Ann. Hum. Genet. (2002), **66**, 419–429 © *University College London*
DOI: 10.1017/S0003480002001318 *Printed in the United Kingdom*

Association analysis for quantitative traits by data mining: QHPM

P. ONKAMO[1,2], V. OLLIKAINEN[3], P. SEVON[1,4], H. T. T. TOIVONEN[4], H. MANNILA[3,5] AND J. KERE[1,2]

[1] *Karolinska Institute, Department of Biosciences at Novum, SE-14157 Huddinge, Sweden*
[2] *Finnish Genome Center, P.O. Box 63, FIN-00014 University of Helsinki, Finland*
[3] *Helsinki Institute for Information Technology, Basic Research Unit, Department of Computer Science, P.O. Box 26, FIN-00014 University of Helsinki, Finland*
[4] *Department of Computer Science, P.O. Box 26, FIN-00014 University of Helsinki, Finland*
[5] *Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FIN-02015 HUT, Finland*

(Received 17.1.02. Accepted 21.6.02)

SUMMARY

Previously, we have presented a data mining-based algorithmic approach to genetic association analysis, Haplotype Pattern Mining. We have now extended the approach with the possibility of analysing quantitative traits and utilising covariates. This is accomplished by using a linear model for measuring association. We present results with the extended version, QHPM, with simulated quantitative trait data. One data set was simulated with the population simulator package Populus, and another was obtained from GAW12. In the former, there were 2–3 underlying susceptibility genes for a trait, each with several ancestral disease mutations, and 1 or 2 environmental components. We show that QHPM is capable of finding the susceptibility loci, even when there is strong allelic heterogeneity and environmental effects in the disease models. The power of finding quantitative trait loci is dependent on the ascertainment scheme of the data: collecting the study subjects from both ends of the quantitative trait distribution is more effective than using unselected individuals or individuals ascertained based on disease status, but QHPM has good power to localize the genes even with unselected individuals. Comparison with quantitative trait TDT (QTDT) showed that QHPM has better localization accuracy when the gene effect is weak.

INTRODUCTION

Interest in association analysis and LD mapping methods for complex and quantitative traits is increasing, especially as experience has shown linkage mapping of complex trait susceptibility loci to be relatively inefficient. This inefficiency is probably due to small locus effects (locus specific λ_s below 2.0) expected to characterize many complex disease loci (Risch, 2000). Power simulations have shown that the numbers of core families or affected sib-pairs needed for mapping loci of such faint effect with linkage methods is beyond any reasonable effort (Hauser *et al.* 1996). Compared to linkage analyses, association methods should be more powerful when penetrance of the genes is low (Abecasis *et al.* 2000).

Despite the obvious need for non-linkage based approaches, serious doubts concerning the feasibility of association mapping have been put forward. The most relevant questions are (1) whether there is enough LD in the populations to make the LD-based gene mapping worthwhile, and (2) whether the statistical methodology for association analysis will be able to cope computa-

Correspondence: Päivi Onkamo, PhD, Karolinska Institute, Department of Biosciences at Novum, SE-14157 Huddinge, Sweden. Tel: +468-6083314; Fax: +46-8-7745538.
E-mail: paivi.onkamo@biosci.ki.se

tionally with the number of markers, disease models, weak gene effects and environmental effects possibly needed to find the true underlying gene effects. Concerning the strength of LD, recent studies of LD in the Caucasian populations show that the genome is composed of blocks of DNA, 20–100 kb in length, inside which the markers are in almost complete LD (Reich *et al.* 2001; Abecasis *et al.* 2001*a*; Daly *et al.* 2001). Further, as few as 2–4 most common haplotypes in a block cover 90% of a population (Daly *et al.* 2001). This block structure seems to imply that an LD-based search of disease loci would perhaps require only a few markers per block to be typed: with a few selected SNPs a block could be identified. Thus, there seems to be enough LD in short distances in order to facilitate LD-based mapping, at least in distances of a couple of hundred thousand kilobases. Currently, a haplotype map describing the block structure is under construction (Helmuth, 2001), but large-scale studies on the extent and range of LD blocks are still needed to evaluate their true usefulness in LD-based gene mapping.

Concerning the ability of computational methods to cope with the expected complexity of statistical models, and the sheer amount of data, it seems that development of methodology is clearly warranted. Several new approaches, both for dichotomous and for quantitative trait association analysis/LD mapping, have been published during recent years. Many of the newer tests are based on TDT (Spielman *et al.* 1993), and thus use family or family trio data. A general test of association for quantitative traits in nuclear families (QTDT, Abecasis *et al.* 2000), including test statistics proposed by many other researchers (Allison, 1997; Rabinowitz, 1997; Fulker *et al.* 1999), is a TDT-based approach, in which the quantitative response is modelled by variance component methods. Contrary to family-based tests, an approach using population samples consisting of unrelated individuals has recently been published by, for example, Zhang & Zhao (2001). Rannala & Reeve (2001) present an interesting approach that uses MCMC methods for the multipoint linkage disequilibrium mapping of a susceptibility gene. Some of the new methods are relevant for bi-allelic markers only (Zhang & Zhao, 2001), and some are relevant to all types of markers (Rannala & Reeve, 2001). Also, attempts to utilize neural networks have been presented (Curtis *et al.* 2001).

