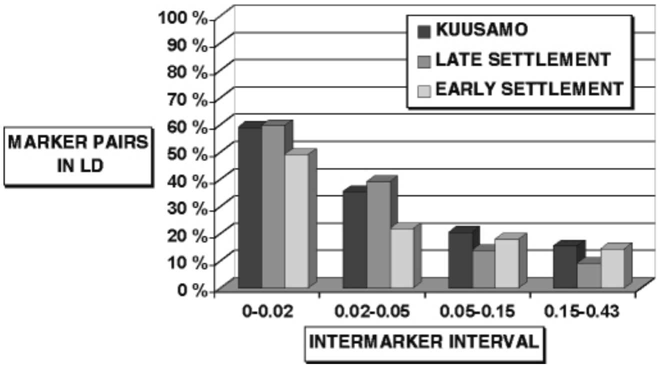

Figure 4. Proportion of 1200 SNP pairs showing LD ($P < 0.05$) at different intervals in 1q in Kuusamo, LS and ES samples. The absolute numbers of marker pairs were, for IS 150/253, 59/166, 68/333 and 69/447, for LS 159/268, 69/176, 48/345 and 41/454, and for ES 122/248, 37/168, 59/328 and 60/412.

Linkage equilibrium was strongly rejected ($P < 0.001$) for 25% of microsatellite marker pairs in IS, 15% in LS, and 6% in ES, again with a distinct distance-correlated pattern. Regression analysis was performed comparing LD [ln (significance level)] to intermarker distance [$\ln(1 - \theta)$]; this analysis demonstrated strong negative correlation between LD and genetic distance, with $r^2 = 0.2844$ ($P \ll 10^{-10}$) in IS and $r^2 = 0.1455$ ($P \ll 10^{-10}$) in LS. These r^2 values increased to 0.311 and 0.150, respectively ($P \ll 10^{-10}$), when marker heterozygosity was included as a covariate in the analysis. Information about the genetic map and polymorphic loci genotyped (Table A) as well as the raw genotype data (Table B) can be found on our website. The difference in the extent of LD interval in the Kuusamo and late settlement chromosomes is presented graphically in Figure 6.

Comparison of microsatellite and SNP markers

We then sought to determine the relative power of microsatellites and SNP markers to detect LD. We looked at the distribution of P-values from LD tests for marker pairs comprising of one microsatellite and one SNP, as well as tests of LD between SNP haplotypes and nearby SNPs. The more informative microsatellites are expected to provide a more powerful test of LD when the degree of LD is equal, but it has been hypothesized that SNPs may generally retain the LD

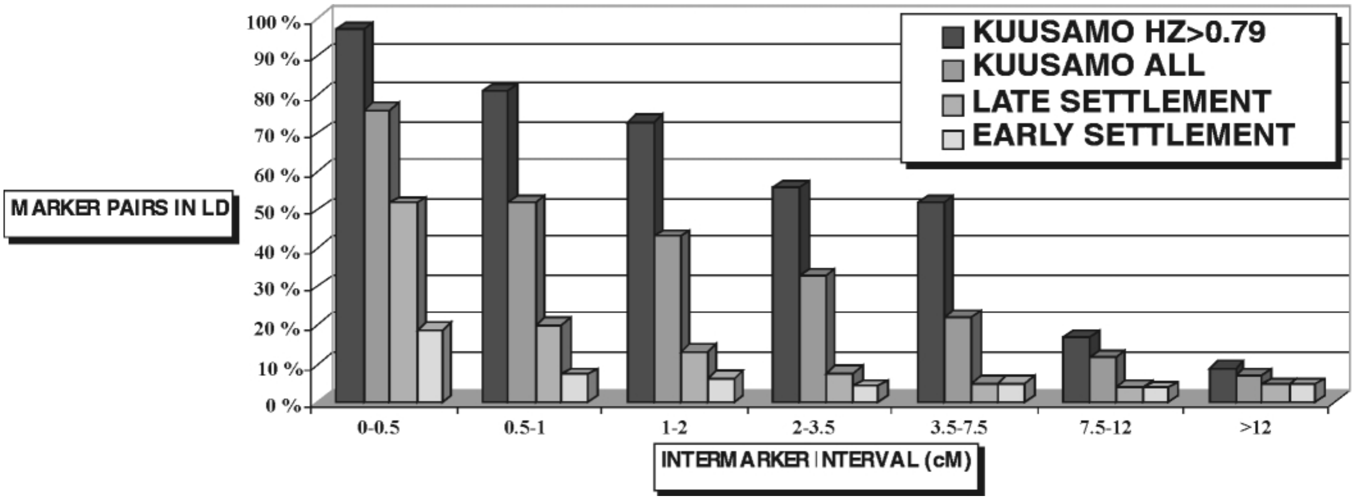

Figure 5. Proportion of microsatellite markers showing LD at different intervals of 1q in Kuusamo, LS and ES. The fine mapping data set of 212 markers and a total of 22 366 marker pairs were analyzed. Data is also provided for the 1353 most informative marker pairs in the Kuusamo population.

