An empirical comparison of case-control and trio based study designs in high throughput association mapping

P Hintsanen, P Sevon, P Onkamo, L Eronen and H Toivonen

J. Med. Genet. 2006;43;617-624; originally published online 28 Oct 2005;
doi:10.1136/jmg.2005.036020

LETTER TO JMG

An empirical comparison of case-control and trio based study designs in high throughput association mapping

P Hintsanen, P Sevon, P Onkamo, L Eronen, H Toivonen

J Med Genet 2006;**43**:617–624. doi: 10.1136/jmg.2005.036020

Motivated by high throughput genotyping technology, our aim in this study was to experimentally compare the power and accuracy of case-control and family trio based approaches for haplotype based, large scale, association gene mapping. We compared trio based and case-control study designs in different disease models, and partitioned the performance differences into separate components: those from the sample ascertainment, the effective sample size, and the haplotyping approaches. For systematic and controlled tests, we simulated a rapidly expanding and relatively young isolated population. The experiments were also replicated with real asthma data. We used computationally efficient methods that scale up to large amounts of both markers and individuals. Mapping is based on a haplotype association test for haplotypes of 1–10 markers. For population based haplotype reconstruction, we use HaploRec, and compare it to both a simple trio based inference and true haplotypes. Firstly and surprisingly, statistically inferred population based haplotypes can be equally powerful as true haplotypes. Secondly, as expected, the effective sample size has a clear effect on both gene detection power and mapping accuracy. Thirdly, the sample ascertainment method does not have much effect on mapping accuracy. Finally, an interesting side result is that the simple haplotype association test clearly outperformed exhaustive allelic transmission disequilibrium tests. The results suggest that the case-control design is a powerful alternative to the more laborious family based ascertainment approach, especially for large datasets, and wherever population stratification can be controlled.

Gene mapping efforts have been criticised for their modest success at finding, and replicating findings of, disease susceptibility genes. While linkage methods are popular in trying to elucidate the genetic basis of complex traits, they have inherent limitations in detecting genes of modest to moderate effect.[1] Association analysis using haplotypes is a good alternative; it has better resolution, and thus also the capability to better utilise high density genotype data produced by high throughput genotyping techniques. Further gene mapping studies that only use triad or case-control data from epidemiological cohorts greatly reduce the amount of effort required to obtain the DNA samples, compared with linkage analyses based on large families.

Haplotype analyses are of increasing interest to a wide variety of investigators, and thus raise the issue of the most relevant study settings, including ascertainment scenarios and computational analysis methods, to be used in practice. In addition, understanding differences between family based and population based methods is useful for those planning new genetic studies. The goal of our study was to shed light on these issues.

We present the results of a simulation study, in which we experimentally compared two strategies for study design and haplotyping in association analysis: (*a*) a trio based setting, where trios are ascertained using affected children as probands, and the non-transmitted haplotypes of the parents are used as controls or additional cases, depending on the phenotype of the parent; and (*b*) a case-control study design, where cases are ascertained as above but independent healthy control individuals are sampled from the population, and the haplotypes are estimated for all individuals with a population based statistical method. The methods we used are computationally efficient and are therefore a valid alternative to the analysis of large datasets. The simulated population was an exponentially and rapidly growing founder population, similar to those successfully used in gene mapping studies.[2]

According to our results, population based reconstruction of haplotypes in case-control datasets combined with association analysis can clearly outperform the trio based approach, even if the haplotypes from all trios could be solved unambiguously. We experimentally analysed the effects of adjusting sample sizes for equal genotyping costs, of haplotyping, and of the sample ascertainment. An unexpected result was obtained for the effect of population based haplotypes; they were found to be very powerful in association studies, and in our experiments virtually equally good as the true (unknown) haplotypes. The results were confirmed with a real dataset used for localisation of an asthma gene.[3] Based on these results, we claim that case-control study designs may serve in general as powerful starting points for genetic association analyses, without a requirement to genotype the families. Furthermore, we suggest that the relatively simple but efficient haplotype association analysis can be sufficiently powerful in high throughput analysis.

METHODS

We used simulated datasets to carry out the study, as they allow power analysis using a large number of replicate datasets with controlled values of parameters. Different sampling and haplotyping methods and eventually a haplotype association mapping algorithm were then applied to each of these replicates in turn. The power to detect the disease susceptibility gene and the mapping accuracy were then analysed. By controlling each of the parameters separately, we were able to analyse their effects in isolation. A recent asthma dataset was used to confirm the results with real data.

Abbreviations: EATDT, exhaustive allelic transmission disequilibrium tests; SNP, single nucleotide polymorphism

Table 1 Tested alternatives of sample ascertainment, haplotyping methods, and total sample sizes (for example, 500 means 250 cases + 250 controls)

	Sample ascertainment and effective sample sizes					
	Case-control		Case-random		Trios	
	Sample size		Sample size		Sample size	
Haplotyping method	500	334	500	334	500 (750*)	334 (500*)
Population based estimation	+	−	−	−	NA	NA
Trio based inference (simple)	−	−	−	−	−	+
True haplotypes (best case)	−	−	−	−	−	+
Randomly phased haplotypes	−	−	−	−	−	−
Maximally wrong phases	−	−	−	−	−	−

*Number of genotyped individuals. +, Primary alternatives to be compared; −, settings used to analyse effects of different factors, some of them unrealistic. NA, not available

The compared settings differed in three major aspects (table 1), as follows.

