Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity

Andrew P. Morris*, John C. Whittaker†, Chun-Fang Xu‡, Louise K. Hosking‡, and David J. Balding†§

*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom; †Department of Epidemiology and Public Health, Imperial College, London W2 1PG, United Kingdom; and ‡Discovery Genetics, GlaxoSmithKline, Stevenage SG1 2NY, United Kingdom

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved September 15, 2003 (received for review August 8, 2003)

Single-nucleotide polymorphism (SNP) genotypes were recently examined in an 890-kb region flanking the human gene *CYP2D6*. Single-marker and haplotype-based analyses identified, with genomewide significance ($P < 10^{-7}$), a 403-kb interval displaying strong linkage disequilibrium (LD) with predicted poor-metabolizer phenotype. However, the width of this interval makes the location of causal variants difficult: for example, the interval contains seven known or predicted genes in addition to *CYP2D6*. We have developed the Bayesian fine-mapping software COLDMAP, which, applied to these genotype data, yields a 95% location interval covering only 185 kb and establishes genomewide significance for a causal locus within the region. Strikingly, our interval correctly excludes four SNPs, which individually display association with genomewide significance, including the SNP showing strongest LD ($P < 10^{-34}$). In addition, COLDMAP distinguishes homozygous cases for the major *CYP2D6* mutation from those bearing minor mutations. We further investigate a selection of SNP subsets and find that previously reported methods lead to a 38% savings in SNPs at the cost of an increase of <20% in the width of the location interval.

Linkage-disequilibrium (LD) mapping using high-density single-nucleotide polymorphism (SNP) maps is already useful in identifying genes involved in complex diseases (1). Its role will grow substantially, as recent reports of the extent of LD in the human genome (2–6) make detecting associations more feasible than predicted (7). Conversely, these findings suggest that intervals displaying association may be relatively wide and hence contain many genes. The challenge is then to refine techniques for fine-mapping of the causal polymorphism(s) within regions of high LD.

Recently, the efficacy of LD mapping has been confirmed for genes involved in drug response (8). Genotypes of 1,018 individuals were obtained for 32 SNP markers located in an 890-kb region flanking the *CYP2D6* gene on human chromosome 22q13. Functional polymorphisms at the *CYP2D6* locus were also typed, and 41 individuals were found to carry two mutant alleles so that they were predicted to have a "poor-metabolizer" phenotype. LD mapping identified a 403-kb region, including the *CYP2D6* locus, displaying strong LD with this predicted phenotype. LD was sufficiently strong that genomewide significance ($P < 10^{-7}$) was achieved for 10 of the markers, even when analyzed individually so that joint information from neighboring markers was ignored. Inferring haplotypes from the SNP genotypes and then analyzing five-marker haplotypes led to even higher levels of significance but not to any narrowing of the 403-kb high-LD interval.

Here, we use the data of Hosking *et al.* (8) to demonstrate the power of multipoint LD mapping to narrow location intervals. Our Bayesian method explicitly models mutation and recombination histories using the shattered coalescent model (9) and allows for allelic heterogeneity at the functional locus. Because it separates out sporadic case chromosomes and delineates the contributions to LD of different founding mutation events, it can identify a 95% location interval much narrower than the interval displaying high LD.

Methods

Full details of the data are given in ref. 8, but note that distances reported here are based on updated SNP locations provided by the authors subsequent to publication and differ slightly from the values reported in ref. 8. Genotypes for 1,081 individuals were obtained for 32 SNPs, spanning an 890-kb region encompassing the *CYP2D6* gene on chromosome 22q13. Five SNPs had sample allele frequencies <10% and were not used in the statistical analyses contained in ref. 8. Because our multipoint method simultaneously analyzes all of the data, there is no problem with sensitivity to allele frequency, and we retain all 32 SNPs for our analyses.