Table 1. Summary of results: LD observed between microsatellite marker pairs at different intervals in 1q and 5q in Kuusamo (IS), late (LS) and early settlement (ES) populations

Intermarker distance (cM)	Skeleton map study			Fine map study			
	IS	LS	ES	IS (HZ < 0.79)	IS	LS	ES
Chromosome 1							
0–0.5	77%	77%	23%	97%	76%	52%	19%
0.5–1	80%	20%	10%	81%	52%	20%	7%
1–2	67%	20%	7%	73%	43%	13%	6%
2–3.5	37%	21%	16%	56%	33%	7%	4%
3.5–7.5	54%	7%	10%	52%	22%	5%	5%
7.5–12	34%	7%	3%	17%	12%	4%	4%
>12	9%	2%	1%	9%	7%	5%	5%
Chromosome 5							
0–0.5	59%	55%	21%				
0.5–1	57%	34%	9%				
1–2	43%	20%	15%				
2–3.5	19%	10%	4%				
3.5–7.5	14%	5%	2%				
7.5–12	15%	9%	6%				
>12	23%	8%	0%				

signature of historical demographic events longer owing to their much slower average mutation rate.

To model the use of microsatellite markers in detection of disease-associated DNA variants, LD between microsatellite–SNP marker pairs was calculated. We used 212 microsatellites (average heterozygosity 0.63) and 18 SNPs (average heterozygosity 0.43). The results are given in Figure 7. The power of SNP haplotypes to detect LD with flanking DNA variants was then compared analogously (Fig. 8). We excluded the most uninformative SNPs (heterozygosity < 0.20) from the analysis, and were left with 55 SNPs in the 1q regions, which formed 18 successive three-point haplotypes (average heterozygosity 0.71), 13 successive four-point haplotypes (average heterozygosity 0.76) and 10 successive five-point haplotypes (average heterozygosity 0.82). We calculated LD between each haplotype and the remaining 50–52 SNPs (average heterozygosity 0.40) in IS. For three point haplotypes, 72% (92/128) of haplotypes showed significant LD ($P < 0.05$) with flanking SNPs closer than <0.035 cM, 29% (106/364) with SNPs separated by $0.035 < 0.38$ cM, and 5% (24/444) over 17.05–17.87 cM intervals (haplotype–SNP pairs spanning the two densely typed regions). For four-point haplotypes the corresponding figures were 81% (68/84), 28% (75/264) and 4% (14/315), and for five-point haplotypes, 78% (47/60), 28% (56/199) and 4% (10/241).

The power of SNPs for detection of LD was further examined by quantifying LD between SNP pairs. There is a well-known upward bias in pointwise estimates of D', owing to the bounded nature of the parameter (7). In fact, for unlinked marker pairs in our data set average point estimates of D' range from 0.09 to 0.12 depending on the population, with IS giving the highest background bias.

In order to estimate the relationship between D' and intermarker distance, a model for decay of D' as a function of intermarker distance was estimated by maximum likelihood. We used the raw data to compute the likelihood of each marker pair (assuming independence as a function of some decay curve for D' with intermarker distance). Since it is known that in general