1. Sample ascertainment: in all our sampling designs, an affected child is used as a proband. In the trio design, the non-transmitted (pseudo) haplotypes of parents are used as controls or additional cases, according to the phenotype of the parent. In the case-control design, healthy individuals are sampled from the population; in the case-random design, random individuals are sampled regardless of their disease status.
2. Haplotyping: haplotypes can be partially deduced when trios are available. We considered two extreme options: (*a*) a simple method, where all ambiguous alleles are marked unknown; and (*b*) the true haplotypes (known from the simulations), as the best possible case where all ambiguous alleles are successfully estimated. In the case-control and case-random designs, a statistical or combinatorial estimation method must be used. We also considered randomly phased haplotypes, and maximally wrongly phased haplotypes for comparison.
3. Sample size: in our study, the default sample size was 500 individuals. In the case-control and case-random designs, we used 250 cases and 250 control or random individuals, whose haplotypes were then estimated. In the trio design we could afford to use 167 trios—that is, 501 individuals, from which we obtained haplotypes for the 167 probands plus 167 pseudohaplotype individuals from the non-transmitted haplotypes, 334 individuals in total.

For association analysis, we used haplotype association with all haplotypes between 1 and 10 markers in length. This simplicity was a deliberate choice; the method easily scales up to very large datasets. While it is simple, it is also quite powerful, as will be shown by comparisons with exhaustive allelic transmission disequilibrium tests (EATDT).[4] A comparison of different association analysis methods is outside the scope of this paper; we aimed specifically to compare study designs and the related haplotyping approaches.

Simulation

We used a two phase simulation procedure to mimic a true founder population. In the first phase, the founder haplotypes were simulated using a coalescent model with a recombination and infinite sites mutation model, to produce a realistic polymorphism structure and realistic allele frequency distributions for the founder chromosomes. In the second phase, the final population was simulated using forward in time simulation to obtain a larger population with a realistic recombination history (see appendix for more details).

The population model we used is an exponentially and rapidly growing founder population, which starts with 100 founder individuals randomly chosen from the coalescence simulation. The final size, 100 000 individuals, is reached in 15 generations. This approximately corresponds to a recently founded subpopulation living in isolation, such as Kainuu region in northeastern Finland.[2]

The marker map in our study consists of single nucleotide polymorphisms (SNPs) separated by approximately 33 kb from each other, corresponding to the average density of a genome wide 100 kb microarray SNP chip. We simulated 451 markers—that is, a stretch of 15 000 kb (15 cM). The SNPs were chosen to have a minor allele frequency of at least 0.05. As a result of picking markers based on these two criteria, the average minor allele frequency was 0.21 and the average distance between markers was 33 kb (with 2.63 kb SD on average).

Three disease models were designed to resemble interesting and challenging cases (table 2). ''Common'' is a common disease variant, with susceptibility allele frequency of 20% in the population and low penetrance (20%), created to correspond to a typical common complex disease, such as asthma or diabetes in human populations. ''Rare'' is a model with low susceptibility allele frequency (1%) and low prevalence (5%) in the population, and with 40% penetrance (in accordance to the rare variant, rare disease hypothesis). This corresponds to a typical inherited disease studied since the late 1990s. Finally, a disease model called ''intermediate'' with susceptibility allele frequency between those of the rare and common models was created (10%), with intermediate penetrance (30%). Other disease model parameters were set so that the proportions of phenocopies were 28% (common model), 58% (intermediate), or 78% (rare). All these models result in relatively difficult mapping problems where differences between methods and approaches can be more easily observed than with easier models.

After diagnosing the final simulated generation using the disease model, we ascertained samples of cases, random individuals, and healthy controls (corresponding to the columns of table 1). Siblings were not allowed in samples.

Table 2 Disease model parameters

	Disease model		
Parameter	Common	Inter	Rare
Susceptibility allele frequency Pr(M)	0.20	0.10	0.01
Carrier frequency in population Pr(M^+)	0.36	0.19	0.02
Penetrance Pr($D^+ \mid M^+$)	0.20	0.30	0.40
Prevalence Pr(D^+)	0.10	0.14	0.05
Phenocopies Pr($M^- \mid D^+$)	0.28	0.58	0.78
Penetrance for non-carriers Pr($D^+ \mid M^-$)	0.04	0.10	0.04

Inter, intermediate. D^+, affected; M, disease susceptibility allele; M^+, susceptibility allele carrier (genotype MM or MW, where W is a wild type allele); M^-, susceptibility allele non-carrier (genotype WW).

In order to minimise the random effects caused by random sampling of individuals, the overlap of the samples for different strategies was maximised for a given replicate: the set of cases was identical in all the strategies, and the healthy individuals of the case-random design were a subset of the controls of the case-control design. This held for both sample sizes. Furthermore, the smaller sample was a subset of the larger one. As we report results of over 100 independent simulations in each setting, we believe the random effects are well controlled. We did not handle the issue of population stratification in this study. The trio design is generally robust to stratification effects; in case-control studies false positives resulting from population substructure can be reduced by genomic control (see for example, Devlin *et al*[5]).

Haplotyping

Trio based inference of haplotypes was performed simply by deducing the phase of each marker from the genotypes of the child and the parents. If all members of the trio were heterozygous at a marker, the respective alleles of the child and the pseudohaplotype were denoted as unknown in the haplotypes. This caused some of the data to be missing. More complex methods for haplotype inference would estimate the phases of these markers and eventually improve the quality of the haplotypes. We also used the best possible case, true haplotypes, to approximate an upper limit for (family based) haplotyping.