Previously, we introduced a data mining inspired algorithmic approach, Haplotype Pattern Mining (HPM), for genetic association analysis of binary traits (Toivonen *et al.* 2000; Sevon *et al.* 2001). HPM utilises linkage disequilibrium between close genetic markers in relatively densely mapped data: all trait-associated haplotype patterns, potentially with small gaps, are searched from the data. The strength of association of the patterns is measured by a suitable statistic, such as a simple χ^2 test, and a scoring function is used for combining the information about strongly associated patterns into a prediction of a susceptibility gene location. We showed that with such an association method, one is able to find disease genes with moderately small sample sizes and low gene effects: for instance, with $\lambda_s = 1.7$, sample size 200 affected individuals and 200 controls, the probability of finding the right genetic area (prediction error less than 4 cM) is still greater than 80% (Toivonen *et al.* 2000).

In this paper, we extend the HPM method to utilise information from quantitative traits, either as a response variable or covariates. This is accomplished simply by measuring the strength of association with a linear model. Results from power analysis with simulated data are presented, with comparison to QTDT.

METHODS

The QHPM analysis for quantitative traits is carried out as follows. We assume that either (a) family trios with genotypes or (b) case-control data with haplotypes are available. The family trios are haplotyped, and the trait and covariate values of the offspring are assigned to the transmitted haplotypes of a trio, whereas the values of the parent are assigned to the corresponding non-transmitted haplotype. Now, all haplotype

patterns that occur at least a specified minimum number of times are searched for: given a marker map M with k markers $m_1, \ldots, m_k$, a *haplotype pattern* P on M is defined as a vector $(p_1, \ldots, p_k)$, where each p_i is either an allele of m_i or the 'don't care' symbol ($*$). The haplotype pattern P occurs in a given haplotype vector (chromosome) $H = (h_1, \ldots, h_k)$ if $p_i = h_i$ or $p_i = *$ for all i, $1 \leqslant i \leqslant k$. The patterns are allowed to include gaps ($*$), in order to account for missing and erroneous data (Toivonen *et al.* 2000). For each pattern we fit a linear model,

$$Y_j = \alpha + \beta I_{pj} + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_n X_{nj} + \epsilon_j,$$

where Y_j is the (quantitative) trait value for individual j, I_{pj} is the indicator variable for the occurrence of the haplotype pattern P in a chromosome of individual j, X_{nj} is the value for nth explanatory variable for individual j, and ϵ_j is the error term. The explanatory variables, or covariates, might be environmental factors, sex, age at examination, etc. In this model the trait is normal and the effects of covariates linear. The significance of a pattern as a covariate was obtained from a t-test comparing the model to the best fitting model in which the corresponding coefficient is zero. These nominal significances (p values) form the basis for the scoring function for markers used in QHPM. An example how the method works is given in Figure 1.

Scoring function

For each marker in turn, all haplotype patterns that include that marker are considered, the number of which is denoted by s in the following. Informally, the markers in which the overlapping haplotype patterns show strong association to the phenotype are those of most interest. The distribution of observed nominal p values of all haplotype patterns overlapping a marker are compared to a uniform distribution between 0 and 1. Uniform distribution would be expected for mutually independent patterns not associated with the trait, under the null hypothesis of no trait association. We acknowledge that the patterns we observe and their p values are not mutually independent, but the uniform distribution is a useful approximation for the expected distribution. As a measure of the distance between the observed distribution of p values and the uniform distribution the following heuristic scoring function was used: Let t_r be the rth p value in the sorted list of s observed p values for a given marker, and q_r the expectation of the rth p value $(r/(s+1))$, if s p values were randomly picked from the uniform distribution. The score was defined as the mean of the distances $(t_r - q_r)\log(t_r/q_r)$, which is an *ad hoc* statistic that was proven to perform very well in simulation experiments. This measure yields larger distances when the observed distribution is skewed towards lower p values. In the experiments, the disease gene is predicted to be at the marker with the largest distance measure.

In addition to the distance described above, the Kolmogorov–Smirnov goodness-of-fit test (comparing the observed and expected distributions of the p values) was tested for scoring function. However, Kolmogorov–Smirnov did not work even nearly as well as the heuristic distance measure, which was therefore used for all analyses.

For evaluation of empirical significance levels of the scores, we propose permutation tests to be carried out, permuting randomly the trait values associated with the chromosomes.