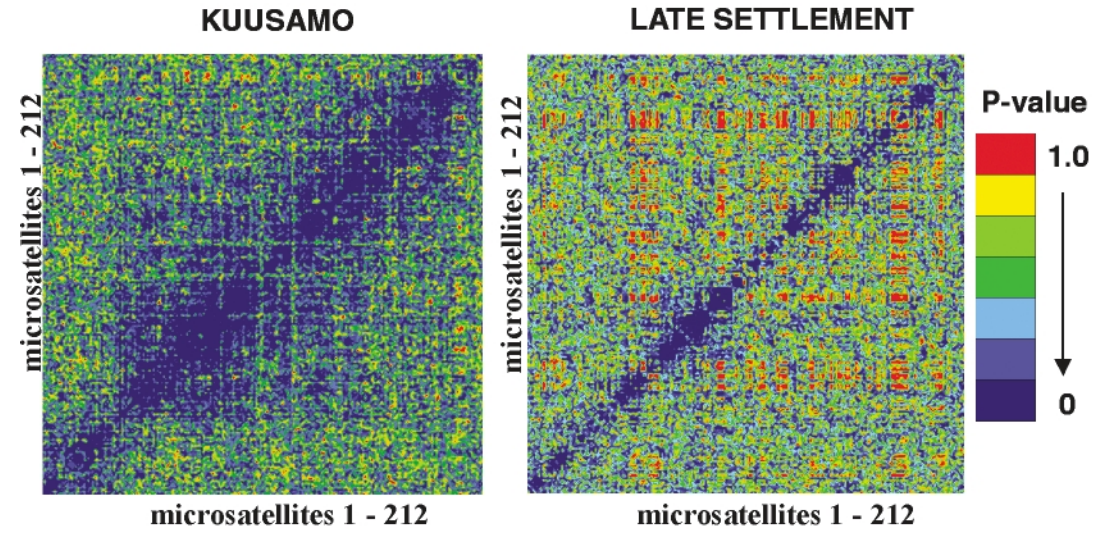

Figure 6. LD observed in Kuusamo (left) and late settlement (right). LD is plotted in 212 microsatellite markers against each other, forming 22 366 dots. Color reflects the significance of the *P*-value from deep blue ($P = 0.00$) to red ($P = 1.00$). The closest marker pairs are in the diagonal line. The figure was drawn using the GOLD 1.0 program.

D' decays as $D' = D'_0\,(1 - \theta)$, we estimated D'_0 and n using all the raw haplotype data jointly. In each case, the null hypothesis of no LD was rejected at $P \ll 10^{-100}$, as was the hypothesis of constant LD over all markers in favor of the exponential decay model.

The expected maximum likelihood estimate of D', $\hat{D}'$, is a joint function of the true value of D' and the allele frequencies of the two markers. When $D' = 1$, there is no bias, but when $D' = 0$ the bias can be enormous, especially when the SNP alleles are rare (Figures A–J on our website).

To correct for this bias in the above analysis, we first tried to estimate the likelihood of the observed data for each true value of D' via simulation of the density functions for $\hat{D}'$ in each marker pair, conditional on the marginal allele frequencies in the sample and on various observed values of D'. This procedure turned out to be extremely time-consuming, and, in the end, an approximate solution was employed; this approximation yielded almost identical results for a small subset of the data to which it was feasible to apply both methods. The procedure we used was to first take all marker loci separated by more than 14 cM (i.e. all pairs of SNPs with one from each of the two disjoint genomic regions typed), and estimate a constant value of $\hat{D}'_0$. Then, for the markers which are linked it was assumed that $D' = \hat{D}'_0 + D_0(1 - \theta)^n$. Since the bias is essentially 0 when $\theta = 0$, we estimated a bias-corrected $D'(0)$ as $D'_0 = D_0 + \hat{D}'_0$, while elsewhere the bias was intermediate between that at 0 and that for unlinked markers, such that $D'_{\text{corr}} = (\hat{D}'_0 + D_0)(1 - \theta)^n$. While this approach averages the bias over all marker heterozygosities, it nonetheless gave nearly identical numerical results for the estimated decay curve, as did the full simulation-based bias correction on the small subset of the data for which that was feasible (Fig. 9).

The effect of microsatellite informativeness on detection of LD

The heterozygosity of microsatellite markers in our study varied from 0.09 to 0.91. To determine the effect of marker informativeness on apparent LD we divided all marker pairs into two categories according to their mean heterozygosity in the Kuusamo sample. As expected, in the more polymorphic half (mean heterozygosity = 0.73) 89% of microsatellite pairs separated by <0.5 cM were in significant LD ($P < 0.05$), as were 41% of pairs spaced from 0.5 to 7.5 cM, 14% from 7.5 to 12 cM, and 8% at greater than 12 cM intervals. In contrast, for the less polymorphic half of the markers (mean heterozygosity = 0.53), the corresponding figures were 63, 21, 10, and 7% (Fig. 10). The same phenomenon is reflected in the smaller proportion of SNPs exhibiting significant LD with other SNPs, versus microsatellites showing LD with the same set of SNPs. In regression analysis, marker heterozygosity explained 18.4% of the variance of In(*P*-values) ($r^2 = 0.184$, $P \ll 10^{-10}$). In conclusion, the more informative the markers, the more LD was detectable.

