Association Mapping and Significance Estimation via the Coalescent

Gad Kimmel[1,2], Richard M. Karp[1,2], Michael I. Jordan[1,3] and Eran Halperin[2]

Affiliation:

1. Computer Science Division, University of California Berkeley, Berkeley, CA 94720, USA
2. International Computer Science Institute, 1947 Center St., Berkeley, CA 94704, USA
3. Department of Statistics, University of California Berkeley, Berkeley, CA 94720, USA

Running title: Association Mapping via the Coalescent

Corresponding author: Gad Kimmel, kimmel@cs.berkeley.edu

Abstract

The central question asked in whole genome association studies is how to locate associated regions in the genome and how to estimate the significance of these findings. This is usually done by testing each SNP separately for association, and then applying a suitable correction for multiple hypothesis testing. However, SNPs are correlated by the unobserved genealogy of the population, and a more powerful statistical methodology would attempt to take this genealogy into account. Leveraging the genealogy in association studies is challenging since the inference of the genealogy from the genotypes is a computationally intensive task, in particular when recombination is modeled, as in ancestral recombination graphs. Furthermore, if large numbers of genealogies are imputed from the genotypes, the power of the study might decrease if these imputed genealogies create an additional multiple hypothesis testing burden. Indeed, we show in this paper that several existing methods that aim to address this problem suffer from either low power or from a very high false positive rate; their performance is generally not better than the standard approach of separate testing of SNPs. We suggest a new genealogy-based approach, CAMP (Coalescent based Association MaPping), which takes into account the trade-off between the complexity of the genealogy and the power lost due to the additional multiple hypotheses. Our experiments show that CAMP yields a significant increase in power relative to previous methods, and that it can more accurately locate the associated region.

1 Introduction

Recent advances in genotyping technologies have considerably improved our understanding of common complex diseases through whole-genome association studies. In these studies a population of cases and controls is collected, in which hundreds of thousands of single nucleotide polymorphisms (SNPs) are genotyped. These studies search for SNPs that are associated with the studied disease, by measuring the difference in the SNP-allele distributions between the cases and the controls (e.g., [1,2]).

Since complex diseases are caused by multiple environmental and genetic factors, the differences in allele frequencies between the cases and the controls for any given SNP can be expected to be quite small. Therefore, analyses that achieve high statistical power are essential for these studies. Additionally, although current technology (e.g., the Affymetrix SNP Array 6.0, and the Illumina human1m-duo beadchip) allows measurement of nearly two million genetic variants for each individual, this is still only a fraction of the set of genetic variants, and statistical methods are needed to cope with this partial assessment of genetic variation.

The statistical analysis of a typical association study involves the testing of individual SNPs or genomic regions for association, and the evaluation of the significance of the findings. The simplest approach to significance testing is to test each marker separately for association [3,4]. Many attempts have been made to move beyond separate testing by leveraging the unobserved genealogy of the chromosomes (e.g., [5,6]). These proposals aim to increase statistical power by taking into account the dependency among SNPs. Model-based approaches in particular try to infer aspects of the unobserved genealogy. In practice, however, this is a non-trivial task since the genealogy has to be inferred from the genotypes. As we show in this paper, the loss of information caused by erroneous inference of the genealogy can be detrimental to the association, and thus genealogy-based methods are not always desirable.

Previous methods that use genealogies in association studies face two main challenges. First, the number of possible genealogies is very large, and even more so when recombination events are taken into account; thus, it is infeasible to examine all possible genealogies. Second, an inferred genealogy determines a large set of genealogy-based association tests (these can be expressed as tests of SNP interactions); a major challenge is how to choose a subset of these tests such that the increased number of hypotheses tested will not decrease the power. If the tests are not chosen properly, the statistical power can be reduced considerably due to the burden of multiple hypotheses, even when the genealogical modeling is accurate.

In this paper, we suggest a new genealogy-based approach that takes into account the trade-off between the complexity of the genealogy and the power lost because of multiple hypotheses. The approach we present seeks to avoid excessive loss of power due to multiple testing, while still testing the observed mutations and selected putative unobserved mutations suggested by plausible genealogies. As with previous genealogy-based methods, we test selected SNP interactions. The core of our method is to exploit properties of the coalescent to decide which interactions can be ignored. In a nutshell, we construct a perfect phylogeny graph which represents the genealogy of the haplotypes, and restrict attention to observed mutations and to unobserved mutations that are consistent with that graph.

Many genealogy-based association tests have been suggested in earlier work. One popular way of using genealogy in association studies is through the use of Ancestral Recombination Graphs[7] (ARGs). These graphs aim to model the coalescence and the recombination events explicitly. Several studies have proposed performing full-likelihood or Bayesian inference under the ARG model (e.g.,[8,9]). This is, however, a technically challenging problem, and the proposed solutions are feasible computationally only on relatively small data sets. Zollner and Pritchard[5] suggested an approximation to this inference problem in which testing for association is done by a likelihood

ratio test which is obtained by calculating the probability of the disease mutation given the genotypes and the disease status. The inference is performed by a Markov Chain Monte Carlo (MCMC) algorithm. This approach has the advantages of model-based procedures, but it is too expensive computationally to be used in a large-scale whole-genome association study.

A different approximate approach to association mapping was suggested by Durrant et al.[10]. Their main idea is to perform a cladistic analysis of SNPs. The cladogram captures the successive partitioning of SNP haplotypes into clusters. At each partition, clusters of haplotypes from the previous partition are merged such that the mean pairwise haplotype diversity is minimized within the new clade. The cladogram is built using a sliding window of SNPs. In each window the best partition of haplotypes is chosen. This procedure incorporates two levels of multiple testing, which are adjusted by a Bonferroni correction.

Minichiello and Durbin[6] introduced another approximation scheme for the inference of ARGs. There are two stages to their analysis: First, they attempt to infer all plausible ARGs, using a heuristic algorithm. Second, a genealogical tree at each locus is built, and a possible causative mutation at each branch is tested. Since the true ARG is unknown, this analysis is averaged over a set of inferred ARGs.

