The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes

Iain A. Eaves[1], Tony R. Merriman[2], Rachael A. Barber[2], Sarah Nutland[1], Eva Tuomilehto-Wolf[3], Jaakko Tuomilehto[3], Francesco Cucca[4] & John A. Todd[1]

The choice of which population to study in the mapping of common disease genes may be critical[1,2]. Isolated founder populations, such as that found in Finland, have already proved extremely useful for mapping the genes for specific rare monogenic disorders[3,4] and are being used in attempts to map the genes underlying common, complex diseases[5–8]. But simulation results suggest that, under the common disease-common variant hypothesis[9–13], most isolated populations will prove no more useful for linkage disequilibrium (LD) mapping of common disease genes than large outbred populations[12]. There is very little empirical data to either support or refute this conclusion at present[14–16]. Therefore, we evaluated LD between 21 common microsatellite polymorphisms on chromosome 18q21 in 2 genetic isolates (Finland and Sardinia) and compared the results with those observed in two mixed populations (United Kingdom and United States of America). Mean levels of LD were similar across all four populations. Our results provide empirical support for the expectation that genetic isolates like Finland and Sardinia will not prove significantly more valuable than general populations for LD mapping of common variants underlying complex disease.

We genotyped 21 polymorphic microsatellite loci, which map to a 6.5-cM interval on chromosome 18q21, in four large, diabetic family data sets (Table 1) and reconstructed 800 unrelated haplotypes for each. Although the putative type 1 susceptibility locus *IDDM6* maps to this region[17,18], the pattern of LD in our samples was representative of the general population as it was no different from that observed in affected family-based control[19] (AFBAC) data sets (data not shown). This was not surprising, as any ascertainment bias introduced would have been minimal because the risk conferred by the putative *IDDM6* locus is very small[17] (λ_s=1.1).

Initially, we quantified the LD between all 210 possible pairs of loci using a multiallelic extension of Lewontin's standardized measure of disequilibrium, D´ (ref. 20). The patterns of LD in the four populations were similar, although mean levels were slightly elevated in the Finns (Fig. 1). In particular, we observed higher mean levels of LD in the Finnish data set between markers separated by more than 0.5 cM. As LD was at best modest (mean D'_m<0.35) over this distance regardless of the population, this increase was unlikely to be of significance for mapping disease genes. Over short distances (<0.5 cM), we observed high levels of

Table 1 • Features and relative locations of 21 microsatellites

		Finland		Sardinia		UK		USA	
Marker	cM	locus diversity	no. alleles	locus diversity	no. alleles	locus diversity	no. alleles	locus diversity	no. alleles
D18S851	0.0	0.74	7	0.76	6	0.76	6	0.76	6
D18S484	0.4	0.73	6	0.68	6	0.69	6	0.68	6
AFM318xd5	1.0	0.45	5	0.36	5	0.40	5	0.38	5
D18S1156	1.2	0.59	6	0.61	5	0.59	6	0.57	6
252,16	1.3	0.50	2	0.49	2	0.50	2	0.50	2
88,21	1.5	0.59	10	0.65	10	0.65	9	0.64	9
114,1	2.3	0.76	9	0.76	10	0.73	10	0.74	10
30T7	2.4	0.69	6	0.66	7	0.70	7	0.69	6
129,6	2.5	0.86	14	0.88	14	0.88	14	0.87	14
129,12	2.7	0.60	3	0.54	4	0.56	4	0.57	4
129,11	2.8	0.84	17	0.88	16	0.86	16	0.86	16
IO43,56	2.9	0.86	12	0.85	13	0.86	13	0.87	13
D18S487	3.0	0.82	9	0.79	8	0.80	9	0.80	9
A181,2	3.0	0.41	6	0.32	7	0.47	8	0.47	6
49,22	3.3	0.39	5	0.37	4	0.42	4	0.42	5
296,11	4.3	0.49	2	0.48	2	0.49	2	0.50	2
296,7	4.4	0.70	8	0.78	7	0.74	8	0.77	8
D18S35	4.8	0.68	6	0.66	7	0.69	6	0.70	6
D18S69	5.2	0.75	6	0.79	8	0.78	8	0.77	7
D18S39	5.8	0.76	8	0.80	10	0.79	10	0.80	9
D18S41	6.5	0.67	7	0.71	7	0.66	7	0.68	7
mean value	–	0.66	7.3	0.66	7.5	0.67	7.6	0.67	7.4