Our analysis is based on the shattered coalescent model (9) of the genealogy underlying a sample of case chromosomes in the vicinity of a putative disease locus, together with a simpler first-order Markov assumption for the controls. The shattered coalescent model is more realistic than the star-shaped tree model implicitly assumed by many existing multipoint methods. Within our modeling framework, we can allow for missing marker information and uncertainty about both the true underlying genealogy and the makeup of ancestral marker haplotypes. The model is implemented by means of Markov chain Monte Carlo within the COLDMAP software (COalescent LD MAPping). The output of COLDMAP leads to approximations of the posterior distributions of the location of the disease locus and of the marker allele frequencies. In addition, the output permits construction of a cladogram that can indicate genetic heterogeneity by means of clusters of individuals corresponding to different genotypes at the disease locus. Simulation studies (9) demonstrate that inferences about the location of the disease locus are robust to many of the modeling assumptions.

The COLDMAP analyses presented here feature two substantial improvements over those described in ref. 9. First, our previous algorithm was restricted to haplotype data; COLDMAP directly analyzes genotypes. Initially, the phase of each genotype at each locus is assigned at random. Periodically, a new phase allocation is proposed by selecting an individual at random and then swapping the current phase allocation at all loci either to the right or to the left of a randomly chosen location. The proposal is accepted or rejected according to the usual rules for a Metropolis sampler (10). Because there exist good algorithms for inferring haplotypes from genotype data (11), this may seem

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: LD, linkage disequilibrium; SNP, single-nucleotide polymorphism.

§To whom correspondence should be addressed at: Department of Epidemiology and Public Health, Imperial College, St. Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom. E-mail: d.balding@ic.ac.uk.

www.pnas.org/cgi/doi/10.1073/pnas.2235031100

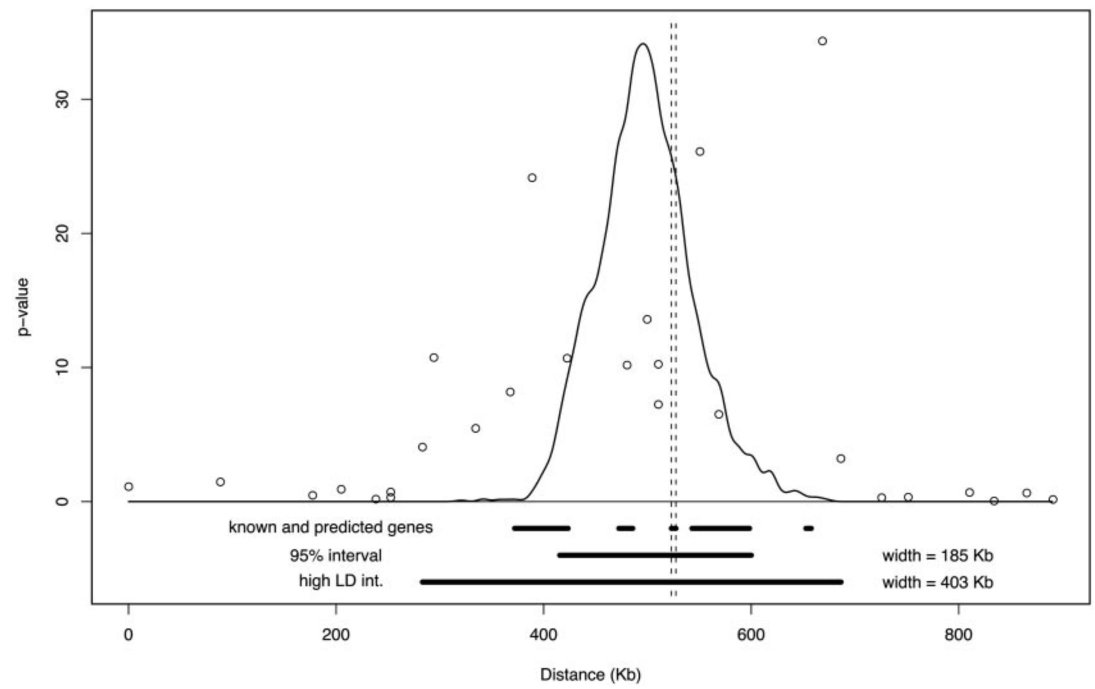

Fig. 1. Location of a functional polymorphism underlying the predicted poor-metabolizer phenotype within the 890-kb candidate region studied by Hosking *et al.* (8). The curve shows the posterior probability distribution estimated by using COLDMAP, applied to the 41 cases only. Beneath the *x* axis, horizontal bars indicate the locations of the known and predicted genes in the high-LD interval (based on NCBI34 build), 185-kb 95% posterior interval, and the 403-kb high-LD interval (8). The circles indicate $-\log_{10}(P\,\text{value})$ for the 27 SNPs tested in ref. 8 (data from table 2 in ref. 8). The vertical dashed lines indicate the location of *CYP2D6*.