In general, the genealogy-based methods are meant to improve upon the naive approach to association testing in which each SNP is tested separately using a χ^2 test and the tests are adjusted for multiple hypothesis testing using a permutation test (we will refer to this approach as *standard* χ^2). To assess the extent to which this goal has been realized by existing methods, we compared these methods to the naive approach. In our experiments, we found that the naive approach has more power and a lower false positive rate than any of the tested methods. This surprising result motivates our new genealogy-based method, which we refer to as *CAMP* (Coalescent based Association MaPping).

Like previous methods, CAMP tests for interactions of SNPs or haplotypes with disease. To address the issues of computational complexity and multiple hypothesis testing, our emphasis is on reducing the number of tests. The core of our method is to exploit properties of the coalescent to decide which interactions can be ignored. Briefly, we construct a perfect phylogeny graph which represents the genealogy of the haplotypes, and restrict attention to observed mutations and to unobserved mutations which are consistent with that graph, in the sense that each of the unobserved mutations is consistent with a larger graph that retains the perfect phylogeny property. The larger graph represents a genealogy of the haplotypes with the unobserved mutation.

As opposed to ARGs, our method does not model the recombination events explicitly in detail. Indeed, we begin our presentation by making the simplifying assumption that there are no recombinations across the studied region, and that there are no recurrent mutations (this is often referred to as a *perfect phylogeny model*, or a coalescent model with the infinite site assumption). It is well known that in order to satisfy the assertion that a region is consistent with a perfect phylogeny model, the region has to comply with the four gamete test; put differently, every pair of SNPs has at most three out of the four possible haplotypes. We use this characterization to define a simple version of our method for generating unobserved mutations. We then back off from the simplifying assumption of no recombination and consider a model that allows some deviation from the four gamete condition. This yields the CAMP algorithm, which can be viewed as a procedure for defining tests based on an approximate genealogy. A similar approach has been taken by[11] in their work on haplotype phasing.

In order to evaluate the power achieved by CAMP, we have tested CAMP on an extensive number of simulated data sets. Our experiments show that CAMP yields a significant increase in power relative to previous methods. In particular, unlike previous methods, CAMP achieves an increase of more than 10% over the standard χ^2. This advantage was observed with different sampling

distances of SNPs and with different numbers of individuals. Thus, by using our method in association studies, we expect that more associated SNPs will be discovered due to the increased power.

2 Methods

2.1 The General Framework

We begin by sketching the main idea of our approach. As in previous approaches, our goal is to exploit the unobserved genealogy of the population in order to map and evaluate the significance of associations. This is done by performing additional tests of interactions between SNPs; these tests correspond to unobserved mutations along the genealogical tree.

The basic idea of our approach is to restrict attention to interactions between pairs of SNPs that may represent a plausible mutation along a genealogy. Our approach relies strongly on the theory of *perfect phylogeny* of SNPs and haplotypes. There are several studies in the literature that have focused on this combinatorial object, yielding theoretical characterizations that provide the basis of our approach; for background see, e.g., [12,13,11,14].

2.2 Notation and Definitions

Let n be the number of individuals tested, and m the number of markers. The $2n \times m$ haplotype matrix is denoted by H. Hence, $H_{i,j} = s$ if the i-th haplotype has type s at the j-th marker, where s can be 0 or 1. Let the vector of the disease status be d. The entries of d are 0 (for a healthy individual) or 1 (for an individual that has the disease).

For a pair of discrete vectors x, y, let $\Omega(x, y)$ denote their contingency table; i.e., $\Omega(x, y)$ is a matrix in which $\Omega(x, y)_{i,j} = |\{k|x(k) = i,\ y(k) = j\}|$ (in our case, the matrix is of size 2×2, since a SNP is two-valued and there are two disease states). An *association function* A is a function that assigns a positive score to a contingency table. Typical examples of association functions are the Pearson score and the Armitage trend statistic. We have used the Pearson statistic in our work; however, it is important to point out that our algorithm does not use any specific

properties of the association function, apart from the property that the score is a function of the contingency table, and the following symmetry property (which holds for the Pearson score):

$$A\left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right\} = A\left\{\begin{pmatrix} a & c \\ b & d \end{pmatrix}\right\}. \quad (1)$$

An *association score* is a function of the haplotype matrix and an arbitrary disease vector e (a binary vector of dimension $2n$), and is defined by $S(H,e) = \max_j A[\Omega(H_{\cdot,j}, e)]$; i.e., the value of the association function at the most associated locus.

The goal is to calculate the significance of a pair (H,e), which is defined as the probability of obtaining an association score at least as large as $S(H,e)$ under a null model. Formally, if e is a random disease vector, the p-value is $\Pr[S(H,e) \geq S(H,d)]$. In addition, we want to accurately find the location of the associated SNPs in the genome. In our case, the null model is defined according to the randomization model in which $e = \pi(d)$ is a permutation of the disease vector and all instances $\pi(d)$ are equiprobable.

In addition to testing all the SNPs of H, we also test selected SNP interactions. For a set of SNPs $j_1, j_2, \ldots, j_k$, suppose that there are Ψ different haplotypes induced by these SNPs, and let $h(j_1, j_2, \ldots, j_k)$ be the $2n$-dimensional haplotype vector induced by these SNPs (so that each element of $h(j_1, j_2, \ldots, j_k)$ is an integer between zero and $\Psi - 1$). A *combinatorial interaction* of the SNPs $j_1, j_2, \ldots, j_k$ is a binary vector v of dimension $2n$ which corresponds to a partition of the Ψ haplotypes into two sets S_1 and S_2. Formally, let S_1, S_2 be disjoint sets of integers such that $S_1 \cup S_2 = \{0, 1, 2, \ldots, \Psi - 1\}$. Then, $v(i) = 0$ if $h(j_1, j_2, \ldots, j_k) \in S_1$, and $v(i) = 1$ if $h(j_1, j_2, \ldots, j_k) \in S_2$.