[1]*Wellcome Trust Centre for Molecular Mechanisms in Disease, Department of Medical Genetics, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge, UK.* [2]*Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, Oxford, UK.* [3]*Diabetes and Genetic Epidemiology Unit, National Public Health Institute, Helsinki, Finland.* [4]*Instituto di Clinica e Biologia dell'Eta' Evolutiva, University of Cagliari, Italy. Correspondence should be addressed to J.A.T. (e-mail: john.todd@cimr.cam.ac.uk).*

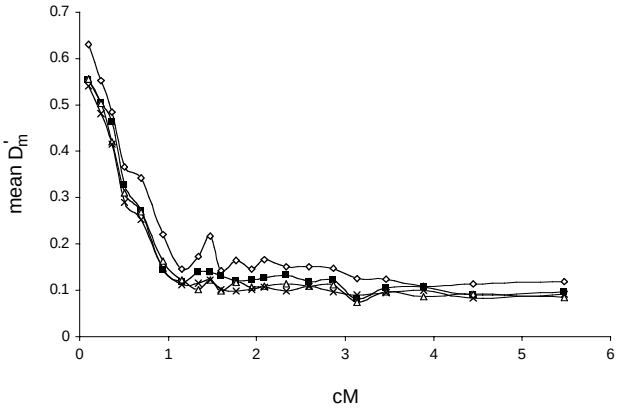

Fig. 1 Relationship between mean level of LD and marker separation for the four populations studied. Each point represents the average of 10 marker pairs: ◊, Finland; ■, Sardinia; Δ, UK; ×, USA.

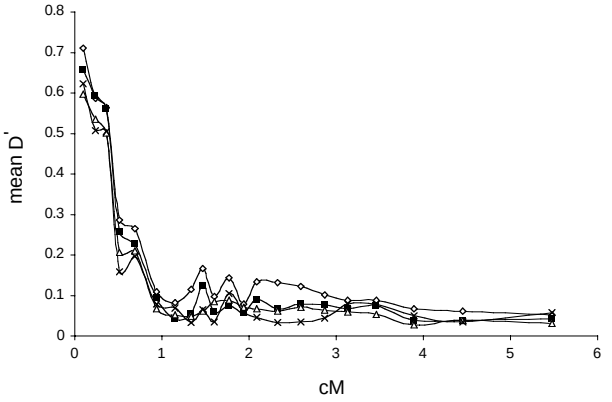

Fig. 2 Relationship between mean level of LD and marker separation after classifying alleles into two groups. Each point represents the average of 10 marker pairs: ◊, Finland; ■, Sardinia; Δ, UK; ×, USA. For each pair of loci the unsigned absolute value of D´ was calculated.

LD in all four populations, and the percentage increase in mean D'_m values seen in the Finnish data was small.

We dichotomized each microsatellite to compare our data with previous results[12] (Table 2). The pattern of LD between these 'artificial' biallelic markers mirrored that seen before the pooling of alleles (Fig. 2). In all four data sets, mean LD decayed rapidly with increasing distance before reaching a very low or background level for markers separated by approximately 1 cM or more. Because Kruglyak[12] used d^2 as the measure of association in his simulations, we also calculated this statistic (Fig. 3). As expected, the distribution of LD across the four populations was very similar. In contrast, the actual levels of LD were considerably higher than predicted for comparable single-nucleotide polymorphisms[12] (SNPs), with maximum mean d^2 values ranging from 0.24 to 0.29. We had expected the higher mutation rate at microsatellites to result in lower LD between our 'artificial' biallelic markers than between common SNPs. Others[21] have also reported higher levels of LD than those predicted, and it may be that because these simulations[12] were based on a simplified scenario of population history, they underestimated the degree of LD that should be expected.