For population based statistical reconstruction of haplotypes we used the HaploRec algorithm.[6] The method works by fitting a variable order Markov chain model of haplotypes to the observed genotype data (see the appendix for more details). Unlike most other methods for population based haplotyping, HaploRec allows recombinations in the marker map and does that without assuming a specific haplotype block structure. As it is also computationally efficient, we chose it for our study involving a large number of individuals and markers.

Association analysis methods

Allelic association was estimated by predicting the disease susceptibility locus to be at the marker that has the highest value for the χ^2 test statistic. For haplotype association, each haplotype of 1–10 markers in length was evaluated with the χ^2 test, then the disease susceptibility gene was predicted to reside in the middle point of the best haplotype. In case of two or more equally good results, one of the best markers or haplotypes was chosen at random.

For each dataset, the p value of the best allele or haplotype was computed by a permutation test; the disease association statuses of the original haplotypes were randomly shuffled 9000 times, the χ^2 values recalculated each time, and the best χ^2 value of each permutation used as the test statistic. This procedure takes into account the multiple testing of haplotypes and produces a corrected p value.

We estimated the statistical power to detect the disease susceptibility gene at a significance level of 0.05. Because the simulated chromosomes were only 15 cM long (approximately 1/187 of the human genome) we adjusted the p values for genomewide analysis by Šidák correction, essentially assuming that the genome consists of 187 independent 15 cM blocks. The genomewide p value (p^*) was estimated as $p^* = 1-(1-p)^{187}$, where p is the p value obtained for the simulated 15 cM region using permutation tests.

In addition to the statistical power to detect the disease susceptibility gene, the association analysis methods described above give a point estimate for the locus as a result, and the results over 100 simulation replicates give a sample of prediction errors. The accuracy of association analysis can be then visualised as a cumulative distribution function of (absolute) prediction error.

Throughout the results, the maximum prediction error shown is 1000 kb (1 cM in the simulated data) as prediction errors over 1 cM to either side of the true disease susceptibility gene location are not particularly interesting for fine mapping. With the population and disease models used, the methods are reasonably accurate, so that most of the interesting differences lie within this range. In addition to the experimental error curves, the prediction error of uniform random guesses (without any data) is given in all figures for reference. With our disease models, detecting the presence of a gene is more difficult than locating it, and correspondingly the powers are much less than 1. However, this design is intended to bring out differences between the methods.

RESULTS

A direct comparison of the case-control and trio designs, under the assumption of equal genotyping costs, showed that the case-control design is substantially more accurate for association analysis across the three different disease models (fig 1A, common and rare models; fig 1B, intermediate model). Even in the best possible case (with true haplotypes) the accuracy of the trio design was clearly inferior to the case-control designs. The differences in the statistical power to detect the gene are even more striking (table 3): the powers were 0.72 *v* 0.07 (rare), 0.48 *v* 0.08 (intermediate), and 0.76 *v* 0.30 (common) for the case-control and trio designs, respectively. Using the true haplotypes did not significantly improve powers in the trio designs.

To validate the use of the relatively simple haplotype association mapping method in this study, we compared the results with those obtained by allelic association and by the recently introduced EATDT (applicable to trio based data only).[4] The haplotype association method used in this study actually outperforms both allelic association and EATDT (fig 1C, intermediate model; table 3, last two lines of each block), indicating that it is powerful and accurate, and therefore suitable for measuring the differences between different study design and haplotyping approaches.

We next perform a detailed experimental analysis of the differences between the case-control and trio based approaches. In particular, we isolate the effects caused by sample size, sample ascertainment, and haplotyping method, and study each of them separately. For illustration, we only show prediction accuracy curves for the intermediate disease model (the most difficult one); the power and prediction results for all three disease models are summarised in table 3.

Sample size

The most obvious cause for the performance differences above is the sample size: with equal genotyping costs, the effective sample size in the trio based approach is only two thirds of that in the population based approach. However, the difference in sample size only explains a part of the difference (fig 1D, intermediate model; table 3).

Sample ascertainment

The two sample ascertainment methods are quite different. In the trio based approach, the non-transmitted chromosomes of the parents are used as additional data: as a pseudo-control if the parent is healthy, or as a pseudo-case if the parent is diseased. This procedure often results in somewhat unbalanced numbers of cases and controls, but this does not seem to have much effect on the mapping accuracy (results not shown).

To test the effect of sample ascertainment, we conducted experiments where the sample size and haplotyping method were fixed, and only the origin of controls varied. In