The perfect phylogeny tree: Our method is based on the construction of a perfect phylogeny tree, and a specific choice of interactions among the SNPs based on the tree. A perfect phylogeny tree is a genealogical tree in which every node corresponds to a haplotype, and every edge

corresponds to a mutation (Figure 1). In a perfect phylogeny tree, we assume no recombination events, and no recurrent mutations. Thus, such a genealogy is equivalent to the coalescent tree with the infinite site assumption.

In a perfect phylogeny model, every pair of SNPs satisfies the four gamete test. Formally, for two SNP vectors $H_{\cdot,i}, H_{\cdot,j}$, we consider the haplotype counts $C_{a,b}(i,j) = |\{H_{x,i}, H_{x,j} | H_{x,i} = a, H_{x,j} = b\}|$. For example, $C_{0,0}(i,j)$ is the number of haplotypes in which both SNP i and j equal 0. We say that the pair of SNPs (i,j) satisfies the four gamete test if there exists at least one pair (a,b) for which $C_{a,b}(i,j) = 0$.

2.3 The Algorithm

The intuition for our method is based on the case where the data is consistent with the perfect phylogeny model. Our algorithm can be applied also to cases where there are deviations from the perfect phylogeny, as discussed in Section 2.4. The algorithm can be outlined as follows:

1. Build a perfect phylogeny tree using a method such as the one developed by Eskin et al.[11].
2. Select all pairs of SNPs that correspond to adjacent edges in the tree (edges that share a common vertex).
3. For each selected pair of SNPs, add a combinatorial interaction vector to the haplotype matrix H as a column.
4. Perform an association test using the augmented haplotype matrix H.

The newly added columns represent putative unobserved SNPs that are plausible given the observed SNPs. In the algorithm described above, we added all pairwise interactions of SNPs but no higher-order interactions of SNPs. It is straightforward to extend this algorithm to also test higher orders of interactions, i.e., haplotypes with more than two SNPs. We restrict the discussion

of this paper to pairwise interactions since we observed experimentally that higher-order interactions did not attain statistically significant improvement of the power (data not shown). We note that in the case of pairwise interactions, the association tests are practically done on haplotypes; the extension of this method to higher order interactions cannot however be expressed as a haplotype test.

The algorithm finds the value of the association function for each SNP and each interaction, with the corresponding association score. We use a permutation test to determine the significance of this score (corrected for multiple hypotheses). Since permutation tests can be quite inefficient, we use an importance sampling method for efficient calculation of the permutation test[4]. Note that the algorithm we use is generic, and we could use any test for association for each of the interactions (e.g., a two-by-two χ^2 test, or a three-by-two trend test).

Even though the above algorithm is quite simple, it is not immediate to see where the gain in power comes from. In the remainder of this section, we will describe the rationale for the algorithm. In order to do so, we will begin with the case where the perfect phylogeny model is consistent with the data. We will explain later how we deal with deviations from the perfect phylogeny model. We begin by describing in detail the process for adding combinatorial interaction vectors and the interpretation of this process as imputed unobserved SNPs.

2.3.1 Selecting the SNP Interactions

Each edge of a perfect phylogeny corresponds to a mutation in some SNP. Contracting an edge in the tree corresponds to the removal of the SNP associated with the edge from the dataset. Thus, every unobserved SNP corresponds to a contracted edge. More generally, we can view a perfect phylogeny on a set of observed SNPs as the result of a series of edge contractions on a larger perfect phylogeny determined by both observed and unobserved SNPs. It follows that the effect of

adding an unobserved SNP to a perfect phylogeny must be to reverse an edge contraction; i.e., to split a node into two copies and insert an edge joining the two copies.

Every putative unobserved SNP that our algorithm constructs corresponds to such an edge insertion. Here, we limit ourselves to the simplest kind of edge insertions: those resulting from the interaction between a pair of observed SNPs corresponding to adjacent edges of the perfect phylogeny.

Any two observed SNPs correspond to edges in the tree, and the deletion of those edges induces three subtrees, corresponding to three different joint values of the two SNPs. For instance, in Figure 1(a), the deletion of SNPs 1 and 5 induces three subtrees, where the first contains the haplotypes $S_1 = \{11000, 10000, 10100\}$, the second contains the haplotypes $S_2 = \{00000, 00010\}$, and the third contains the haplotype $S_3 = \{00011\}$. By our definition of an interaction (a partition of the set of haplotypes), an interaction between the two SNPs corresponds to a partition of the haplotypes into a set S_k versus the rest of the haplotypes. Thus, there are three possible interactions defined by a pair of SNPs (i, j). However, two of the three interactions corresponds to testing one of the SNPs i or j. For instance, in the case of the pair (1,5) described above, testing the interaction $(S_3, S_1 \cup S_2)$ is equivalent to testing SNP 5. In general, under the perfect phylogeny assumption, every pair of SNPs has at most one non-trivial interaction that does not correspond to testing one of the SNPs. This can be shown by case analysis of all possible interaction of SNPs in such a scenario. When SNPs i and j correspond to adjacent edges, the non-trivial interaction corresponds to splitting a node and inserting an edge between the two copies. Indeed, our algorithm imputes precisely the unobserved SNPs that correspond to non-trivial interactions between adjacent edges.

Consider for example the case presented in Figure 1. Assume that SNPs $1, \ldots, 5$ are genotyped, and SNP 6 is the causal SNP. In this case, testing the interaction between SNPs 4 and 5 in the

original tree is equivalent to testing SNP 6. Similarly, testing the interactions between SNPs 2 and 3 is equivalent to testing potential causal SNPs that mutated after SNP 1 has mutated, but before SNPs 2 and 3 have mutated. The CAMP algorithm restricts the set of tested interactions to interactions that correspond to such cases.

We note that there are other edge insertions that are not induced by the interaction of two SNPs. For instance, in the case of a starlike perfect phylogeny in which every leaf is adjacent to the root, any subset of the SNPs may correspond to a mutation that occurred after the root, but before this set of SNPs. In CAMP, we do not consider such higher-order combinatorial interactions, although in theory they may potentially increase power.