QTDT

QTDT version 2.2.1 was used for the comparative analyses of power with quantitative responses. QTDT is a variance component model approach, where all available offspring can be included in the analysis, with or without parental information. The test is not biased on the presence of linkage or familiality. The approach is suitable for dense maps. For each of our own simulated data sets, the parents of the 200 affected individuals were included, yielding in total 1200 chromosomes per set. The test was performed as described in Abecasis *et al.* (2000). The prediction of the location of the disease gene for each data set was made to the marker with the highest

A) Haplotype data:

	M1	M2	M3	M4	M5	M6	M7	M8
H1	1	2	2	3	3	2	4	1
H2	1	2	2	4	2	2	1	2
H3	1	2	2	4	2	1	1	1
H4	2	3	2	3	3	2	4	2
H5	2	1	1	1	1	2	4	2
H6	1	1	2	3	3	2	1	1
...								

B) Haplotype patterns:

	M1	M2	M3	M4	M5	M6	M7	M8	Freq.
P1	1	2	2	*	*	*	*	*	20
P2	1	2	2	4	2	*	*	*	16
P3	*	*	2	3	3	2	*	*	15
P4	*	*	*	*	*	2	4	2	14
P5	*	*	*	*	1	2	4	2	14
P6	*	*	*	3	3	2	4	2	13
...									

C) Phenotype data:

	Y	X	P1	P2	P3	P4	P5	P6	...
H1	1.5	2.0	1	0	1	0	0	0	
H2	1.5	2.0	1	1	0	0	0	0	
H3	2.0	2.8	1	1	0	0	0	0	
H4	0.4	2.5	0	0	1	1	0	1	
H5	0.2	2.0	0	0	0	1	1	0	
H6	0.2	2.0	0	0	1	0	0	0	
...									

D) Linear regression:

Pattern	*P* value
P1	0.023
P2	0.042
P3	0.178
P4	0.621
P5	0.448
P6	0.778
...	

E) Scoring (marker 3)

Observed(t)	Expected(q)
0.023	0.250
0.042	0.500
0.178	0.750

$$\text{Score} = \frac{1}{3}\sum_{r=1}^{3}(t_r - q_r)\ln\frac{t_r}{q_r} \approx 0.833$$

Fig. 1. An example of the QHPM method. (A, B) First, all haplotype patterns exceeding a given frequency threshold are searched for from haplotype data, where markers are denoted by M and chromosomes with H. (C) Then, for each frequent pattern (P) in turn, a linear model is fitted to the data containing the response variable Y and, in the example, a single covariate X and the dummy variable for the pattern. (D) p value for each pattern is yielded by t-test. (E) The markers are scored by our heuristic distance measure, which compares the p value distribution of the patterns overlapping the marker under consideration to the expected uniform distribution under the null hypothesis.

value of F-test statistic (lowest uncorrected p value).

RESULTS

Simulated data sets

We simulated data in order to evaluate the performance of the extended HPM, and to compare it with the QTDT method. The simulations were carried out with the population simulator Populus (Ollikainen, 2002). The simulated data corresponds to a moderately sized, isolated population, which has grown from 100 founder individuals to 100000 in 20 generations. The genetic length of the simulated chromosomes was 100 cM for both males and females. Within this region, the disease locus was randomly selected, and 6 founder mutations were randomly assigned to the

initial population, all of which were then associated to a different founder haplotype. No chiasma interference was modelled. The simulated microsatellite markers had 4 alleles with frequencies of 0.4, 0.2, 0.2, and 0.2 in the founder population. The markers were spaced 1 cM apart. We computed the liabilities to the disease for each individual using two alternative models:

$$DM_1 = 2x_g + x_{e1} + x_{e2} + x_r + C_1,$$

and

$$DM_2 = 5x_g + x_{e1} + x_{e2} + x_r + C_2,$$

where x_g is an indicator variable for the presence of at least one of the disease-predisposing mutations, yielding a dominant disease model with reduced penetrance. Variables x_{e1} and x_{e2} are simulated environmental components affecting the liability, which are observed for each individual in analysis, and x_r is an unobserved random component, all of which follow a standard normal distribution $N(0,1)$. Constants C_1 and C_2 represent the baseline liability, and they are adjusted in an extra pre-sampling phase to make the prevalence of the disease as close to the target value of 5% as possible. When the liability of an individual has been computed, the disease statuses are statistically defined: an individual's probability of being affected is obtained from formula

$$\log \frac{p}{1-p} = DM_i,$$

where $i = 1$ for model DM_1 and $i = 2$ for model DM_2.

For quantitative analysis, five traits, Q_1–Q_5, were simulated. The value for each trait Q_j, j denoting the strength of genetic effect, was computed from the formula

$$Q_j = jx_g + x_{e1} + x_{e2} + r,$$

where x_g, x_{e1}, and x_{e2} are the genetic and environmental liability components described above, and r is a random value between zero and unity. Since the trait values differ only by the coefficient of the genetic component, equation $Q_1 = \cdots = Q_5$ holds when no disease gene is present, whereas equation $Q_j = Q_{j-1} + 1$ holds ($j \in \{2, \ldots, 5\}$) when the individual has inherited a disease-predisposing mutation.