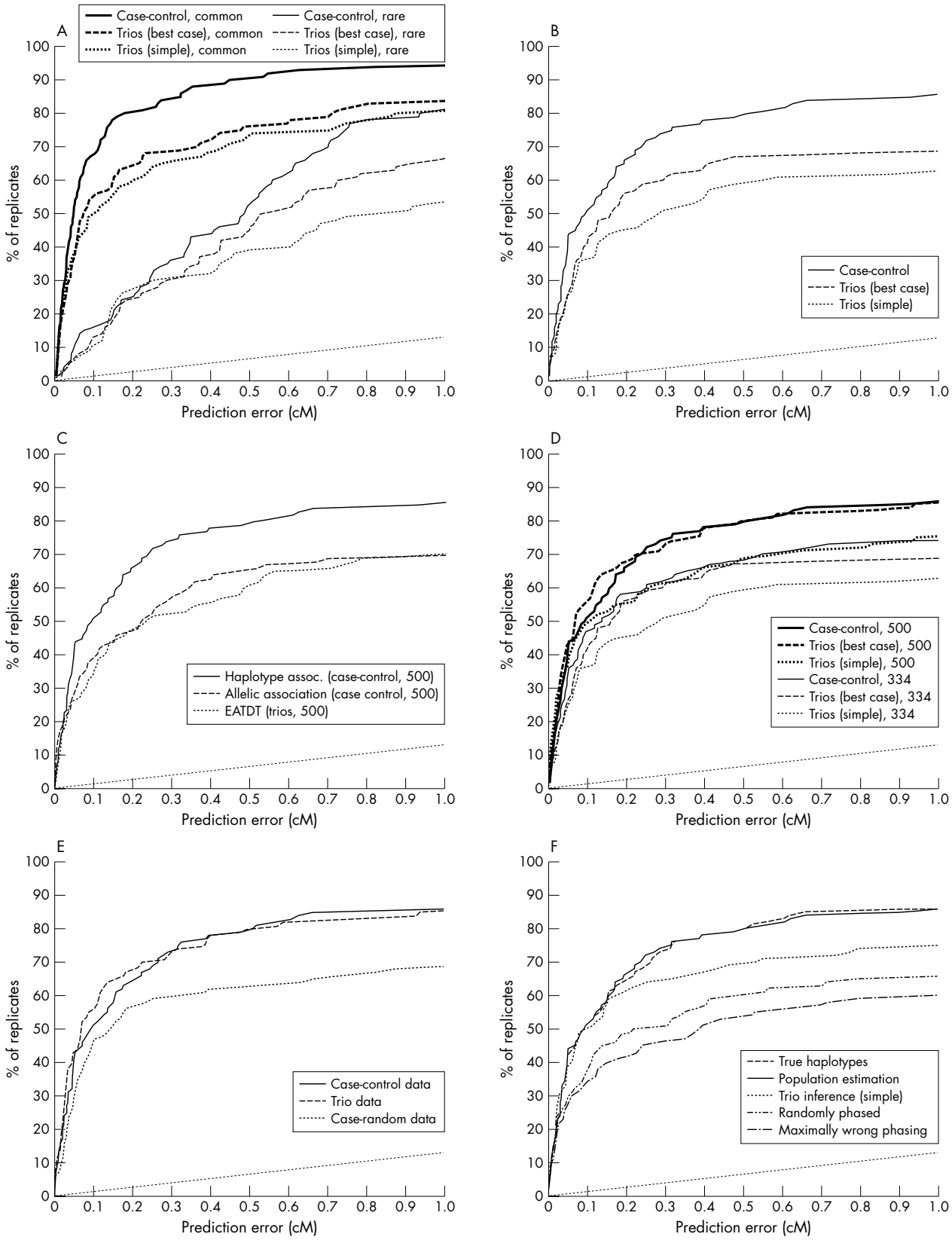

Figure 1 Prediction errors with simulated data. Where not otherwise stated, the disease model is intermediate, the mapping method is haplotype association, the case-control approach consists of 500 individuals haplotyped with a population based method, and the trio based approach has 167 trios (effective sample size 334) with trio inferred and true (best case) haplotypes. Case-control and trio based approaches for (A) common and rare, and (B) intermediate disease model. Effect on prediction error of (C) association mapping method, (D) sample size, (E) sample ascertainment method (known haplotypes, sample size 500), and (F) haplotyping method (case-control data, sample size 500). In all panels, the dotted line close to the bottom is the expected error for random predictions.

Table 3 The power to detect the disease susceptibility gene using different sample ascertainment methods, haplotyping methods and sample sizes for rare, intermediate and common mutation models

	Sample ascertainment and effective sample sizes					
	Case-control		Case-random		Trios	
	Sample size		Sample size		Sample size	
Haplotyping method	500	334	500	334	500 (750*)	334 (500*)
Rare						
Population based estimation	**0.72 (0.81)**	0.22 (0.68)	0.56 (0.78)	0.16 (0.57)	NA	NA
Trio based inference	0.37 (0.69)	0.05 (0.55)	0.31 (0.64)	0.05 (0.48)	0.33 (0.62)	**0.07 (0.53)**
True haplotypes	0.74 (0.79)	0.17 (0.68)	0.56 (0.75)	0.11 (0.59)	0.62 (0.74)	**0.10 (0.66)**
Randomly phased haplotypes	0.07 (0.53)	0.01 (0.44)	0.07 (0.45)	0.02 (0.40)	0.05 (0.52)	0.01 (0.39)
Maximally wrong phases	0.08 (0.45)	0.01 (0.29)	0.05 (0.37)	0.01 (0.27)	0.03 (0.48)	0.00 (0.30)
Allelic association	*0.06 (0.47)*	0.01 (0.38)	0.05 (0.48)	0.02 (0.39)	0.04 (0.51)	0.01 (0.33)
EATDT	NA	NA	NA	NA	0.49 (0.84)	*0.13 (0.62)*
Intermediate						
Population based estimation	**0.48 (0.85)**	0.21 (0.74)	0.20 (0.68)	0.08 (0.52)	NA	NA
Trio based inference	0.39 (0.74)	0.13 (0.56)	0.17 (0.61)	0.06 (0.43)	0.25 (0.75)	**0.08 (0.62)**
True haplotypes	0.46 (0.85)	0.21 (0.70)	0.19 (0.68)	0.07 (0.53)	0.34 (0.85)	**0.11 (0.68)**
Randomly phased haplotypes	0.22 (0.65)	0.06 (0.52)	0.13 (0.57)	0.04 (0.38)	0.06 (0.48)	0.02 (0.35)
Maximally wrong phases	0.20 (0.60)	0.09 (0.45)	0.08 (0.50)	0.04 (0.29)	0.10 (0.47)	0.02 (0.39)
Allelic association	*0.24 (0.69)*	0.07 (0.55)	0.08 (0.56)	0.06 (0.38)	0.06 (0.62)	0.04 (0.47)
EATDT	NA	NA	NA	NA	0.15 (0.70)	*0.07 (0.47)*
Common						
Population based estimation	**0.76 (0.94)**	0.47 (0.87)	0.59 (0.95)	0.25 (0.84)	NA	NA
Trio based inference	0.72 (0.94)	0.46 (0.82)	0.54 (0.91)	0.30 (0.75)	0.61 (0.95)	**0.30 (0.80)**
True haplotypes	0.77 (0.95)	0.47 (0.88)	0.60 (0.97)	0.27 (0.83)	0.62 (0.94)	**0.29 (0.83)**
Randomly phased haplotypes	0.61 (0.85)	0.35 (0.73)	0.39 (0.80)	0.20 (0.69)	0.34 (0.83)	0.19 (0.68)
Maximally wrong phases	0.50 (0.79)	0.30 (0.67)	0.35 (0.75)	0.20 (0.67)	0.35 (0.77)	0.08 (0.64)
Allelic association	*0.57 (0.90)*	0.33 (0.83)	0.42 (0.87)	0.27 (0.73)	0.36 (0.82)	0.19 (0.70)
EATDT	NA	NA	NA	NA	0.47 (0.86)	*0.23 (0.79)*