The number of tests performed by our algorithm can be quadratic in the number of SNPs (e.g., if the perfect phylogeny tree is a star). However, in practice the great majority of pairs of edges will not be adjacent. In particular, if the tree is degree-bounded (i.e., the maximum number of edges that touch one vertex is below some constant number), the number of imputed unobserved SNPs will be linear in the number of observed SNPs.

2.4 Handling Recombination Events

The algorithm that we have described thus far is based on the perfect phylogeny model, a model which assumes no recombination events. We now describe a modification of our algorithm that pulls back from this simplifying assumption and attempts to provide a partial accounting for recombination events. One may view this modification as an approximation of the perfect phylogeny model.

In the modified algorithm, in place of the perfect phylogeny tree, we instead construct a *perfect phylogeny graph*. Each node in this graph represents a SNP. The edges in the graph are directed and are defined below.

There are two possible relationships between SNPs in the perfect phylogeny: 1. The haplotype 00 can have two descendant haplotypes: 01 and 10, which corresponds to a *brotherhood* relation between the two SNPs. 2. The haplotype 00 can have a descendant 01, which have a descendant 11, which corresponds to a *parenthood* relation between the two SNPs.

We say that two SNPs i, j are in brotherhood relation if $C_{0,0}(i,j)C_{1,1}(i,j) < C_{0,1}(i,j)C_{1,0}(i,j)$; otherwise these SNPs are in parenthood relation. If SNPs i, j are in parenthood relation, then i is defined to be an *ancestor* of j if $C_{1,0}(i,j) > C_{0,1}(i,j)$. It is easy to see that, in the case of a perfect phylogeny in which the root is the haplotype for which all alleles have value 0, this definition agrees precisely with the notion of ancestry in the phylogeny tree. Similarly, if i and j are in brotherhood relation, then neither of them is an ancestor of the other. We now define the edges in the perfect phylogeny graph as follows: There is a directed edge from vertex v_i to v_j if v_i is an ancestor of v_j and there is no other vertex v_x such that v_i is an ancestor of v_x and v_x is an ancestor of v_j. Such a graph can be built using a topological sorting of the vertices.

In this construction we assume that the root of the tree is the haplotype for which all alleles have value 0; we can justify this by rooting the tree in one of the existing haplotypes, and renaming the alleles of each SNP so that the root will satisfy this assumption.

Similarly to the original algorithm described in Section 2.3, we test the interaction of two SNPs if they have a common parent in the perfect phylogeny graph or if one of them is the parent of the other. In constructing the perfect phylogeny graph, we do not consider relations of pairs of SNPs of physical distance in the genome higher some threshold c. We call this threshold the *linkage upper bound*.

Observe that if there are no recombination events, the modified algorithm described in this section is equivalent to the algorithm described in Section 2.3. In a perfect phylogeny, at least one of the four $C_{0,0}(i,j), C_{1,1}(i,j), C_{0,1}(i,j), C_{1,0}(i,j)$ equals zero (i.e., the four gamete test holds), while

here we do not require this property to decide the relationship of the SNPs.

3 Results

3.1 Data Sets

In order to test our approach, we needed a large data set that contains a sequence of several megabases for thousands of individuals. Currently, such a data set does not exist, and therefore we generated simulated population data as follows. We used the SNPs obtained from the HapMap data set as a starting point. To amplify this data, we assumed a fixed population size of 10,000, a mutation rate of 10^{-8} and a recombination rate of 10^{-8}. In each generation, individuals are mated randomly to produce the next generation. The number of children generated by two individuals is a random variable with a predefined distribution. We used 30,000 generations to generate the final population sample. This process was done for 15 megabase pairs along one chromosome. Note that we did not use an approximation-based approach to simulate the population (such as the coalescent model with recombination events or the Li and Stephens model[15]), but rather an explicit forward simulation of the population, which is initiated from a real data set.

We used a multiplicative model to generate samples of cases and controls. We simulated experiments with 1000 cases and 1000 controls. A *panel* is defined to be one experiment. For each panel, a SNP was randomly chosen to be the causal SNP, and was then removed from the panel. We set the disease prevalence to 0.01, and the relative risk to 1.5. We set the linkage upper bound (c) to 50 kb, which has been shown to be a good estimate in humans (e.g.,[4]). We used the perfect phylogeny graph algorithm described in Section 2.4.

The running time of CAMP for 500 cases and 500 controls for 38,864 SNPs on chromosome 1 (corresponding to the Affymetrix SNP chip) on a Sun workstation (with a Quad 2.4GHz AMD

Opteron 850 Processor) is 4 minutes to calculate the scores for each SNP, and an additional 18 minutes for a standard permutation test.

3.2 Evaluation of Previous Methods

Many of the existing genealogy-based methods are computationally inefficient, and thus a large-scale evaluation of these methods is prohibitive. Our experiments involved thousands of panels, each containing thousands of haplotypes with thousands of SNPs, and thus we concentrated on the evaluation of methods that are efficient enough to handle data sets of this size. In particular, coalescent based methods such as LATAG [5] are not computationally feasible for large scale data sets. The Margarita [6] algorithm is also too computationally intensive: it took more than two weeks to analyze a data set of 500 cases and controls with 10,000 SNPs (we used the recommended parameters by the developers: 30 ARGs and 100 permutations).

We did, however, test the power of Margarita on a small number of SNPs and individuals. To our surprise, the power of Margarita was not as good as the standard χ^2 test under several of the scenarios that we studied: Since Margarita assumes the coalescent model we tested it on null data produced using a coalescent model, upon which phenotypes have then been randomly assigned (i.e., in which there is no causal SNP). We generated 100 different panels of size 25 kb using the ms software (Hudson), which simulates data under the coalescent model. In these experiments we generated 50 individuals, which were randomly assigned to be cases or controls. We found that the false positive rate (using a p-value cutoff of 0.05) was 1%. We used the following strong association model – one of the SNPs was arbitrarily set as a strong causal SNP, by declaring an individual to be a case if the corresponding SNP is either heterozygous or homozygous 1, and control otherwise. Using this model, we observed that the power of Margarita is 17%, while the power that is obtained by a standard permutation test was much higher: 69%.