Care is required when drawing general conclusions from individual LD studies, as LD is thought to be distributed heterogeneously across the genome[22]. Therefore, the absolute levels of LD reported here may be less informative than the relative levels across the four populations. Although we have only surveyed one genome region, our large set of data lends empirical support to the expectation that genetic isolates such as Finland and Sardinia will not prove significantly more valuable than general populations of European descent for LD mapping common variants that underlie common disease[12]. Neither the Finnish nor the Sardinian samples studied here displayed greatly elevated levels of LD compared with those from the UK and USA. The most likely explanation for this pattern is the introduction of most common variants (>1%) to the two isolates by multiple founders, such that the recombinational histories of the variants extend to their origin in the general popu-

Table 2 • Minor allele frequencies after classifying alleles into two groups

Marker	FIN	SAR	UK	USA
D18S851	0.37	0.46	0.45	0.40
D18S484	0.30	0.36	0.41	0.37
AFM318xd5	0.26	0.17	0.23	0.20
D18S1156	0.48	0.41	0.49	0.46
252,16	0.46	0.43	0.49	0.45
88,21	0.37	0.44	0.44	0.43
114,1	0.35	0.32	0.31	0.31
30T7	0.29	0.46	0.40	0.35
129,6	0.24	0.25	0.27	0.26
129,12	0.42	0.32	0.32	0.33
129,11	0.47	0.40	0.44	0.44
IO43,56	0.22	0.19	0.26	0.24
D18S487	0.33	0.29	0.25	0.26
A181,2	0.13	0.08	0.15	0.13
49,22	0.18	0.17	0.21	0.20
296,11	0.42	0.41	0.43	0.46
296,7	0.45	0.43	0.48	0.48
D18S35	0.30	0.26	0.32	0.34
D18S69	0.42	0.50	0.49	0.50
D18S39	0.33	0.21	0.31	0.29
D18S41	0.39	0.42	0.37	0.41
mean value	0.34	0.33	0.36	0.35

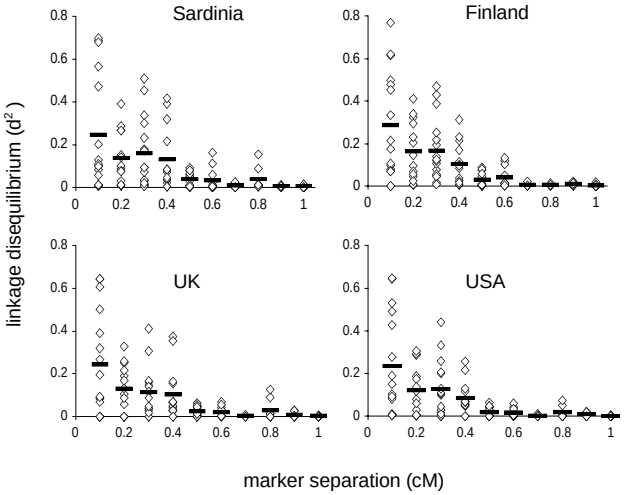

Fig. 3 Distribution of d^2 values for marker separations of ≤1 cM. Each open diamond represents an individual d^2 value and mean values are represented by horizontal bars. Both possible values of d^2 are plotted for each marker pair.

lation[12]. In support of this, we observed almost identical mean locus diversities across our four data sets (0.66–0.67). Our results do not exclude the possibility that LD may extend further in other genetic isolates or sub-isolates[14]; however, these populations are likely to be small, making the collection of adequate samples difficult[15]. Moreover, although it is likely that some of the genetic variation underlying common disease can be attributed to relatively uncommon alleles (<1%) that will be surrounded by higher levels of LD in populations such as Finland, their low frequency means that enormous data sets will be required to demonstrate statistically significant disease associations. Given our results and these additional considerations, we favour establishing large, well-ascertained collections of clinical material from non-founder populations such as the UK and from isolated founder populations. These resources will ultimately help identify many of the common variants that underlie common disease. Given the current limitations in throughput and cost for SNP typing, however, present studies will probably have to focus on identifying those in known functional and positional candidate genes.