*Number of genotyped individuals. Numbers in parenthesis are fractions of predictions for which prediction error was less than 1 cM. Results obtained with allelic association and EATDT[4] are included for comparison; their primary alternatives are printed in italics.

particular, population based controls were in some experiments haplotyped using their parents to form trios. According to the results, there is virtually no difference in the prediction accuracies between the two major sample ascertainment methods (fig 1E, intermediate model). In terms of the power to detect a gene, on the other hand, population based controls tend to be more powerful than trio based data (table 3).

Sampling random instead of healthy individuals has a clear negative effect. For the most difficult disease model, the intermediate model, the effect is comparable with the effect of sample size. For simpler models, the effect is smaller. A trio based strategy where all non-transmitted haplotypes are labelled as controls yields similar results to those found by sampling random individuals (results not shown).

Haplotyping method

The final aspect in which the case-control and trio based approaches differ is the haplotyping method. Population based reconstruction is based on estimated phases, which can contain errors, especially with the long maps used in this study (451 markers, 15 cM). The simple trio based approach we used, on the other hand, cannot always resolve the phase of a marker, in which case we marked the alleles as unknown. The true haplotypes used in our experiments represent the best case scenario for haplotyping and can be viewed as an upper bound for the performance of any trio haplotyping method.

Controlled experiments, where other factors (in particular, sample size and sample ascertainment) are constant, show that the mapping results obtained with population based haplotyping are virtually identical to those obtained with the true haplotypes across all tests (fig 1F, table 3). In other words, the (few) phasing errors did not affect mapping power or accuracy.

The simple trio based haplotyping, in turn, is clearly inferior. This is explained by the number of missing data, as all inferred phases have to be correct. On average, trio based haplotyping resulted in 5.4% missing alleles in our simulated datasets. In the best case, a sophisticated trio based haplotyping method would give mapping accuracy equal to the true and the population estimated haplotypes (fig 1D). However, the power of the trio design to detect a gene remains inferior to the case-control design even in this case (table 3).

Example: asthma data

We demonstrate the power of the case-control design and population based haplotyping on real asthma data, consisting of 194 small families.[3] Our goal was to compare different approaches with this real data, not to reproduce the original results of Laitinen *et al*. The subjects were genotyped by Laitinen *et al* for 91 microsatellite and 64 SNP markers spanning a 20 cM region in chromosome 7. The original genotyping process had been iterative; new markers were added where the intermediate analyses showed strongest association. In the final stage, a 133 kb risk conferring haplotype was identified in the region densely covered by SNP markers.[3]

To mimic the case-control study design, we used families with an affected and a healthy parent, constituting an effectively independent case-control pair. For the trio based approaches, we also randomly sampled one child in each of the families. We sampled at most one such trio per pedigree, giving 93 trios from the available 194 pedigrees.

All markers with at least 20% of genotypes missing were rejected. The remaining marker map consisted of 73 microsatellite and 15 SNP markers. We haplotyped the case-control dataset of affected and healthy parents using HaploRec. For comparison, we inferred the haplotypes of a random subsample of 62 trios (two thirds of 93) using the genotypes of the children. Additionally, we subsampled the haplotypes of the parents in the 62 trios from the set of 93