a small advantage. However, for the goal of locating causal variants, the phase of genotype data is not of central interest, and there are substantial benefits in analyzing genotype data directly, avoiding this intermediate step. COLDMAP averages over phase allocations, weighted according to their plausibilities under the shattered coalescent model. This allows the genotype data to be fully used, even if there is uncertainty about the correct haplotypes. Perhaps more importantly, phenotype information can inform the haplotyping, which in turn makes the gene mapping more precise.

Our second innovation is that, in addition to a standard analysis, we ran COLDMAP with some controls mislabeled as cases. Such mislabeling would be seriously detrimental to some alternative methods of analysis, but because COLDMAP can identify and exclude sporadic cases, it suffers little from the misclassification. The purpose of this deliberate misclassification was to allow us to test the null hypothesis of no causal locus within the interval, because under this hypothesis the mislabeled controls should not be distinguishable from cases. To test this hypothesis, we identify a major cluster of individuals labeled as cases and count the number of true cases within this cluster. This number has a hypergeometric null distribution, for which *P* values are readily computed. For example, if there are 100 individuals labeled as cases in the analysis, 50 true cases and 50 mislabeled controls, and the largest cluster size is 40 and includes 35 true cases, then the *P* value is the probability of observing ≥35 white balls in a sample size of 40 drawn from an urn containing 50 white balls and 50 red balls, which is 3.6×10^{-10}. In the analyses below, we chose the number of mislabeled controls to equal the number of cases, giving good statistical efficiency relative to computational cost.

Results and Discussion

The curve in Fig. 1 shows the posterior density of the location of a functional polymorphism underlying the poor-metabolizer phenotype, for the COLDMAP analysis of the 41 cases only. The 95% equal-tailed posterior probability interval runs from 415 to 600 kb (distances measured from the first SNP), less than half the length of the high-LD interval (8). Four of the ten SNPs displaying genomewide significance (8) are correctly excluded from our interval, including that at 668 kb, which displays the strongest single-marker association ($P < 10^{-34}$), and that at 389 kb, which is the third strongest ($P < 10^{-24}$). Three of the seven known or predicted genes within the high-LD region, other than *CYP2D6*, are excluded from our interval.

Fig. 2 shows a cladogram of the 82 case chromosomes at the putative functional locus. The cladogram was formed by average-link hierarchical cluster analysis (12) implemented in S-plus, with the distance between any two chromosomes defined to be the proportion of COLDMAP outputs for which they were not assigned to the same subtree of the shattered coalescent. The figure shows a strong separation between the chromosomes from the 32 individuals homozygous for the major *CYP2D6* mutation (G1846A) and those from the individual carrying two minor mutations.

The chromosomes from the eight individuals carrying one major and one minor mutation are less easy to interpret in Fig. 2, because the data consist of unphased genotypes. If LD is sufficiently strong in the region of CYP2D that phase is correctly resolved, we would expect one chromosome from each of these individuals to cluster with the G1846A chromosomes and one with the minor chromosomes (although the latter clustering may be weaker). This is indeed broadly the case (Fig. 2).

Because of the problem of interpretation for unphased genotypes, we also clustered individuals, rather than chromosomes. At each iteration of the algorithm, the distance between two individuals was taken to be 1 if no pair of chromosomes, one from each individual, was in the same subtree of the shattered coalescent. If this held for one pair of chromosomes, the distance was 0.5, and if two nonoverlapping pairs could be found satisfying this condition, the distance was 0. Fig. 3 shows the