The sampling from the simulated population was done on the basis of affection status for both quantitative traits and the disease status: 200 independent trios with an affected offspring were randomly sampled. For the analysis of quantitative traits no further sampling based on values of quantitative traits was done. This ascertainment scheme closely resembles real studies in the sense that often data are collected through an affected proband, and there are correlated quantitative traits which have been measured and need to be analysed in the data.

The effect of fixed prevalence is that there are more carriers of the liability alleles in a sample of affected individuals given by DM_2 than those created by DM_1. Thus, the affection status as well as the quantitative traits will be more 'genetically' determined in the group DM_2 compared to DM_1, rendering the data that have been simulated under the model DM_2 easier for gene mapping purposes.

GAW12 data

In addition to our own simulated data, GAW12 data (Almasy *et al.* 2001) was used, in order to compare the performance of the QHPM method with different simulated data sets, and to ensure its robustness against varying population and disease model parameters. GAW12 data consist of 50 replicates of a simulated isolated population. Each replicate includes 1497 individuals in 23 extended pedigrees. For each individual, five quantitative traits, Q_1–Q_5, and affection status have been measured. The genetic background for the traits is complex, involving five major genes for Q_1–Q_5, and one for affection status, with complex gene–gene and gene–environment interactions. In total 2855 marker genotypes, spanning the whole genome (22 chromosomes), were given per each individual alive. Marker mean heterozygosity was 0.81, and the average marker spacing 1 cM. Because QHPM uses haplotype data as input, we designed a

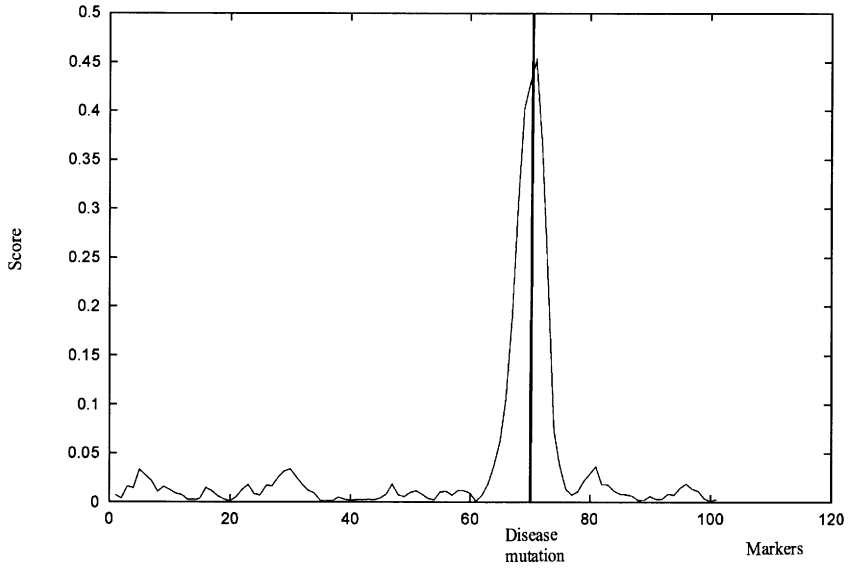

Fig. 2. An example of a successful disease gene localization. The vertical line shows the true location of the disease gene (70.36 cM from the start of the region). The curve corresponds to the marker-wise scores.

simple algorithm for ascertaining independent family trios from the pedigrees, based on (1) affection status of the offspring or the value of the quantitative trait being studied in the offspring, and (2) the genotype information availability in the trios. The quantitative trait values were chosen either from (I) the upper 10% tail of the quantitative trait distribution, and the control individuals from the lower 10% end of distribution or (II) the upper quartile and lower quartile. Only one trio per pedigree was chosen, to ensure the independency of the trios. The choices were made in 25 replicates in order to achieve a reasonable number of affected individuals from the data. The actual number of individuals sampled was a little less than 600, depending on the trait. The family trios were haplotyped by an algorithm defining the allele phases for each trio. Alternatively, haplotypes can be obtained by, e.g., using the haplotyping feature of the GENEHUNTER (Kruglyak *et al.* 1996) software package.

The GAW12 data differs from our own simulated data in the following: (1) the GAW12 isolated populations are smaller; (2) the disease model is more complicated with dichotomous disease status affected by one major gene, 5 quantitative traits, and a household effect; (3) high (25%) disease prevalence; (4) each quantitative trait has a complex background with both genetic and environmental effects, as well as interactions affecting them; (5) age-dependent penetrance for affection status.