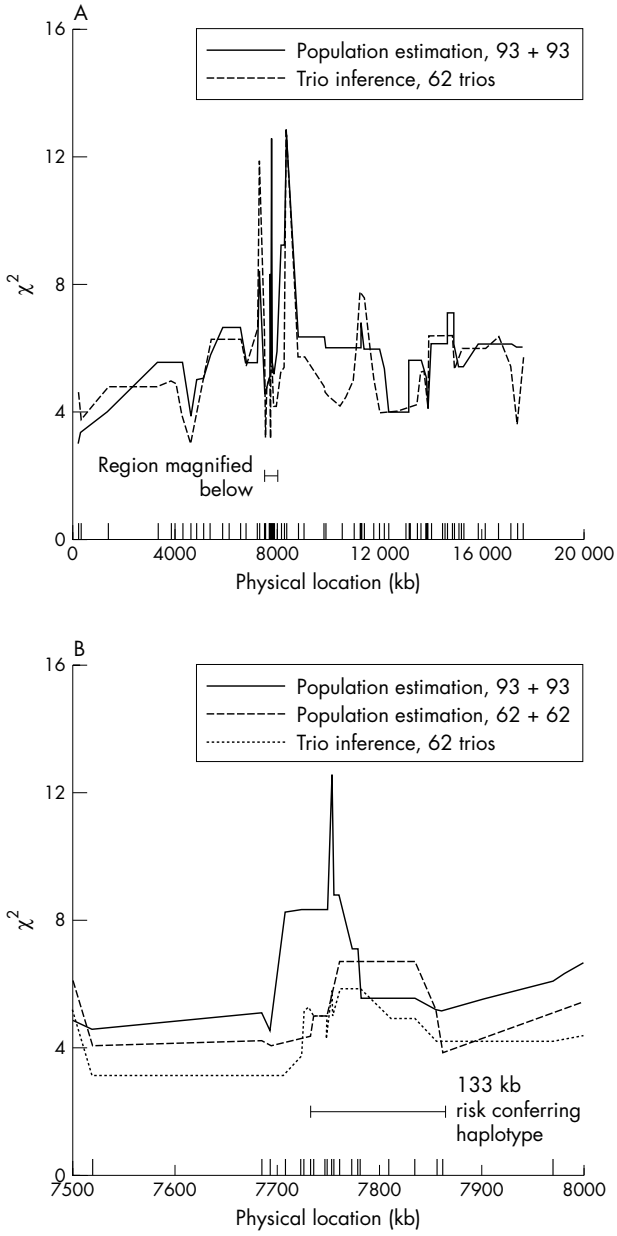

Figure 2 Haplotype association in asthma data,[3] with the population based and trio based approaches. (A) Entire 20 cM region; (B) 500 kb region containing the identified 133 kb haplotype (7733–7865 kb). The curves show the highest χ^2 value for each marker from the set of all 1–10 marker haplotypes spanning over the marker. The locations are reported relative to an arbitrarily chosen origin. The bars at the bottom show the locations of the markers.

case-control pairs haplotyped using HaploRec. This dataset represents our best estimate of the haplotypes in the trios.

With the case-control sample, there was a strong association peak within the correct 133 kb region, but also a false positive about 500 kb to the right (fig 2). In contrast, the trio based setting did not show any associations within the correct region, but two false positives instead. Increasing the sample size to include all 93 trios did not produce different results (data not shown). This is a strong indication of the potential of the case-control approach to association mapping.

DISCUSSION

We reported on simulation experiments where we compared sample ascertainment strategies and related haplotyping approaches in association mapping studies. We considered two main alternatives: ascertainment of family trios with an affected child, from which it is easy to infer the haplotypes partially, and ascertainment of a case-control sample of unrelated individuals, for which the haplotypes were estimated statistically. We conducted the mapping step using haplotype association analysis, a simple but powerful and efficient method. Case-control samples were haplotyped with HaploRec.[6] Both haplotype association and HaploRec scale up to large amounts of markers and individuals and are suitable for high throughput association studies. Finally, we isolated and experimentally analysed the effects of three separate factors: sample size, sample ascertainment method, and accuracy of haplotyping.

For an equal number of genotyped individuals, the effective sample size in the trio based approach is only two thirds of that of the case-control design. According to our experiments, this difference has a major effect on the mapping power and accuracy, as expected.

In order to maximise the genetic effect in the sample, the controls should be ascertained from unaffected individuals to minimise the frequency of disease susceptibility allele. In a population based study, it can be possible to ascertain healthy controls for this purpose. Alternatively, random population controls are often used since their ascertainment is easier, especially if they are stratified to match the cases. For rare diseases, the difference between healthy and random controls is marginal.

We ascertained the cases in a standard manner, using an affected child as a proband. If all affected individuals are equally likely to be chosen as probands, then families with several affected children will be over-represented in the sample (known as classical ascertainment bias; see for example, Fisher[7] or Cannings and Thompson[8]). For association analysis, this bias increases the power as the expected frequency of the disease susceptibility gene is then further elevated in the cases.

According to our results, the case-control setting is more powerful than the trio based design, but they are roughly equally accurate. The case-random design was clearly inferior, but this depends on the disease model, in particular the susceptibility allele frequency. Trios have the benefit that they produce controls that are well paired with cases for population substructure and other factors that are not uniform over the population. Ascertainment of matching controls from the population is more difficult, but genomic control can be used to correct for stratification.[5] We did not simulate any population substructure effects.

In our experiments, haplotypes inferred with a population based statistical method were equally powerful as using the true haplotypes. This is in contrast to the result of Morris *et al*, who concluded that statistically inferred haplotypes are inferior to the true haplotypes, and genotype based approaches should be preferred.[9] There are several differences between our studies that can partially explain the results. Firstlt, the results of Morris *et al* are based on a shorter map (950 kb interval and 20 SNPs versus 15 000 kb and 451 SNPs in our study). Secondly, Morris *et al* used their own COLDMAP software for mapping (we could not evaluate the effect of this choice because, according to Morris *et al* themselves, COLDMAP[9] is not feasible for datasets of the size of our study). Thirdly, they used SNPHAP (designed by D Clayton) to estimate haplotypes. SNPHAP is not well suited for long maps with recombinations, and is likely to be a suboptimal choice here (we could not evaluate it in our study because it often stopped without haplotyping all individuals). As a fourth reason, different simulation techniques were applied.