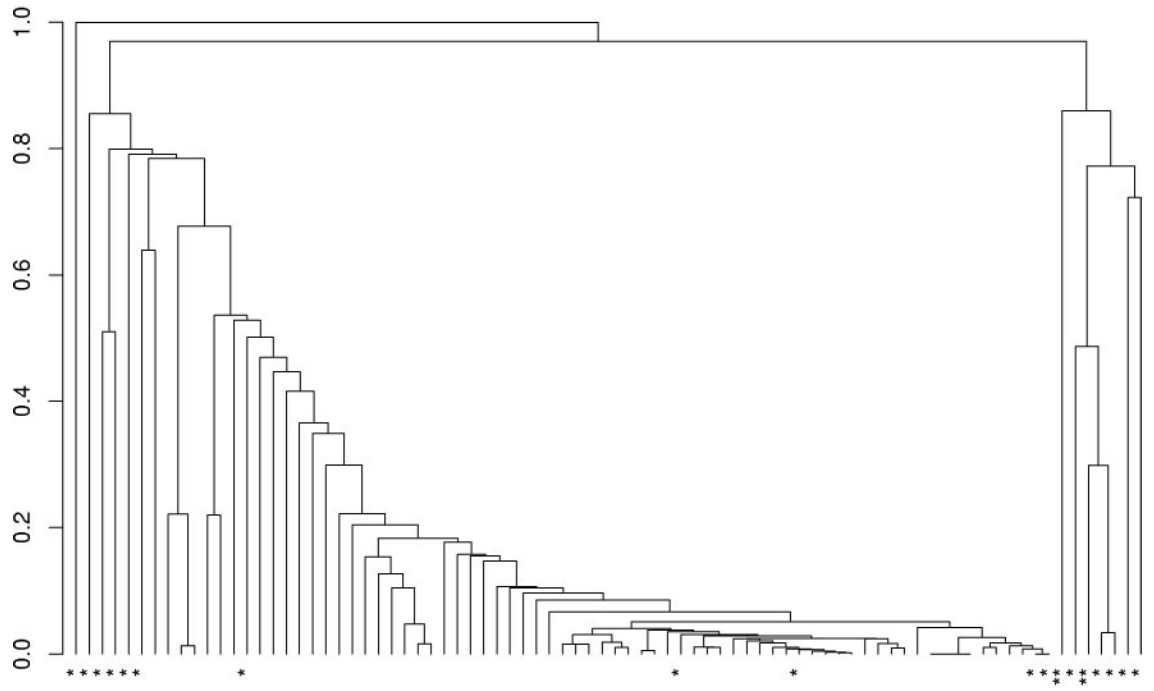

Fig. 2. Cladogram of the 82 chromosomes from the 41 predicted poor-metabolizer individuals (cases). The pairwise distance measure (*y* axis) is the proportion of COLDMAP outputs in which the chromosomes are allocated to different subtrees. Chromosomes from individuals carrying one minor *CYP2D6* mutant allele are marked with *, whereas those from the individual carrying two minor mutant alleles are marked with **.

cladogram based on these distances. It shows a clear separation of the G1846A homozygotes from those carrying, respectively, one and two minor *CYP2D6* mutations. This ability to characterize the allelic heterogeneity at the causal locus is the source of COLDMAP's ability to narrow the interval of plausible locations within the LD interval.

The above analysis assumes that there is a causal locus within the candidate region and hence always finds a location for it. To provide an assessment of the weight of evidence for or against the presence of a causal locus within the region, we ran COLDMAP again with 41 controls, chosen at random from the 977 available, mislabeled as cases. Eight of the controls selected were het-