Methods

Families. All families in this study were white European or of white European origin and were selected for the existence of at least one type 1 diabetic offspring. The UK data set comprised 264 multiplex families; the USA data set, 264 largely North American (a few were from Europe itself) multiplex families from the Human Biological Data Interchange repository of type 1 diabetic families (http://www.hbdi.org/). The Sardinian data set comprised 257 and the Finnish data set, 239 simplex families.

Genotyping and mapping data. Genotyping PCRs for 21 polymorphic microsatellites were performed and analysed as described[23]. Primer sequences for *D18S851*, *D18S484*, AFM318xd5, *D18S1156*, *D18S35*, *D18S69*, *D18S39* and *D18S41* are available from the Genome Database (http://www.gdb.org/). We rescued microsatellites 252,16, 296,7, 296,11 and 88,21 using a described PCR-based method[17] (mapping data and primer sequences for these are available on request). Sequences for amplifying all other microsatellite markers have been described[17,18]. With the exceptions of *D18S35*, *D18S39* and *D18S41*, we were able to order all markers using our own physical contigs[18] or the Whitehead YAC contig WC18.4 (http://www-genome.wi.mit.edu/). To establish the best genetic order for all 21 markers, we used the software Genome Analysis System (http://users.ox.ac.uk/~ayoung/gas.html), which uses a version of the Metropolis algorithm based on simulated annealing. There were no discrepancies between the physical and genetic maps. Having ordered the markers, we double checked all suspect recombination events before using the ASPEX software (ftp://lahmed.stanford.edu/pub/aspex/index.html) to generate multipoint maximum likelihood estimates of the map distances.

Haplotype reconstruction. We constructed haplotypes using software available from our ftp site (ftp://ftp-gene.cimr.cam.ac.uk/pub/software/). Briefly, for each family we determined the parental origin of the offspring alleles for all possible loci. Where appropriate we used the genotype information of a tightly linked adjacent marker to resolve phase. Cases where phase could not be resolved unambiguously were treated as if the genotype information was missing. By choosing a random offspring from each family, it was possible to identify four unrelated chromosomes: the two received paternally and maternally, respectively, and the two not transmitted to the offspring. Under the assumption of no recombination, these would be equivalent to the parental chromosomes themselves. To ensure that we were analysing equivalent information for each marker pair in each of the four populations, we discarded those haplotypes which held no typing information at multiple loci, such that no more than 10% of the remaining 800 haplotypes were untyped at any particular locus. We constructed an affected family-based control (AFBAC) data set for each of the four populations. These comprised haplotypes selected as described[19] for single alleles, with the modification that for the multiplex UK and USA data sets only non-recombinant chromosomes could be included. The AFBAC population gave an unbiased estimate of the population haplotype frequencies[19].

Pooling of alleles. To emulate biallelic markers and allow comparisons with previous results, we classified the alleles at each microsatellite marker into two groups. Of the 21 microsatellites we studied, 2 were already biallelic and 10 exhibited a clear bimodal distribution of alleles in all 4 populations. We grouped the alleles at these 10 loci according to the 2 modes, as we wanted to maintain ancestral relationships. Alleles at the remaining loci were approximately unimodally distributed and we classified these into two groups defined by the two most common alleles central to the distribution. Where there were discrepancies between the populations as to which two alleles were the most common, grouping was performed according to the most consistent pattern. For example, the 6 alleles of *D18S851* were found to be distributed as follows in the UK data set: allele 1, 6.3%; 2, 14.2%; 3, 34.5%; 4, 26.9%; 5, 15.2%; 6, 2.9%. In the USA and Sardinian data sets, alleles 3 and 4 were also the 2 most common alleles central to the distribution, but in the Finnish samples allele 2 was the second most common. On balance, we grouped alleles 1–3 and 4–6 together, respectively.