The effect of haplotyping accuracy on fine mapping is greatest near the disease susceptibility locus in the haplo-

types carrying the disease susceptibility allele. As this is exactly where linkage disequilibrium is increased in the cases, this area is likely to be best haplotyped by methods that give emphasis on local structure, such as HaploRec. It was found to have excellent switch accuracy (defined as the fraction of neighbouring phases, between each pair of consecutive heterozygous markers, reconstructed correctly): 99.6% and 99.3% on average for sample sizes of 500 and 334, respectively.

With the trio based haplotype inference, the only source of uncertainty is similar heterozygotes, for which we marked the alleles as unknown in our simple method. This resulted in a poorer association mapping performance compared with the population based estimates. Based on this result, an obvious recommendation is to use population based techniques to augment trio based inference when trios are available, to avoid missing alleles. However, our experiments with the true haplotypes suggest that even the best possible haplotyping method would give the same mapping accuracy as the population based haplotyping, and the power to detect the gene would remain inferior compared with the case-control design, owing to the smaller effective sample size.

Despite some contrasts with the conclusions of Morris *et al*,[9] we do agree with many of their points. In particular, mapping results based on inferred haplotypes are likely to be optimistic in terms of confidence or credibility intervals, owing to exaggeration of linkage disequilibrium.

In further research, it is important to test the utility of the case-control study design in more real life experiments, to verify that it does not suffer from unexpected effects from genotyping errors, missing data, or varying disease models. The effect of more elaborate association methods (for example, those proposed by Purcell *et al*[10]) on the power and mapping accuracy should be investigated, although our comparisons to EATDT suggest that the simple haplotype association with a population based haplotyping is a powerful method for association analysis.

CONCLUSIONS

For future high throughput association mapping studies, case-control design combined with efficient population based haplotyping method can be more powerful alternative than the trio design. To enrich the specific disease susceptibility genes in the cases, the cases should rather be familial than sporadic. Naturally, the opposite holds for the controls: it is useful if the presence of the specific disease or trait in question can be reliably excluded. The downside of using population controls is that it may prove difficult to find controls matched for ethnic origin in studies of stratified or mixed populations; genomic control alleviates some of the issues.

According to our results, data from existing epidemiological research projects, consisting of cases only, or case-control pairs, can be easily and effectively used in large scale haplotype based association analysis. This opens great opportunities for utilising the vast blood sample or biopsy collections maintained in many university hospitals and national health institutes.

APPENDIX

Two phase simulation procedure

We used a two phase procedure for the simulation of relatively young, isolated founder populations. In the first phase, the founder chromosomes were simulated using a coalescent model with recombination and the infinite sites mutation model.[11] We used a coalescent simulation procedure based on the spatial algorithm of Wiuf and Hein.[12] The spatial algorithm produces a graph similar to the ancestral recombination graph of Griffiths and Marjoram.[13] The algorithm first generates a tree genealogy for one end of the sequence using the standard coalescent model, and then iteratively proceeds over the sequence until it is covered in its entirety. In each iteration, the distance to the next crossover in any of the lineages in the current graph is first drawn from an exponential distribution. A new recombination node is then generated to a randomly chosen point in the graph, splitting the lineage at the chosen point. The new lineage is coalesced with a randomly chosen lineage.

In our implementation, when updating the graph, we ignore the lineages of the graph that are no longer part of the local genealogical tree. This way the graph is a tree at all times, and, consequently, the algorithm simulates long sequences much more rapidly. Although our model is an approximation, experiments show excellent agreement with the true coalescent model with recombination, in terms of linkage disequilibrium (D') and allele frequency distributions (data not shown).

We simulated recombination at a flat rate of 10^{-8} crossovers per bp per meiosis, with no chiasma interference. Mutations were generated at rate 10^{-8} per bp per generation. The generated mutations constitute the set of all possible SNP markers and disease susceptibility loci for the study. The effective size of the population was 10 000 individuals.

In the second phase, inheritance of chromosomal segments in the founder population was simulated. We used the populus simulator of Ollikainen for simulating random mating in distinct generations.[14] Recombinations were modelled as in the first phase; mutations were not simulated during the second phase. As a result, segment compositions of the individuals were obtained; for each individual, the two homologous chromosomes consisted of separate segments that have been inherited from the founder chromosomes intact.

The disease susceptibility mutation and minimum minor allele frequencies for markers were fixed for each study setting. For the choice of the disease susceptibility locus and marker map, we therefore needed to compute the frequencies of the minor alleles. This was performed for all SNPs in the final population by using the simulated chromosomal segment decompositions and the founder haplotypes acquired from the first phase simulation.

Next, a set of approximately equidistantly spaced markers with a minor allele frequency exceeding a specified minimum frequency was chosen from the set of SNPs. At the same time, the disease susceptibility locus was chosen from the remaining SNPs so that the frequency of the disease susceptibility mutation was closest to the desired value. If there were more than one such candidate available, one of those was randomly chosen.

Individuals were diagnosed based on their genotype at the disease susceptibility locus and the specified disease model. After diagnosis, a desired sample was picked. Finally, allelic data were generated for the sampled individuals and their parents.