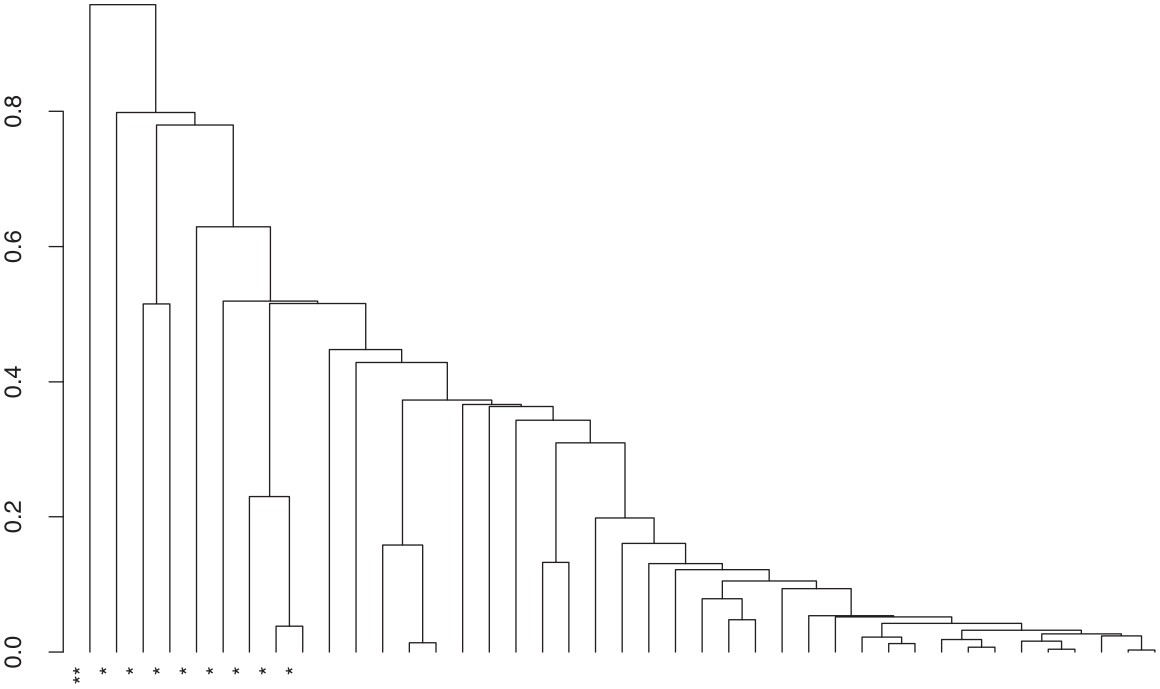

Fig. 3. Cladogram of the 41 cases; see *Results and Discussion* for definition of distance. * denotes individuals with one minor mutant allele, and ** denotes the individual with two minor mutant alleles.