DISCUSSION

We sought to investigate the effects of population history and marker informativeness on detection of intermarker LD. We compared the recently isolated subpopulation of Kuusamo, Finland, characterized by a restricted number of founders, to more diverse early and late settlement Finnish population samples. When the most informative microsatellite marker pairs on chromosome 1q were analyzed, the majority of marker pairs <7.5 cM apart were found to exhibit LD in the Kuusamo isolate. In analyses of all 22 366 microsatellite marker pairs, most marker pairs separated by less than 3 cM revealed LD in the Kuusamo isolate; detectable LD was observed over significantly smaller intervals on the chromosomes from the broader Finnish population. However, in the late settlement sample LD was evident between most marker pairs separated by <0.5 cM. Similarly, on chromosome 5q LD was detected between most microsatellite marker pairs <2 cM apart in the Kuusamo isolate. Our study samples from the late settlement area of Finland, and especially those from the regional

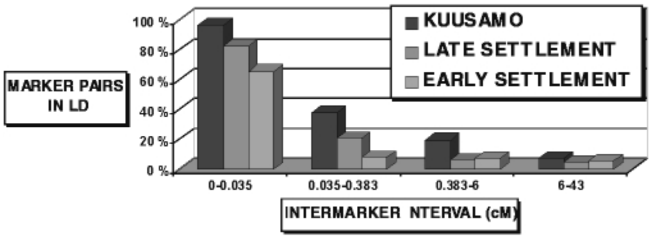

Figure 7. Proportion of 3816 microsatellite–SNP marker pairs showing LD ($P < 0.05$) at different intervals in 1q in Kuusamo, LS and ES samples. The absolute number of marker pairs were, for IS 28/29, 39/103, 155/805 and 209/2879, for LS 24/29, 21/103, 48/805 and 135/2879, and for ES 19/29, 8/103, 57/805 and 148/2879.

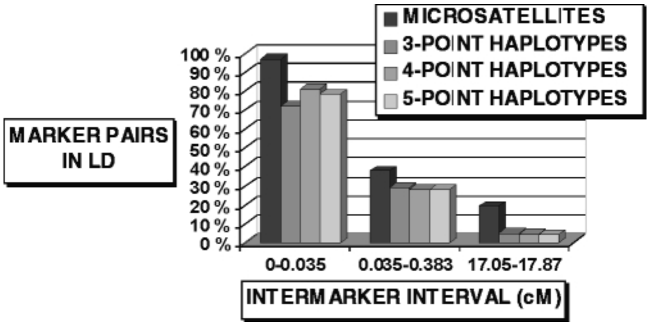

Figure 8. Proportion of marker pairs in LD at different intervals in 1q in Kuusamo: the data set contains 3816 microsatellite–SNP marker pairs, 936 three-point haplotype–SNP marker pairs, 663 four-point haplotype–SNP marker pairs and 500 five-point haplotype–SNP marker pairs.

Kuusamo subpopulation featuring a small number of founders, thus revealed significant LD over much wider chromosomal regions than was observed in the sample of early settlement chromosomes; this result demonstrates a clear genetic reflection of the known demographic history of these populations.

We have previously compared the extent of LD in the general Finnish population to that in Kuusamo using a sparse set of microsatellite markers located in five chromosomal regions (8). In another study of the Finnish population that used 43 microsatellites spanning 100 cM, all marker pairs separated by <1 cMs revealed LD, versus 78% spaced at 1–3 cM and 39% separated by 3–4 cM (9). Our results here using denser maps are consistent with and substantially extend these findings.

When we compared the genetic and physical maps of the analyzed 62 cM region of chromosome 1, they did not suggest the presence of anomalously high or low overall recombination rates (10); neither the Marshfield nor the more recent DeCODE map revealed any substantive discrepancy between genetic and physical maps of the analyzed segment (11,12).

Our results reveal that LD is affected by factors other than recombination rate and population history alone, as marker heterozygosity had a distinct impact on the magnitude of observed LD. Importantly, single informative microsatellites provided more power to detect LD than did SNPs, even when information from three to five SNPs was combined. In all three Finnish populations LD patterns detected using SNP markers were very similar, and even in the Kuusamo sample only 61% of SNP pairs separated by less than 0.02 cM were found to exhibit detectable LD, and the observed LD did not seem to increase when the intermarker interval diminished, with significant disequilibrium occurring at 64% of intervals under 0.01 cM, 66% under 0.005 cM, and 71% under 0.001 cM. We would also expect similar haplotype blocks identified with SNPs in all Finnish populations since they all originate from the same set of founder chromosomes of early settlement.