Haplotyping method

For population based statistical reconstruction of haplotypes we used the HaploRec algorithm, which is targeted especially for large numbers of relatively sparsely spaced markers.[6] HaploRec assumes Hardy-Weinberg equilibrium, which means that the probability of a haplotype pair is modelled as a product of the probabilities of the two individual haplotypes. The probability of a haplotype is broken into a product of conditional probabilities of individual alleles, with each allele conditional on a varying number of its immediate neighbours. When computing the probability of haplotype H

of length l, the distribution of alleles at marker i is estimated by conditioning on the longest observed haplotype fragment that (a) matches haplotype H and ends at marker i–1, and (b) has an estimated relative frequency of at least 0.2%:

$$P(H) = P(H(1)) \prod_{i=2,\ldots,\ell} P(H(i) \mid H(s_i, i-1)), \quad (1)$$

where $H(i)$ is the allele at marker i, $H(i, j)$ is the haplotype fragment covering markers i–j and $s_i = \min\{s \mid \Pr(H(s, i\text{–}1)) \geqslant 0.2\%\}$. Use of a frequency threshold is motivated by the fact that a long haplotype fragment is likely to be shared by several individuals only if it is inherited from the same ancestor, and thus is useful in estimating haplotypes. Given parameter estimates for the variable order Markov model, the haplotypes are reconstructed by choosing the phases such that the resulting pair of haplotypes has the maximum product of probabilities. Because in practice, neither the model parameters or the haplotypes are known in advance, the original HaploRec algorithm was adapted to apply an EM-like algorithm for simultaneously learning the model and reconstructing the haplotypes. The algorithm starts with a uniform model, and alternates between steps of reconstructing the haplotypes and estimating the model parameters.

ACKNOWLEDGEMENTS

The asthma data[3] were kindly provided to us by Dr T Laitinen of Oy GeneOS Ltd (www.geneos.fi). An implementation of EATDT[4] was kindly provided by Dr D J Cutler of Johns Hopkins University School of Medicine. We would like to thank Dr P Uimari of Oy Jurilab Ltd (www.jurilab.com) for constructive discussions.

Authors' affiliations

P Hintsanen, P Sevon, L Eronen, H Toivonen, Helsinki Institute for Information Technology, Basic Research Unit, Department of Computer Science, University of Helsinki, Finland
P Onkamo, Department of Biological and Environmental Sciences, University of Helsinki, Finland

This research has been supported by Tekes, the National Technology Agency of Finland.

Competing interests: there are no competing interests

Data access: HaploRec software for population based reconstruction of haplotypes and the simulated datasets are available at www.cs.helsinki.fi/group/genetics/.

Correspondence to: Dr H Toivonen, Department of Computer Science, Gustaf Hällströmin katu 2b, PO Box 68, FI-00014 University of Helsinki, Finland; hannu.toivonen@cs.helsinki.fi

Received 17 June 2005
Revised version received 3 October 2005
Accepted for publication 18 October 2005
Published Online First 28 October 2005

REFERENCES

1 **Risch N**, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;**273**:1516–17.
2 **Laitinen T**, Daly MJ, Rioux JD, Kauppi P, Laprise C, Petays T, Green T, Cargill M, Haahtela T, Lander ES, Laitinen LA, Hudson TJ, Kere J. A susceptibility locus for asthma-related traits on chromosome 7 revealed by genome-wide scan in a founder. *Nat Genet* 2001;**28**:87–91.
3 **Laitinen T**, Polvi A, Rydman P, Vendelin J, Pulkkinen V, Salmikangas P, Makela S, Rehn M, Pirskanen A, Rautanen A, Zucchelli M, Gullsten H, Leino M, Alenius H, Petays T, Haahtela T, Laitinen A, Laprise C, Hudson TJ, Laitinen LA, Kere J. Characterization of a common susceptibility locus for asthma-related traits. *Science* 2004;**304**:300–4.
4 **Lin S**, Chakravarti A, Cutler DJ. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* 2004;**36**:1181–8.
5 **Devlin B**, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;**60**:155–66.
6 **Eronen L**, Geerts F, Toivonen H. A Markov chain approach to reconstruction of long haplotypes. *Pac Symp Biocomput* 2004:104–15.
7 **Fisher R**. The effect of methods of ascertainment upon the estimation of frequencies. *Ann Eugen* 1934;**6**:13–25.
8 **Cannings C**, Thompson E. Ascertainment in the sequential sampling of pedigrees. *Clin Genet* 1977;**12**:208–12.
9 **Morris A**, Whittaker J, Balding D. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* 2004;**74**:945–53.
10 **Purcell S**, Sham P, Daly MJ. Parental phenotypes in family-based association analysis. *Am J Hum Genet* 2005;**76**:249–59.
11 **Kimura M**. The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. *Genetics* 1969;**61**:893–903.
12 **Wiuf C**, Hein J. Recombination as a point process along sequences. *Theor Popul Biol* 1999;**55**:248–59.
13 **Griffiths RC**, Marjoram P. An ancestral recombination graph. In: Donnelly P, Tavaré S, eds. *Progress in population genetics and human evolution. IMA volumes in mathematics and its applications*. Berlin: Springer-Verlag, 1997;**87**:257–70.
14 **Ollikainen V**. *Simulation techniques for disease gene localization in isolated populations*, PhD thesis. Helsinki: University of Helsinki, Department of Computer Science, 2002.