QHPM parameters

The QHPM analyses were made mostly using one set of parameter values: the maximum length of the haplotype patterns to search for was set to 7 markers, and the maximum number of gaps per pattern to 1, with the maximum gap length being 1 marker. The minimum number of occurrences of a pattern was 10 (frequency threshold), to exclude patterns for which significant association could not be obtained. Experiments with other parameter settings, analysing dichotomous response variable, have been described in our previous paper (Toivonen *et al.* 2000), in which it was shown that the method is robust to different choices of parameters.

Localization accuracy

To illustrate the QHPM method, a typical example with correct disease gene localization is shown in Figure 2. True vs. predicted locations

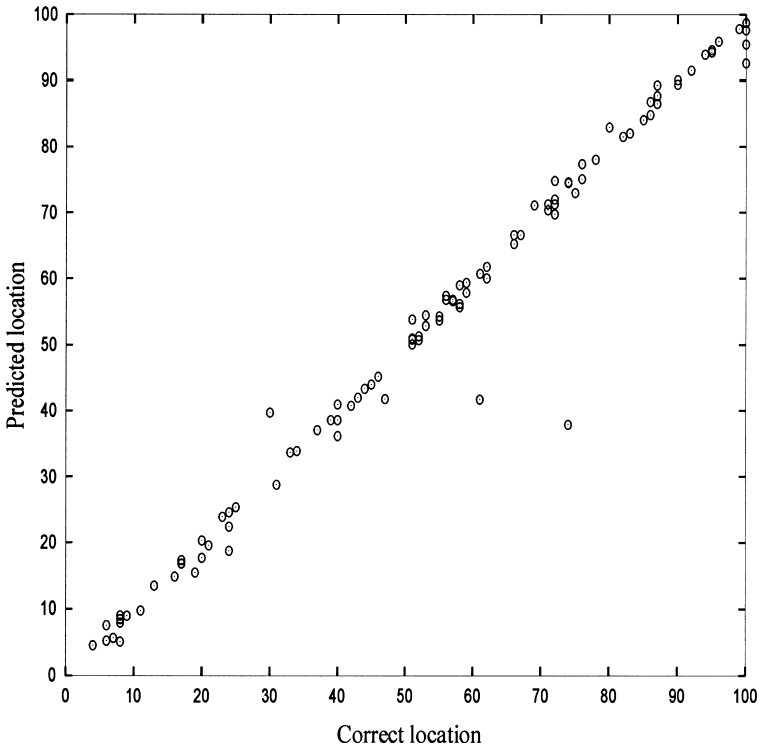

Fig. 3. Predicted vs. correct location of the disease gene in 100 simulated data sets for the trait with the strongest genetic determinant (Q5) and the disease model with a strong genetic effect (DM_2). The sample of 200 trios had been ascertained based on affectation status alone.

for 100 simulated data sets with simulation parameters as described above, with the quantitative trait model DM_2, trait Q_5, are shown in Figure 3.

The power of the QHPM was analysed with 100 replications of data in all simulation settings: both models DM_1 and DM_2, with all five quantitative traits Q_1–Q_5 each. The power is illustrated by the cumulative percentage of data sets in which localization error was the same or less than that given from the x-axis (Fig. 4A, C). Clearly, the simulated data varied from practically impossible to very easy for this method. Comparisons with QTDT (Fig. 4B, D) clearly show that QHPM has a slightly better localization accuracy than QTDT. The effect was seen with both disease models DM_1 and DM_2, and is pronounced with the more difficult quantitative traits, Q_2 and Q_3, but not with Q_1, which seems to have been too difficult for both approaches (Fig. 4).

Next, the quantitative traits were dichotomized and QHPM was compared to the original HPM approach. The comparison was done with our own simulated data sets, with model DM_1 ('difficult' model) and all five quantitative traits. The dichotomization was made by rearranging the data with respect to the values of the quantitative trait in question, and dividing the arranged data set into two parts of equal size. The half with the lower values is labelled as 'controls', and that with higher values as 'cases', to keep the sample size the same for both approaches. HPM was run with the following parameter settings: threshold for the strength of association was set to $\chi^2 = 6$; maximum pattern length 7; and maximum number and maximum length of gaps was 1 (for parameter settings, see Toivonen *et al.* 2000). The analysis on dichotomized variables, compared with quantitative analysis by QHPM (Fig. 5), revealed that the probability of correct prediction is very similar with both methods when there is sufficiently high genetic effect for the traits (in our example, traits Q_3–Q_5). However, when the genetic control of the trait decreases (traits Q_1 and Q_2), the advantage of the genuinely quantitative analysis becomes clear, as