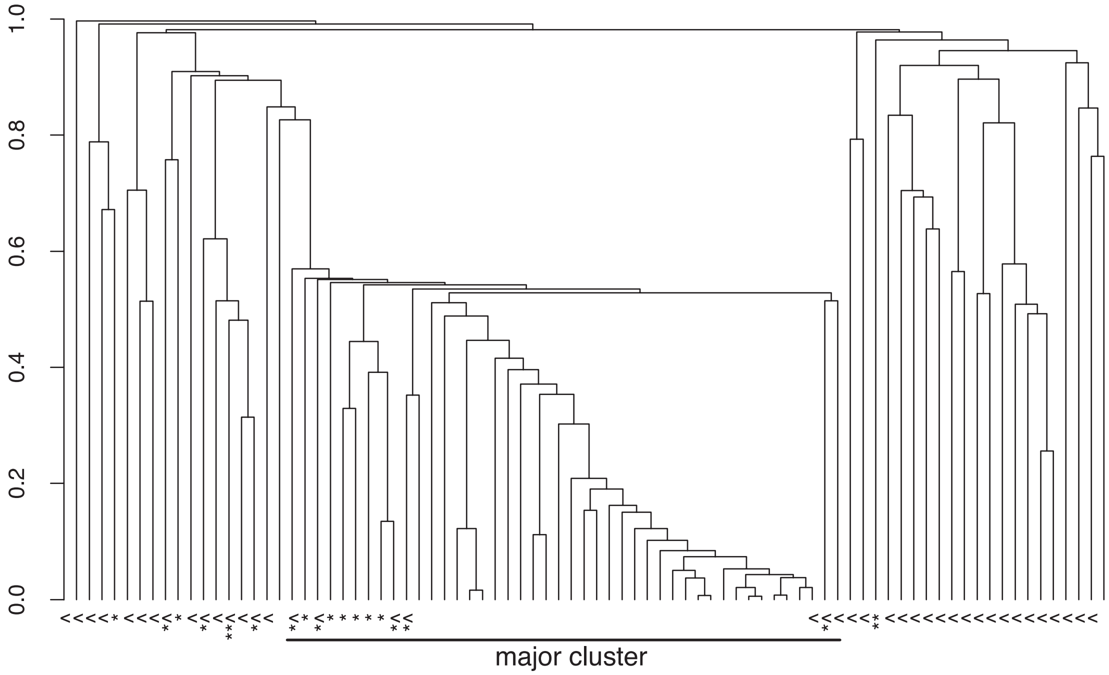

Fig. 4. Cladogram of 41 cases and 41 controls, the latter selected at random and analyzed as cases. The cases are labeled as for Fig. 3. Controls homozygous for the normal *CYP2D6* allele are indicated with ∧; those carrying one G1846A mutant allele are denoted with *∧, and **∧ denotes the control carrying a minor *CYP2D6* mutation.

erozygous carriers of the G1846A mutant, and one carried a minor *CYP2D6* mutation.

The 95% location interval given by COLDMAP in this analysis has a width 196 kb, slightly wider than that obtained for the cases-only analysis (186 kb). The algorithm recognizes that the majority of the control chromosomes do not bear the major mutation and excludes them from the genealogy. Thus, the location estimate is only slightly affected by the misclassified controls. For this analysis, we chose to define the major cluster of the cladogram by the longest internal branch separating the cluster from the rest of the cladogram, subject to a minimum cluster size of 10. An alternative approach is to optimize the product of internal branch length and number of leaf nodes. Using our definition, the cladogram of the 82 individuals (Fig. 4) has a single large cluster of size 44, which includes 38 of the 41 cases. Five of the six controls in the cluster are G1846A carriers, whereas the three cases not included in the cluster carry between them only two G1846A alleles. Thus, the figure illustrates again that COLDMAP is highly specific, as well as sensitive, in discriminating individuals carrying one or two copies of the major *CYP2D6* mutant allele from those bearing only the normal or minor mutant alleles. This discrimination is crucial because one of the key doubts about the efficiency of LD-based methods concerns their ability to cope with genetic heterogeneity (13).

The major cluster of Fig. 4 can be used to obtain a *P* value for the null hypothesis that no causal locus is present in the region. Under this hypothesis, cases and controls are equally likely to be represented in the cluster, and the probability of observing 38 or more cases within a cluster size of 44 is 1×10^{-13} (hypergeometric distribution). We replicated this analysis with two other random choices of 41 misclassified controls and in each case achieved genomewide significance: a major cluster size of 45 including all but five of the cases ($P = 9 \times 10^{-10}$) and a major cluster size of 42 including all but three of the cases ($P = 3 \times 10^{-15}$). The 95% location intervals from these analyses were 189 kb and 192 kb.

Finally, we applied the spectral decomposition and diversity selection procedures (14) to obtain two different 20-SNP subsets of the original 32 markers. We also investigated a strategy of removing one of any pair of SNPs found to have correlation coefficient $r^2 > 0.8$, which led to a 22-SNP subset. Fig. 5 illustrates the 95% intervals generated by COLDMAP using each of these three SNP subsets. We see that the 22-SNP subset leads to substantial loss of accuracy for fine mapping, with an increase in interval width of >50% compared with the full 32-SNP analysis. However, for the two 20-SNP subsets, the increase in interval width is more modest, averaging 18%. Analysis of cladograms (results not shown) shows that there is also some loss of accuracy in identifying clusters of chromosomes or individuals sharing the same *CYP2D6* mutations.

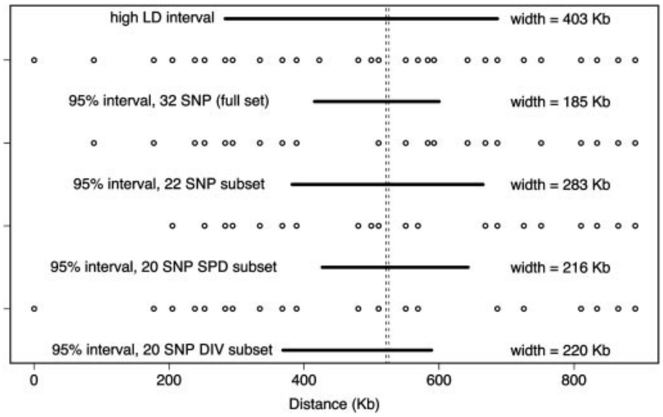

Fig. 5. The horizontal bars indicate (top to bottom) the high-LD interval (8) and 95% intervals from the analysis of all 32 SNPs, a 22-SNP subset, and two 20-SNP subsets. SPD, spectral decomposition; DIV, diversity. Each row of circles shows the locations of the SNPs used to calculate the 95% interval immediately below it.