We repeated the same experiments described above, by simulating additional panels of SNPs using the ms software with recombination events. We found that the false positive rate in this case is 89%. The power for the strong association model was 69% compared with 75% obtained by the standard permutation test.

These results show that there are serious problems with existing methods either with respect to running time or power. Indeed, our results show that these methods are dominated in the scenarios that we studied by the standard χ^2 approach. We thus used the standard χ^2 approach as a baseline in our experimental study of CAMP. We also compared to CLADHC[10], as CLADHC is computationally feasible in our scenarios.

3.3 Power

To study the power of the methods we applied them to case-control panels using the multiplicative model for generating cases and controls described in Section 3.1. A panel is generated from the whole sequence, and then the SNPs are sampled according to the specified sparsity (i.e., density of the sampled SNPs). Therefore, in many cases (but not always) the causal SNP does not exist in the SNP sampled panel. Under a specific significance level (which controls the type I error), the *relative power* is defined as a ratio in which the numerator is the number of SNP sampled panels defined to be significant and the denominator is the number of original panels with the entire sequence, including the causal SNP, defined to be significant.

We simulated panels with 1,000 cases and 1,000 controls. We fixed the sparsity of the sampled SNPs to each of the following values: $(1,000, 2,000, \ldots, 10,000)$. For each value of sampling sparsity 5,000 different panels were generated. Results for the different SNP sampling densities are presented in Figure 2, and for different numbers of cases and controls in Figure 3. The difference in the relative power between CAMP and standard χ^2 testing reaches more than 10%. An even

more prominent difference is observed between CAMP and CLADHC, ranging up to 52.7%.

3.4 Localization of the Causal SNP

We also tested the accuracy in localizing the causal SNP. To obtain an estimate of location from the output of CAMP, note first that CAMP may output more than one SNP. These SNPs presumably represent mutations on the genealogical tree near the causal SNP; thus, we use the average location of these called SNPs to estimate the position of the causal SNP. A comparison to the χ^2 test is presented in Figure 4 for different numbers of individuals. The advantage of CAMP over the standard χ^2 is quite notable; e.g., for 3,000 controls and cases, the percentage of panels for which the distance between the found location and the true causal SNP is below 100Kb was 86% for CAMP and 79% for the standard χ^2.

3.5 Measuring the Advantage of the Coalescent Approach

Our algorithm tests a subset of all possible interactions. This subset, as described before, is determined according to the approximated genealogical relations between the SNPs. Does testing all possible interactions within the linkage upper bound gives similar results? To answer this question we compared CAMP to a procedure that tests all pairwise interactions less than the linkage upper bound. For two SNPs, this corresponds to testing the association between the haplotypes generated by the SNPs and the phenotype, which is calculated by a standard χ^2 test for these two vectors.

The results are presented in Figure 5. As can be seen, CAMP yields significantly greater power than the procedure that tests all pairwise interactions. The difference between these procedures is much larger than the difference between CAMP and the standard χ^2 algorithm, reaching more than 50%.

3.6 Using the Phased Haplotypes

Since CAMP uses phased data to construct the coalescent graph, we tested the effect of phasing errors on our algorithm, considering phasing error rates of 3%, 30% and 50%. The value of 3% corresponds roughly to the error rate reported in the literature for phasing algorithms[16] and 50% corresponds to randomly phasing each one of the heterozygous sites.

Results are presented in Figure 6. As can be observed in the graph, even when the phasing error rate is 50% (which is very unlikely) CAMP has a relative power that is 9% larger than the standard χ^2 test (for a significance level of 5%). With phasing error rates of 3% and even 30%, no significant reduction in the power is observed.

4 Discussion

We have presented a method that leverages the coalescent model to conduct association mapping in whole-genome association studies. We exploit the unobserved genealogy of the chromosomes in order to evaluate more accurately the significance and location of causal SNPs. The genealogy defines a set of haplotypes, and our method consists of a strategy for the selection of tests based on these haplotypes and on the genealogy. As we have demonstrated, selecting these tests carefully gives a large advantage in the power and in the localization of the causal SNP. We have also shown that several existing methods that aimed to address this problem either suffer from low power or suffer from a very high false positive rate when compared to a standard approach in which each of the SNPs is tested separately with a χ^2 test. We have also shown that considering all SNP interactions reduces the power considerably.

Interestingly, we observed that introducing very high rates of phasing errors (30%) does not reduce the power of our method. This can be explained by the fact that when enough individuals are given, the genealogical relationship between SNPs can be determined accurately, even if the heterozygous sites are ignored. The signal is strong enough in the homozygous sites so that phasing accuracy has a minor effect on the results.

Accurate localization of the causal SNP is as important as significance estimation. We have shown that CAMP estimates the location of the real causal SNP more accurately than other methods. Note, moreover, that this was achieved by a relatively naive approach of taking the average of the interacting SNPs. Most likely, the location can be determined even more accurately by a more sophisticated algorithm that uses properties of the coalescent.

There are several important issues that need further attention. We have shown that CAMP is more powerful than the standard χ^2 test, but we have not shown its optimality. The question of whether there exists a more powerful algorithm or strategy for choosing a subset of interactions of SNPs

should be explored. In particular, it is intriguing to study the following optimization problem: For a fixed disease model (say, the multiplicative model with given penetrance and relative risk) and significance level, find the strategy that determines which SNP interactions are tested, such that the power is maximized.

It is also important to study the generalizations of methods such as CAMP to the case where multiple populations may be participating in the study. One of the challenges in drawing causal inferences from whole-genome case-control association studies is the confounding effect of population structure [17,18,19,20,21,22,23,24]. This issue has received much attention in recent years in the literature (e.g., [25,26,24]). Currently, CAMP assumes no stratification effect, i.e., the controls and cases are presumed to be from one population. There is a clear need to explore methods for taking population stratification into account in CAMP.

Although not tested experimentally in this work, since CAMP performs transformation on the SNP data, it can be naturally extended to handle other types of phenotypes such as continuous traits (QTLs).