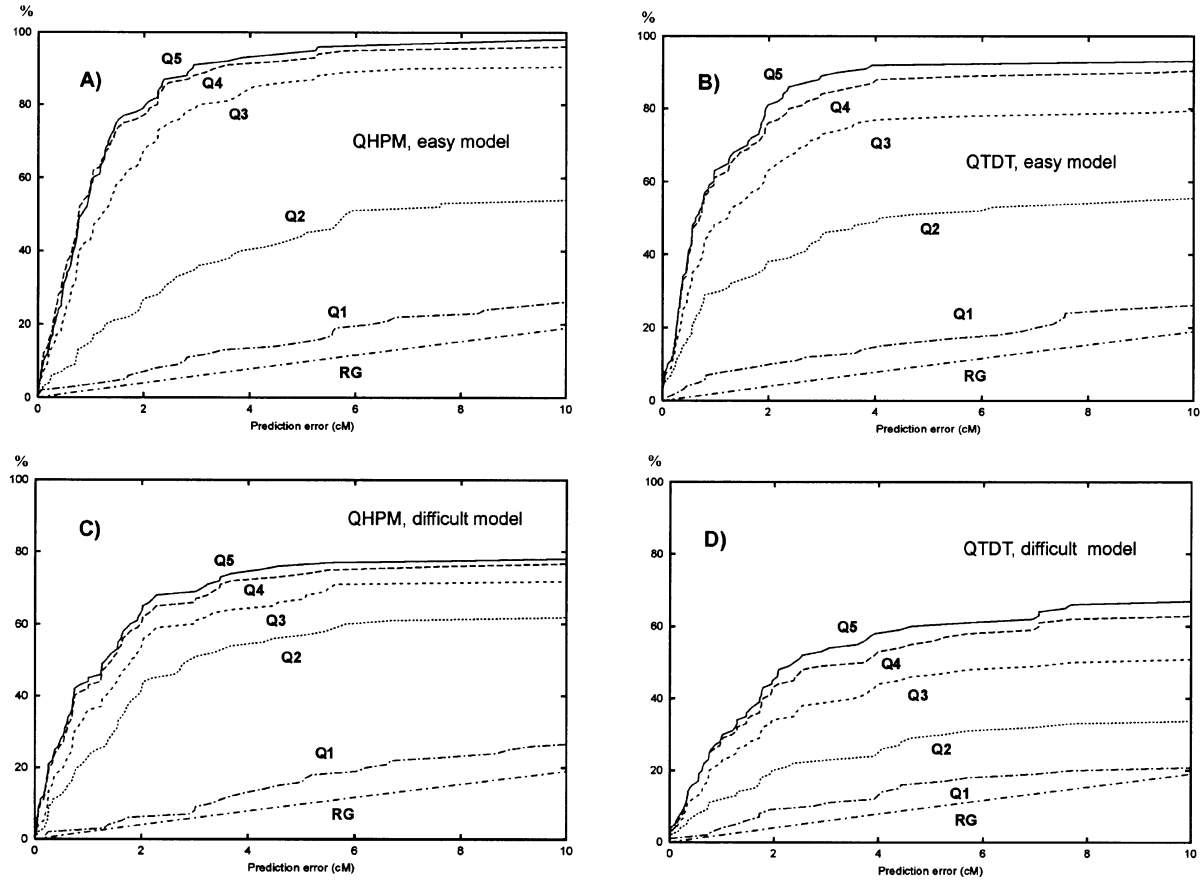

Fig. 4. (a) Probability of correct disease gene localization with QHPM as a function of tolerated localization error (x-axis). We used the disease model DM_2 with a strong genetic effect. Label RG corresponds to random guessing. (b) Same as previous, but for QTDT. (c) Like (a), but utilising disease model DM_1 with a weaker genetic effect than DM_2. (d) Same as previous, but for QTDT.

the probability of correct prediction gets higher using QHPM than HPM.

Analyses with GAW12 data

Ascertainment from the tails of the distribution

The results of QHPM genome scans on the disease status and all quantitative traits Q_1–Q_5 have been published in Sevon *et al.* (2001). Using approximately 600 cases and 600 controls per phenotype, we were able to correctly localize the susceptibility genes. In this paper, the emphasis is on the effect of (1) differing ascertainment schemes and (2) number of cases and controls on the signal of the susceptibility gene when such a gene is known to exist.

The quantitative trait Q_5 was chosen for testing the effect of the ascertainment scheme. First, QHPM was run for a data set in which the 'cases' were chosen from the upper and lower quartile of the trait distribution. Then, the data were dichotomized and HPM was run with these data. We also used a more stringent ascertainment scheme, where cases were sampled from the upper 10% and lower 10% tail of the Q_5 distribution, and made analyses with these data sets both with QHPM and, after dichotomization of the data, with HPM. With the latter, more extreme data set, both methods could localize the correct susceptibility gene. However, with the less extreme sampling scheme, QHPM could correctly pinpoint the susceptibility locus, whereas HPM could not.