Conclusion

We have shown that appropriate genealogical modeling can identify chromosomes bearing the major mutation and hence locate this mutation within an interval much narrower than the interval displaying LD. We thus disagree with the suggestion that the apparent block structure of LD will ". . . reduce the need for sophisticated population-genetic inference in gene mapping" (15). The block structure of LD creates the need for population-genetic inference to refine location within the LD blocks (16). Note also that our results are consistent with recombination occurring occasionally throughout this region, so the punctate nature of recombination reported in the MHC region (17) is not absolute here.

We have also shown that the SNP selection procedures described in ref. 14 work well, in that they have reduced genotyping costs by almost 40% while increasing the width of the location intervals by <20%. These procedures are based on attempting to choose subsets of SNPs that capture as much of the genetic variation as possible. In contrast, choosing a SNP subset based on pairwise LD gives a much greater loss of information (50% increase in interval width), despite including more SNPs.

The major limitation of our COLDMAP software is computational expense. In particular, exploring the possible haplotype assignments of genotype data implies a substantial computational cost. Thus, COLDMAP cannot be applied directly to large numbers of cases or to many SNPs. However, it can be used to investigate regions of interest highlighted by simpler analyses. We have shown here that significance levels can be achieved that compensate for the multiple testing involved in such an approach.

We thank Mike Barnes (GlaxoSmithKline, Stevenage, United Kingdom) for his help in preparing the SNP map. A.P.M. acknowledges financial support from the Wellcome Trust.

1. Rioux, J. D., Daly, M. J., Silverberg, M., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., *et al.* (2001) *Nat. Genet.* **29,** 223–228.
2. Daly, M. J., Rioux, J. D., Scaffner, S. F., Hudson, T. J. & Lander, E. S. (2001) *Nat. Genet.* **29,** 229–232.
3. Johnson, G. C. L., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., DiGenova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., *et al.* (2001) *Nat. Genet.* **29,** 233–237.
4. Pritchard, J. K. & Przeworski, M. (2001) *Am. J. Hum. Genet.* **69,** 1–14.
5. Ardlie, K. G., Kruglyak, L. & Seielstad, M. (2002) *Nat. Rev. Genet.* **3,** 299–309.
6. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.* (2002) *Science* **296,** 2225–2229.
7. Kruglyak, L. (1999) *Nat. Genet.* **22,** 139–144.
8. Hosking, L. K., Boyd, P. R., Xu, C.-F., Nissum, M., Cantone, K., Purvis, I. J., Khakkar, R., Barnes, M. R., Liberwirth, U., Hagen-Mann, K., *et al.* (2002) *Pharmacogenomics J.* **2,** 165–175.
9. Morris, A. P., Whittaker, J. C. & Balding, D. J. (2002) *Am. J. Hum. Genet.* **70,** 686–707.
10. Gamerman, D. (1997) *Markov Chain Monte Carlo* (Chapman & Hall, London).
11. Stephens, M., Smith, N. J. & Donnelly, P. (2001) *Am. J. Hum. Genet.* **68,** 978–989.
12. Gordon, A. D. (1999) *Classification* (Chapman & Hall, London), 2nd Ed.
13. Pennisi, E. (1998) *Science* **281,** 1787–1789.
14. Meng, Z., Zaykin, D. V., Xu, C.-F., Wagner, M. & Ehm, M. G. (2003) *Am. J. Hum. Genet.* **73,** 115–130.
15. Goldstein, D. B. (2001) *Nat. Genet.* **29,** 109–111.
16. Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F. S., *et al.* (2003) *Nat. Genet.* **33,** 382–387.
17. Jeffreys, A. J., Kauppi, L. & Neumann, R. (2001) *Nat. Genet.* **29,** 217–222.