Small, constant size populations like the Saami, which have experienced repeated population bottlenecks but have never dramatically expanded, can be expected to exhibit LD over large genomic intervals and greatly reduced allelic and haplotype diversity, both owing to genetic drift (13–15); such populations may be especially powerful for the initial phase of mapping common trait loci. Expanding or recently expanded populations are similarly expected to undergo rapid decay of LD over a relatively short time span (13,15,16,17); this characteristic may lend itself to finer localization of complex trait loci previously identified through linkage or other studies. Thus in the search for disease genes and similar complex trait loci it has been put forward that substantial advantage can be accrued by gaining access to multiple populations with divergent demographic histories, despite the strategy's obvious limitations. Our results, representing detailed analysis of several Finnish subpopulations, have important implications for the prospect and process of LD-based disease gene mapping. The long-range LD needed for coarse, genome-wide mapping of such traits can be found in carefully selected subpopulations, like the Kuusamo isolate, within an otherwise expanded population. Conversely, the LD decay associated with the main, rapidly growing population may offer an efficient tool for finer-scale gene localization.

Our data further demonstrate that, especially in recently bottlenecked population isolates like Kuusamo, highly polymorphic microsatellite markers can provide much greater power for detection of intermarker LD than can either single SNPs or SNP haplotypes. High heterozygosity microsatellite markers are more useful because one marker can simultaneously mark many haplotypes—both rare as well as common—meaning that they can be informative also for rare variant mapping, whereas common SNPs are only useful for measuring other common SNPs, and measure relatively few haplotypes at a single time. SNP analysis may still be a necessary component of the disease gene localization process, however, in that the far greater density of SNP polymorphisms allows for identification and delineation of ancestral haplotype 'blocks'. Our findings emphasize the value of detailed understanding of a study population's history and genealogy, as well as the importance of careful marker selection when designing studies which intend to use LD in genomewide scans for disease genes or other complex trait loci.

MATERIALS AND METHODS

DNA samples

Early settlement (ES) and late settlement (LS) consisted of families ($n = 16$, $n = 54$, respectively) initially ascertained for our nationwide schizophrenia study of sib-pairs (18). The isolate of Kuusamo (IS) consisted of families ($n = 54$) similarly