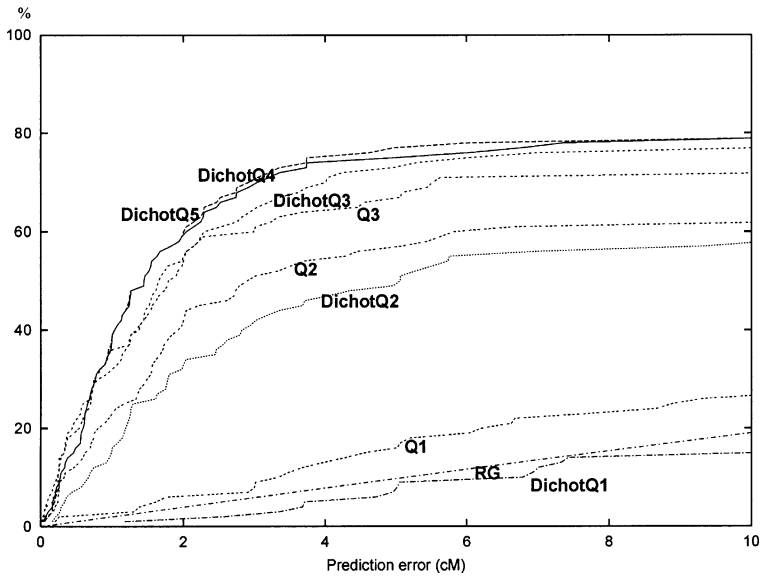

Fig. 5. Performance of QHPM using continuous (Q1–Q3) vs. dichotomised (DichotQ1–DichotQ3) variables. The probability of correct disease gene localisation is shown as a function of tolerated localisation error (x-axis) for the disease model with a weak genetic effect (DM_1). Label RG corresponds to random guessing, again. Curves for traits Q4 and Q5 are omitted, as they closely resemble the upper curves in the figure.

Sample sizes were kept constant with both ascertainment criteria.

Effect of sample size

Given that the genome scans were done on reasonably large data sets, which are not often available in reality, we also tested the effect of differing sample sizes with the GAW12 data using variables Q1 and Q4. We chose chromosomes 2 and 19 for trait Q1, and 9 and 17 for Q4, each of which had been simulated to harbour one susceptibility gene for the above-mentioned traits. The sample sizes were 100, 200, 300 or 600 cases and the same number of controls; the analyses were made with both QHPM and HPM. With $n = 100$ none of the four susceptibility genes was pinpointed; with $n = 200$, 2/4 were localized with HPM and 2/4 with QHPM. 300 cases and controls resulted in only a slightly better performance: 3/4 with QHPM and 2/4 with HPM (data not shown).

DISCUSSION

We have shown that a data mining based approach, combined with a suitable measure of statistical association of a quantitative trait, can produce good localization accuracy even with small data sets. We have also shown that the method is more efficient than the existing 'state-of-the-art' method, QTDT, thus giving improved possibilities for trait gene localization. This probably reflects the ability of our method to take account of all associating patterns in an area, and not just one haplotype.

For complex traits, it is expected that using quantitative scores gives much more information than simple disease status, i.e., the disease status may essentially be just a combination of different symptoms and quantitative measurement scores which has been agreed upon as the clinical diagnosis (for example, as in the definition of rheumatoid diseases). Genetically, such a trait is quite artificial: there may not exist any simple genetic basis for such an entity. Thus, analysis of the quantitative traits on which the diagnosis is based might improve the ability to map genes behind the complex traits. In this paper we showed that in contrast to analyzing such a dichotomized quantitative trait, a truly quantitative analysis offers power gains especially when gene effects are low compared to environmental effects. If dichotomization is carried

out, the choice of cut-off points for dichotomization affects the efficiency of finding genes: if cases and controls can be chosen from the tails of the distribution, the analysis of dichotomized data may become as powerful as real quantitative analysis (based on our experiments with GAW12 data). This is closely related to the use of selection strategies in linkage studies.

In order to maximize the probability of finding true disease associations, one should try to maximize the genotype–phenotype correlation in the study population. To this end, several different selection strategies have been developed, especially for quantitative trait linkage studies. The use of extreme phenotypes has been shown to greatly improve the power to detect QTLs (Lander & Botstein, 1989; Risch & Zhang, 1995, 1996; Zhang & Risch, 1996). Selection strategies include extended family collection and sib-pair designs: concordant, discordant, extreme discordant, and single-selection (in which a sibship is ascertained if one offspring has an extreme phenotype). The same selection strategies are not necessarily optimal for association studies. Usually, the data sets have actually been ascertained for a linkage study, and thus the most relevant problem is probably the lack of power to detect association in data ascertained for other purposes. Effects of linkage analysis selection strategy on the power of an association study of quantitative variables have been studied in Abecasis *et al.* (2001*b*). There, the power to find a QTL with QTDT was best if a discordant sib-pairs or extreme-proband design had been applied. If single selection had been used, the power depended more on allele frequencies of both marker and trait alleles. Another comparison of linkage vs. association methods in the framework of variance component modelling showed that for both the power was mostly dependent on the proportion of phenotypic variance attributable to the QTL (Sham *et al.* 2000). The main difference between the two was that the power declined more rapidly for linkage, as the QTL heritability decreased.