The cost of genotyping is continually decreasing and technology is evolving towards genome resequencing (e.g., Illumina/Solexa 1G and Roche/454). However, it is still very expensive to conduct resequencing of the whole genome as a tool for association studies. As a consequence, it is clear that genotype data of common genetic variants such as SNPs will be the leading approach in association studies in the coming years. Algorithms such as CAMP that yield high statistical power by exploiting aspects of genealogy will play an important role in the analysis of these data.

Acknowledgments

G.K., E.H. and R.M.K. were supported by NSF grant IIS-0513599. M.I.J. was supported by a grant from Microsoft Research and by an appointment as a Miller Research Professor in the Miller Institute for Basic Research in Science.

Web Resources

- Hapmap project, http://www.hapmap.org
- Illumina/Solexa, http://www.illumina.com
- Roche/454, http://www.454.com

References

1. Consortium TWTCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–683

2. Bonnen PE, Pe'er I, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly MJ, Reich DE, et al. (2006) Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat Genet 38:214–217

3. Slager SL, Huang J, Vieland VJ (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genet Epidemiol 18:143–156

4. Kimmel G, Shamir R (2006) A fast method for computing high significance disease association in large population-based studies. Am J Hum Genet 79:481–492

5. Zollner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071–1092

6. Minichiello MJ, Durbin R (2006) Mapping trait loci by use of inferred ancestral recombination graphs. Am J Hum Genet 79:910–922

7. Nordborg M (2001) Handbook of Statistical Genetics. John Wiley and Sons, inc.

8. Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comp Biol 3:479–502

9. Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318

10. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P (2004) Linkage disequilibrium

mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43

11. Eskin E, Halperin E, Karp RM (2003) Large scale reconstruction of haplotypes from genotype data. In Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 03). The Association for Computing Machinery, 104–113

12. Gusfield D (2002) Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB 02). The Association for Computing Machinery, 166–175

13. Bafna V, Gusfield D, Lancia G, Yooseph S (2003) Haplotyping as perfect phylogeny: A direct approach. J Comput Biol 10(3-4):323–340

14. Kimmel G, Shamir R (2005) The incomplete perfect phylogeny haplotype problem. J Bioinform Comput Biol 3(2):359–384

15. Li N, Stephens M (2003) Modelling linkage disequilibrium and identifying recombinations hotspots using SNP data. Genetics 165:2213–2233

16. Kimmel G, Shamir R (2005) GERBIL: Genotype resolution and block identification using likelihood. P Natl Acad Sci USA 102:158–162

17. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, et al. (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388–393

18. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37:1243–1246

19. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

20. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37:90–95

21. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37:868–872

22. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

23. Lohmueller K, Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182

24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

25. Kimmel G, Jordan M, Halperin E, Shamir R, Karp R (2007) A randomization test for controlling population stratification in whole-genome association studies. Am J Hum Genet 81:895–905

26. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Figures Legends

Figure 1:

An example of a perfect phylogeny tree. Each node corresponds to a haplotype. The mutations appear on the edges. (a) A perfect phylogeny with five SNPs. (b) An additional sixth SNP that was mutated and can be expressed as the interaction between SNPs 4 and 5.

Figure 2:

The relative power for different sampling distances of SNPs and for four different significance levels.

Figure 3:

Relative power for different number of individuals.

Figure 4:

A comparison of the cumulative distribution functions for the distance between the discovered SNP and the true causal SNP. The three figures represent the results obtained from different numbers of controls and cases: A - 1,000, B - 2,000, C - 3,000.

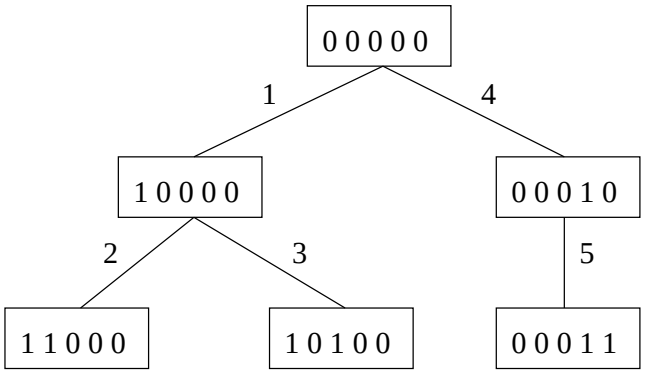
Figure 5:

A comparison of CAMP to a naive pairwise interactions algorithm, in which all pairwise interactions of SNPs with distances smaller than the threshold used in CAMP are tested.

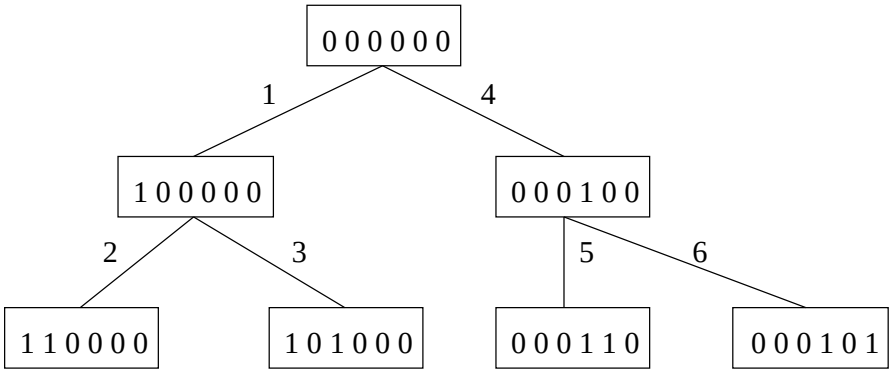
Figure 6:

A comparison of the relative power of CAMP in the presence of different rates of phasing errors.