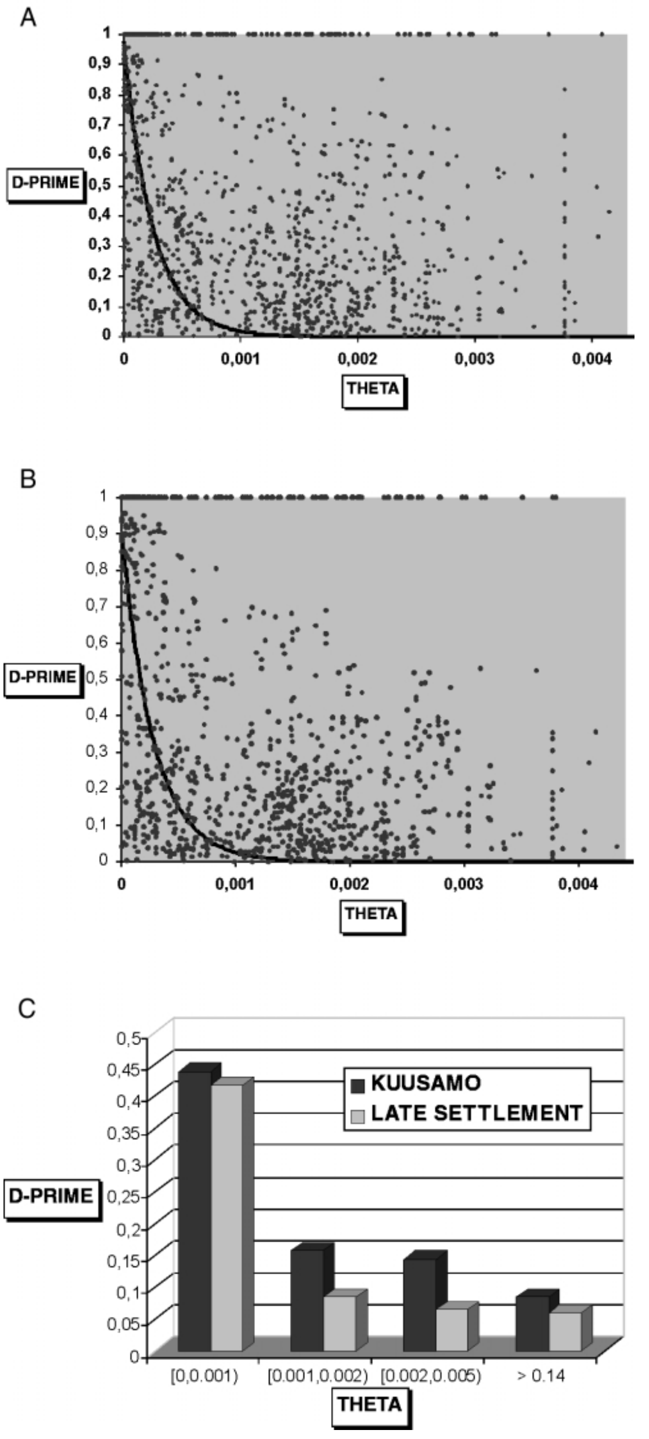

Figure 9. The estimated curves and the uncorrected point estimates of D' are shown for (**A**) Kuusamo, and (**B**) LS, and schematic comparison of the regions is given (**C**). The uncorrected point estimates for $\hat{D}'$ are provided for all marker pairs as a function of the intermarker distance, together with the estimated decay curve from the above equation. The null hypothesis in the global LD test using this procedure was that of no LD, meaning $D' = \hat{D}'_0$ for all values of θ. This was rejected in favor of the simple alternative $D' = \hat{D}'_0 + c$, where c was estimated from all marker loci jointly independent of intermarker distance. The null hypothesis was rejected here at $P \ll 10^{-100}$ as before, with c having a value of ~0.09–0.12 depending on which sample was examined, Kuusamo being the highest. The hypothesis of constant LD was similarly rejected in favor of the exponential model described above, at $P \ll 10^{-100}$.

collected throughout the community for our schizophrenia study (19). Total DNA was extracted from the leukocytes of frozen peripheral venous blood using standard procedures (20). This study has been approved by the ethical review board of the National Public Health Institute of Finland.

Genealogical search

Families were included in the study only if at least three out of four of their grandparents were born in the required region. If two individuals were found to be first cousins, one of them was excluded from the study. The genealogical study was performed in accordance with published criteria (6,21). The names, dates and places of birth of the patients' parents were used to trace ancestors back to the middle of the 1800s from local church and civil registers. Microfilm and microfiche copies of the church records in the Finnish National Archives were used for all the earlier periods.

Microsatellite typing

In the pilot study we chose two chromosomal regions with appropriate marker map according to our experience, a 27 cM region in 1q covered by 22 multi-allelic markers and 14 cM in 5q covered by 31 markers (skeleton maps). In the extended study, a 62 cM region in 1q was covered by all 212 multi-allelic markers available to us (including the markers in the pilot study), containing a 0.43 cM region of 45 SNPs and a 0.38 region of 23 SNPs that were 17 cM apart from each other (Fig. 3). Altogether, the ES and LS and IS were thus genotyped with 243 microsatellites and 68 SNPs.

Marker sequences were obtained from The Genome Database (www.gdb.org). SNPs were identified in public databases (SNP consortium). In addition, we identified novel SNPs by direct sequencing. PCR was performed according to standard procedures and electrophoresis was done on an ABI 377 sequencer (Applera Corporation, Norwalk, CT, USA). Marker order and inter-marker distances were obtained by RH mapping and by utilizing the Human Genome Project sequence data. When sequence data were used, the genetic distance was estimated from the physical distance assuming the equivalence of 1 000 000 bp to 1 cM.

In each family one to two parents, and two to seven children were genotyped. Haplotypes were constructed using Genehunter 2.1 (22) and for each family the first child's most likely haplotype was chosen for the study. Child's haplotypes proved to be more complete and thus more reliable than parents' haplotypes in the analysis mostly because of the sampling. Haplotypes were verified by eye, and the rate of ambiguity did not have an effect on the final LD estimates. LD was thus analyzed from 32, 108 and 108 chromosomes (two independent chromosomes per family) in ES, LS and IS, respectively.