The second important result shown in this paper concerns the effect of the original ascertainment scheme on the efficiency of the QHPM method. We assumed that the sampling has been carried out based on disease status, and other characteristics have been measured as surrogates, or out of general interest. Thus, the distribution of these additional variables depends on the correlation between the original variable used for ascertainment and the quantitative trait in question. This might make the use of such a trait less efficient for gene mapping in which a particular quantitative trait distribution (usually normal) is assumed. However, here we showed that even when the data were indeed ascertained based on a simulated affection status with which the quantitative traits were correlated, the quantitative trait genes could still be localized.

The method of assessment of genetic and environmental variables could be, for example, a variance component (VC) model instead of a linear model, as these are statistically more refined and flexible. VC models allow for more complex (realistic) disease models, i.e. different genes and covariates affect the outcome. They would enable the estimation of allele effects, and the estimation of proportion of variance attributable to different components.

Finally, covariates could be utilized more efficiently: instead of looking at one variable at a time, it might be more interesting to try to find the best combinations of different variables for which there would be reason to believe there to be a common, and strong, genetic basis: 'mining' for the best combinations could be useful.

References

Abecasis, G. R., Cardon, L. R. & Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292.

Abecasis, G. R., Cookson, W. O. C. & Cardon, L. R. (2001*b*). The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am. J. Hum. Genet.* **68**, 1463–1474.

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., Carey, A., Cardon, L. R., Moffatt, M. F. & Cookson, W. O. C. (2001*a*). Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197.

Allison, D. B. (1997). Transmission disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690.

Almasy, L., Terwilliger, J. D., Nielsen, D., Dyer, T. D., Zaykin, D. & Blangero, J. (2001). GAW12: Simulated genome scan, sequence, and family data for a common disease. In *Analysis of complex genetic traits: Applications to asthma and simulated data* (eds E. M. Wijsman, L. Almasy, C. I. Amos, I. Borecki, C. T. Falk, T. M. King, M. M. Martinez, D. Meyers, R. Neuman, J. M. Olson, S. Rich, M. A. Spence, D. C. Thomas, V. J. Vieland, J. S. Witte & J. W. MacCluer). *Genet. Epidemiol.* **21** (Suppl. 1), S332–S338.

Curtis, D., North, B. V. & Sham, P. C. (2001). Use of an artificial neural network to detect association between a disease and multiple marker genotypes. *Ann. Hum. Genet.* **65**, 95–107.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232.

Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267.

Hauser, E. R., Boehnke, M., Guo, W. W. & Risch, N. (1996). Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. *Genet. Epidemiol.* **13**, 117–137.

Helmuth, L. (2001). Genome research: map of the human genome 3.0. *Science* **293**, 583–585.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996). Parametric and non-parametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363.

Lander, E. S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–99.

Ollikainen, V. Simulation techniques for disease gene localization in isolated populations. Academic dissertation. University of Helsinki, Dept. of Computer Science. In: Series of Publications A, Report A-2002-2.

Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum. Hered.* **47**, 342–350.

Rannala, B. & Reeve, J. P. ((2001). High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.* **69**, 159–178.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R. & Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856.

Risch, N. & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589.

Risch, N. J. & Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *Am. J. Hum. Genet.* **58**, 836–843.

Sevon, P., Ollikainen, V., Onkamo, P., Toivonen, H. T. T., Mannila, H. & Kere, J. (2001). Mining associations between genetic markers, phenotypes, and covariates. In *Analysis of complex genetic traits: Applications to asthma and simulated data* (eds. E. M. Wijsman, L. Almasy, C. I. Amos, I. Borecki, C. T. Falk, T. M. King, M. M. Martinez, D. Meyers, R. Neuman, J. M. Olson, S. Rich, M. A. Spence, D. C. Thomas, V. J. Vieland, J. S. Witte & J. W. MacCluer). *Genet. Epidemiol.* **21** (Suppl. 1), S588–S593.

Sham, P. C., Cherny, S. S., Purcell, S. & Hewitt, J. K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-component models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616–1630.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516.

Toivonen, H. T. T., Onkamo, P., Vasko, K., Ollikainen, V., Sevon, P., Mannila, H., Herr, M. & Kere, J. (2000). Data mining applied to linkage disequilibrium mapping. *Am. J. Hum. Genet.* **67**, 133–145.

Zhang, H. & Risch, N. J. (1996). Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. *Am. J. Hum. Genet.* **59**, 951–957.

Zhang, S. & Zhao, H. (2001). Quantitative similarity-based association tests using population samples. *Am. J. Hum. Genet.* **69**, 601–614.