Figures

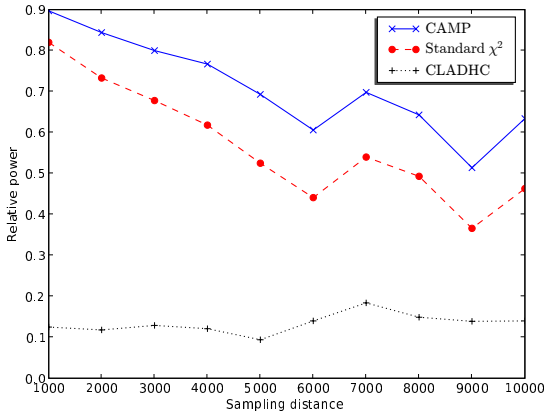


(a)

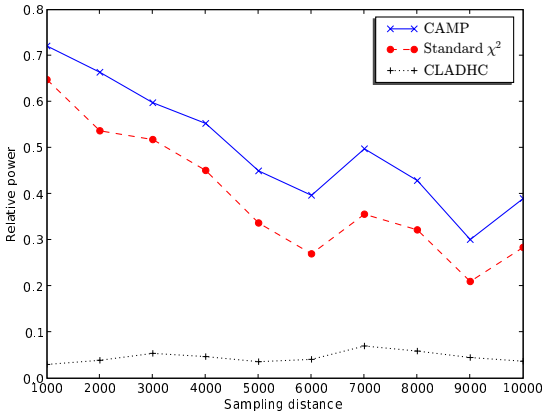


(b)

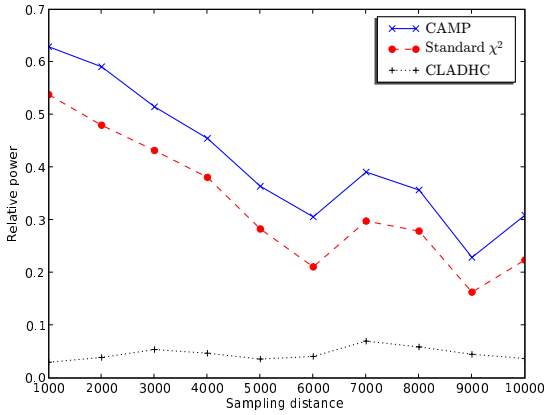
Figure 1:



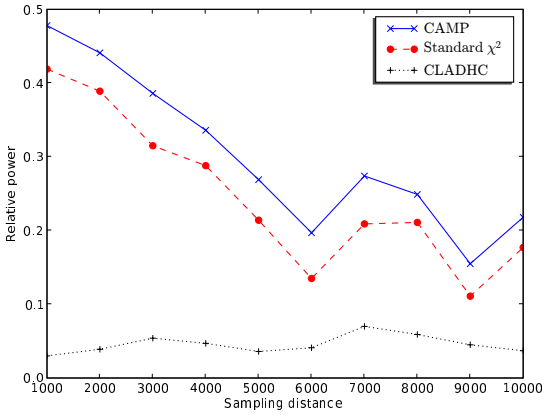
Type I error = 0.05



Type I error = 0.01



Type I error = 0.005



Type I error = 0.001

Figure 2:

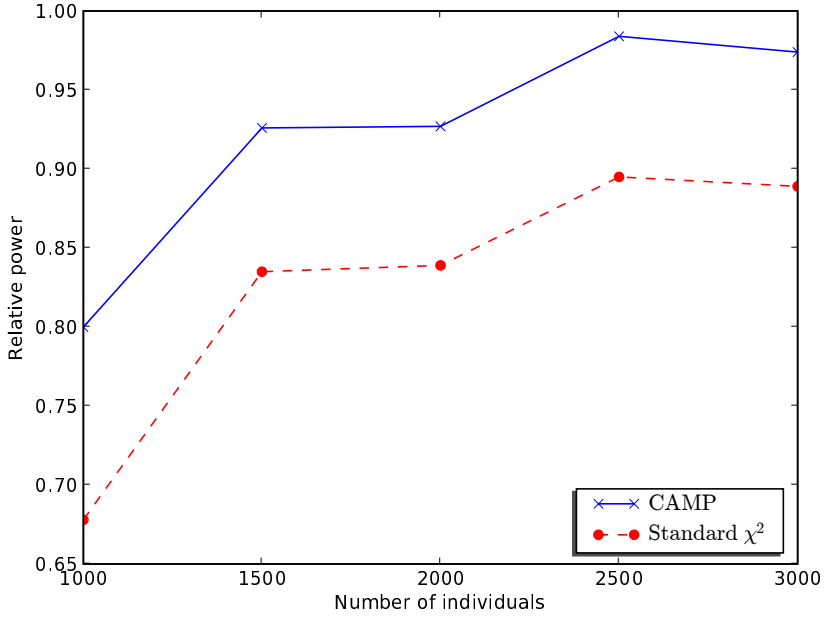
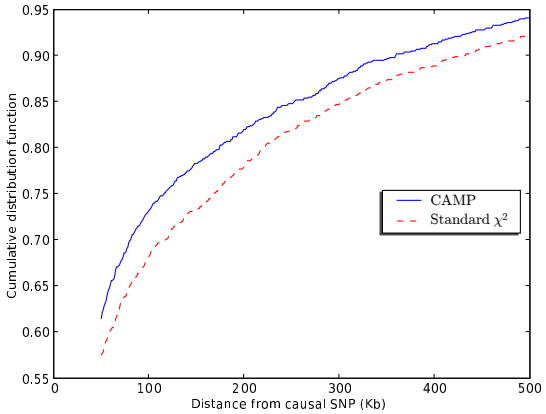
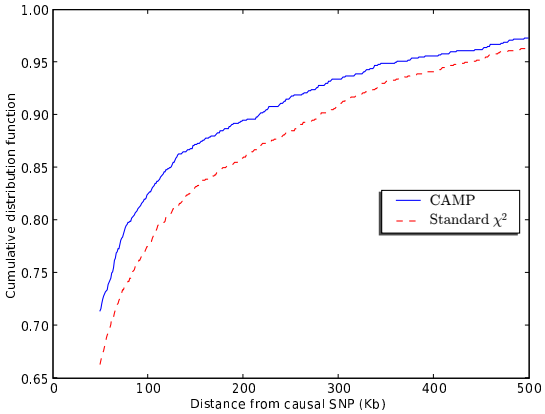