Data analysis

In the study using the skeleton map markers, chromosome 1q contained 231 and chromosome 5q 465 marker pairs. In the fine map study, chromosome 1q contained altogether 22 366 multi-allelic marker pairs and 1243 SNP pairs. To describe the

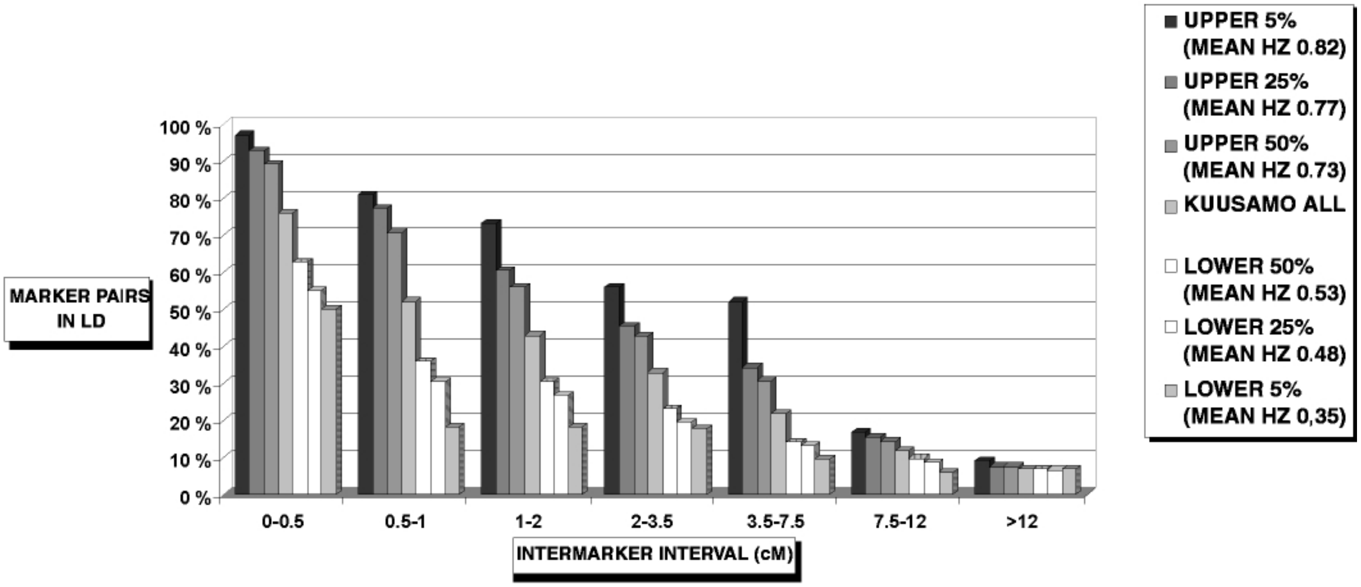

Figure 10. LD observed at different intervals on 1q in Kuusamo using pairs of microsatellites with different mean heterozygosity.

extent of non-random allelic association between pairs of loci, the tail probability (*P*-value) of Fisher's exact test was determined using GenePop 3.3 (March 2001) (23,24). Differences in allele frequency distributions at the microsatellite loci were evaluated by exact tests for population differentiation, as determined by GenePop 3.3. We also applied a likelihood ratio test for LD between the loci (24) (cf. also 24,25). For skeleton map study both these tests gave similar results, and the huge number of marker pairs in the fine mapping study allowed only the latter test to be applied for the fine map study. For D' bias see Weiss and Clark 2002 (7). We applied S-PLUS 2000 Professional Edition for Windows Release 1 (June 1999) for regression analysis.

ACKNOWLEDGEMENTS

We are grateful to the Finnish families for providing the data for this study. We are also indebted to Liisa Moilanen, Margit Keinänen, Outi McDonald, Pirkko Levón, Merja Nissi, Merja Hiltunen, Soili Johansson and Sisko Lietola for blood samples, and Perttu Haimi and Teemu Perheentupa for their excellent technical assistance. The work was supported by Millennium Pharmaceuticals, Inc. M.P. was supported by American Heart Association, Western States Award, Reference no. 0120121Y. The financial support of grant MH63749 from the National Institute of Mental Health (to J.D.T.) is acknowledged.

REFERENCES

1. Peltonen, L., Jalanko, A. and Varilo, T. (1999) Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.*, **8**, 1913–1923.
2. Nevanlinna, H.R. (1972) The Finnish population structure. A genetic and genealogical study. *Hereditas*, **71**, 195–236.
3. Norio, R. (1981) Diseases of Finland and Scandinavia. In Rotschild, H. (ed.), *Biocultural Aspects of Disease.* Academic Press, New York, pp. 359–415.
4. Norio, R., Nevanlinna, H.R. and Perheentupa, J. (1973) Hereditary diseases in Finland; rare flora in rare soul. *Ann. Clin. Res.*, **5**, 109–141.
5. Pastinen, T., Perola, M., Ignatius, J., Sabatti, C., Tainola, P., Levander, M., Syvanen, A.C. and Peltonen, L. (2001) Dissecting a population genome for targeted screening of disease mutations. *Hum. Mol. Genet.*, **10**, 2961–2972.
6. Varilo, T., Savukoski, M., Norio, R., Santavuori, P., Peltonen, L. and Järvelä, I. (1996) 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.
7. Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.*, **18**, 19–24.
8. Varilo, T., Laan, M., Hovatta, I., Wiebe, V., Terwilliger, J.D. and Peltonen, L. (2000) Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur. J. Hum. Genet.*, **8**, 604–612.
9. Mohlke, K.L., Lange, E.M., Valle, T.T., Ghosh, S., Magnuson, V.L., Silander, K., Watanabe, R.M., Chines, P.S., Bergman, R.N., Tuomilehto, J. *et al.* (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Res.*, **11**, 1221–1226.
10. Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A.J., Deloukas, P., Olsen, A., Doggett, N.A., Ghebranious, N., Broman, K.W. *et al.* (2001) Comparison of human genetic and sequence-based physical maps. *Nature*, **409**, 951–953.
11. Broman, K.W., Murray, J.C., Sheffield, V.C., White R.L. and Weber J.L. (1998) Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, **63**, 861–889.
12. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
13. Laan, M. and Paabo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat. Genet.*, **17**, 435–438.
14. Laan, M. and Paabo, S. (1998) Mapping genes by drift-generated linkage disequilibrium. *Am. J. Hum. Genet.*, **63**, 654–656.
15. Terwilliger, J.D., Zöllner, S., Laan, M. and Pääbo, S. (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: "Drift Mapping" in small populations with no demographic expansion. *Hum. Hered.*, **48**, 138–154.
16. Slatkin, M. (1994) Linkage disequilibrium in growing and stable populations. *Genetics*, 331–336.
17. Crouau-Roy, B., Service, S., Slatkin, M. and Freimer, N. (1996) A fine-scale comparison of the human and chimpanzee genomes: linkage, linkage disequilibrium and sequence analysis. *Hum. Mol. Genet.*, **5**, 1131–1137.

18. Ekelund, J., Lichtermann, D., Hovatta, I., Ellonen, P., Suvisaari, J., Terwilliger, J.D., Juvonen, H., Varilo, T., Arajarvi, R., Kokko-Sahin, M.L. *et al.* (2000) Genome-wide scan for schizophrenia in the Finnish population: evidence for a locus on chromosome 7q22. *Hum. Mol. Genet.*, **9**, 1049–1057.
19. Hovatta, I., Varilo, T., Suvisaari, J., Terwilliger, J.D., Ollikainen, V., Arajarvi, R., Juvonen, H., Kokko-Sahin, M.L., Vaisanen, L., Mannila, H. *et al.* (1999) A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am. J. Hum. Genet.*, **65**, 1114–1124.
20. Vandenplas, S., Wiid, I., Grobler-Rabie, A. *et al.* (1984) Blot hybridization analysis of genomic DNA. *J. Med. Genet.*, **21**, 164–172.
21. Varilo, T. (1999) The age of the mutations in the Finnish disease heritage; a genealogical and linkage disequilibrium study. Ph.D. thesis, National Public Health Institute and University of Helsinki, Helsinki (http://ethesis.helsinki.fi/english.html).
22. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.
23. Raymond, M. and Rousset, F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.*, **86**, 248–249.
24. Terwilliger, J.D. and Ott, J. (1994) *Handbook of Human Genetic Linkage.* The Johns Hopkins University Press, Baltimore, MD.
25. Goring, H.H. and Terwilliger, J.D. (2000) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.*, **66**, 1310–1327.

APPENDIX ELECTRONIC-DATABASE INFORMATION

See the authors' website for additional article data, http://oxygen.ktl.fi/molbio/wwwpub/. Table A on our website: information from the genetic map and polymorphic loci genotyped. Table B on our website: the genotype data. Figures A–J on our website: the expected value of the MLE of D', $\hat{D}'$, is a function of the true value of D', and the allele frequencies of the two markers. When $D' = 1$, there is no bias, but when $D' = 0$ the bias can be enormous, especially when the SNP alleles are rare.