Statistics. We calculated locus diversity as:

$$\frac{n}{n-1}\,(1-\sum_i p_i^2)\,,$$

where p_i is the estimated frequency of the i^{th} allele at the locus and n is the number of haplotypes[24]. We calculated a multiallelic extension of the normalized association measure D´ (ref. 20) as:

$$D_m' = \sum_{i=1}^{k}\sum_{j=1}^{l} p_i q_j \left| D_{ij}' \right|\,,$$

where p and q are the observed allele frequencies at the two loci[25]. For comparison with Kruglyak's study[12], we calculated d^2 as $D^2/(f(1-f))^2$, where f represents the frequency of the putative disease variant[12]. As either locus in each pair could be assumed to represent a putative disease locus, f could take the value of either p or q and we calculated both possible values for d^2.

Acknowledgements

This work was funded by the Wellcome Trust, British Diabetic Association (BDA), UK Medical Research Council and the Juvenile Diabetes Foundation International. We are grateful for the assistance of the 'Childhood Diabetes in Finland (DiMe) Study Group' in the collection of the Finnish family material. The Finnish family collaboration was partially funded by grants from NIH (DK 73957) and the Novo Nordisk Foundation. The BDA and the Human Biological Data Interchange are thanked for the collection of families. F.C. was supported in part by a grant from Assessorato Igiene e Sanità, Regione Sardegna.

Received 13 January; accepted 20 April 2000.

1. Terwilliger, J.D. & Weiss, K.M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9**, 578–594 (1998).
2. Wright, A.F., Carothers, A.D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nature Genet.* **23**, 397–404 (1999).
3. de la Chapelle, A. & Wright, F.A. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl Acad. Sci. USA* **95**, 12416–12423 (1998).
4. Peltonen, L., Jalanko, A. & Varilo, T. Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.* **8**, 1913–1923 (1999).
5. Kuokkanen, S. *et al.* Genomewide scan of multiple sclerosis in Finnish multiplex families. *Am. J. Hum. Genet.* **61**, 1379–1387 (1997).
6. Escamilla, M.A. *et al.* Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. *Am. J. Hum. Genet.* **64**, 1670–1678 (1999).
7. Ghosh, S. *et al.* Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc. Natl Acad. Sci. USA* **96**, 2198–2203 (1999).
8. Pajukanta, P. *et al.* Genomewide scan for familial combined hyperlipidemia genes in Finnish families, suggesting multiple susceptibility loci influencing triglyceride, cholesterol, and apolipoprotein B levels. *Am. J. Hum. Genet.* **64**, 1453–1463 (1999).
9. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
10. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
11. Collins, F.S., Guyer, M.S. & Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
12. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
13. Todd, J.A. Multifactorial diseases: ancient gene polymorphism at quantitative trait loci and a legacy of survival during our evolution. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C.R. *et al.*) (McGraw-Hill, New York-London, in press).
14. Laan, M. & Paabo, S. Demographic history and linkage disequilibrium in human populations. *Nature Genet.* **17**, 435–438 (1997).
15. Lonjou, C., Collins, A. & Morton, N.E. Allelic association between marker loci. *Proc. Natl Acad. Sci. USA* **96**, 1621–1626 (1999).
16. Jorde, L.B., Watkins, W.S., Kere, J., Nyman, D. & Eriksson, A.W. Gene mapping in isolated populations: new roles for old friends? *Hum. Hered.* **50**, 57–65 (2000).
17. Merriman, T. *et al.* Evidence by allelic-association dependent methods for a type 1 diabetes polygene (IDDM6) on chromosome 18q21. *Hum. Mol. Genet.* **6**, 1003–1010 (1997).
18. Merriman, T.R. *et al.* Transmission of haplotypes of microsatellite markers rather than single marker alleles in the mapping of a putative type 1 diabetes susceptibility gene (IDDM6). *Hum. Mol. Genet.* **7**, 517–524 (1998).
19. Thomson, G. Mapping disease genes: family-based association studies. *Am. J. Hum. Genet.* **57**, 487–498 (1995).
20. Lewontin, R.C. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
21. Collins, A., Lonjou, C. & Morton, N.E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
22. Huttley, G.A., Smith, M.W., Carrington, M. & O'Brien, S.J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).
23. Reed, P.W. *et al.* Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet.* **7**, 390–395 (1994).
24. Weir, B.S. *Genetic Data Analysis* (Sinauer Associates, Sunderland, Massachusetts, 1996).
25. Hedrick, P.W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341 (1987).