1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Relative power
1000
1500
2000
2500
3000
Number of individuals
CAMP
Standard χ^2

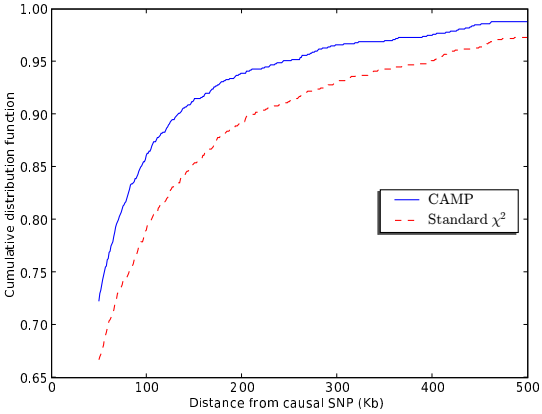
Figure 3:


Cumulative distribution function
Distance from causal SNP (Kb)
CAMP
Standard χ^2

A


Cumulative distribution function
Distance from causal SNP (Kb)
CAMP
Standard χ^2

B


Cumulative distribution function
Distance from causal SNP (Kb)
CAMP
Standard χ^2

C

Figure 4:

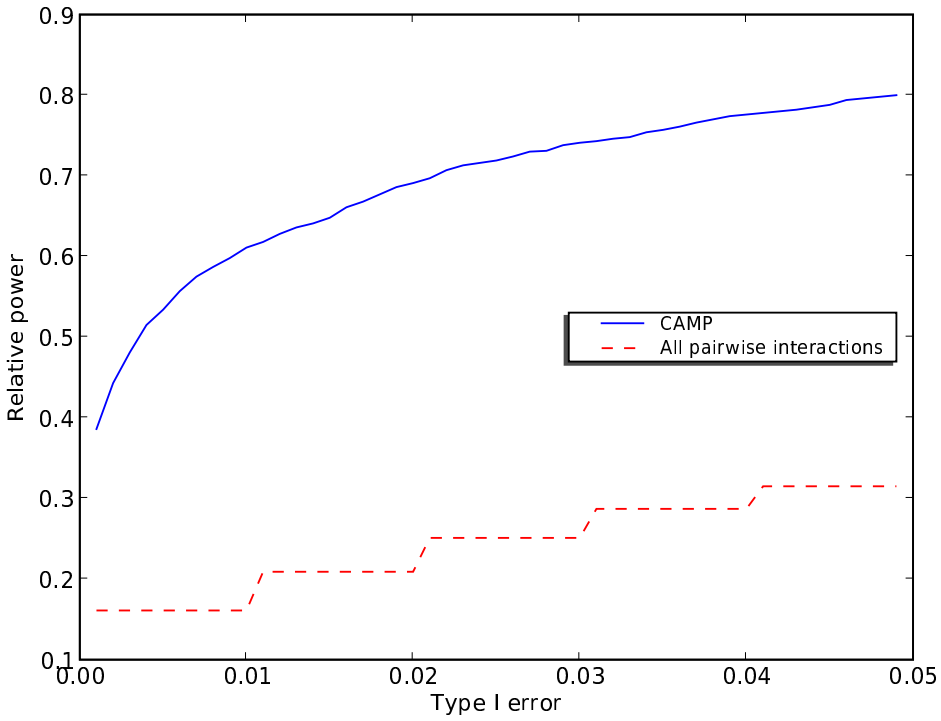

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Relative power
0.00
0.01
0.02
0.03
0.04
0.05
Type I error
CAMP
All pairwise interactions

Figure 5:

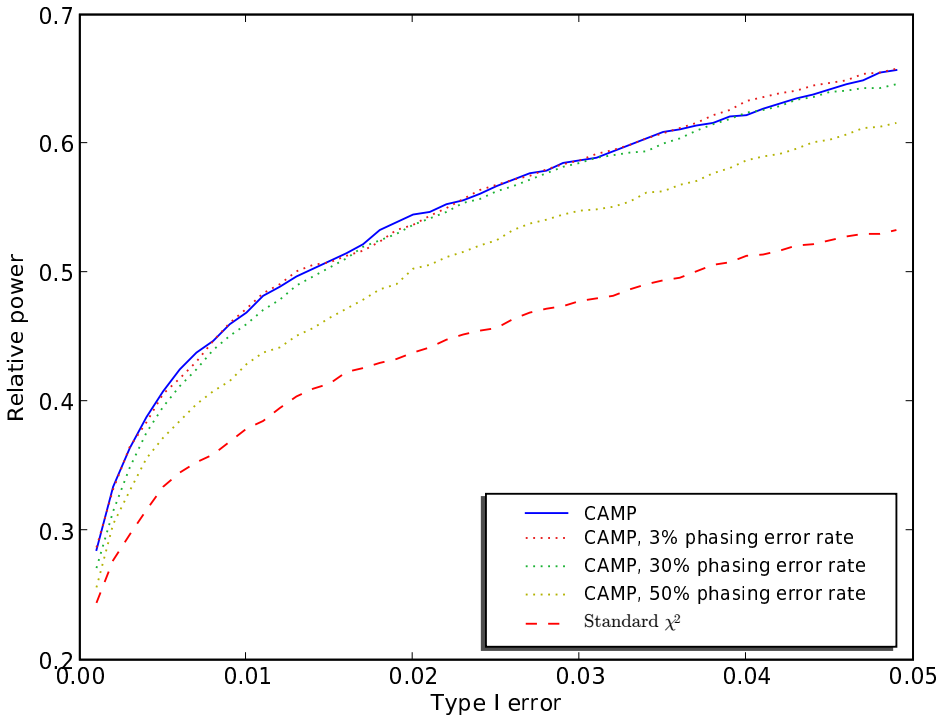

0.7
0.6
0.5
0.4
0.3
0.2
Relative power
0.00
0.01
0.02
0.03
0.04
0.05
Type I error
CAMP
CAMP, 3% phasing error rate
CAMP, 30% phasing error rate
CAMP, 50% phasing error rate
Standard χ^2

Figure 6: