BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 19 2008, pages 2215–2221
doi:10.1093/bioinformatics/btn406

Genetics and population analysis

Efficient whole-genome association mapping using local phylogenies for unphased genotype data

Zhihong Ding[1], Thomas Mailund[2,*] and Yun S. Song[3]

[1]Department of Computer Science, University of California, Davis, USA, [2]Bioinformatics Research Center, University of Aarhus, Denmark and [3]Computer Science Division and Department of Statistics, University of California, Berkeley, USA

Received on May 1, 2008; revised on July 25, 2008; accepted on July 29, 2008

Advance Access publication July 30, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Recent advances in genotyping technology has made data acquisition for whole-genome association study cost effective, and a current active area of research is developing efficient methods to analyze such large-scale datasets. Most sophisticated association mapping methods that are currently available take phased haplotype data as input. However, phase information is not readily available from sequencing methods and inferring the phase via computational approaches is time-consuming, taking days to phase a single chromosome.

Results: In this article, we devise an efficient method for scanning unphased whole-genome data for association. Our approach combines a recently found linear-time algorithm for phasing genotypes on trees with a recently proposed tree-based method for association mapping. From unphased genotype data, our algorithm builds local phylogenies along the genome, and scores each tree according to the clustering of cases and controls. We assess the performance of our new method on both simulated and real biological datasets.

Availability: The software described in this article is available at http://www.daimi.au.dk/~mailund/Blossoc and distributed under the GNU General Public License.

Contact: mailund@birc.au.dk

1 INTRODUCTION

Utilizing the current chip technology, genome-wide scans with hundreds of thousands of single nucleotide polymorphisms (SNPs) in thousands of individuals is affordable for large-scale association studies (Barrett and Cardon, 2006). Several genome-wide studies have already been published (Amundadottir *et al.*, 2006; Arking *et al.*, 2006; Smyth *et al.*, 2006; The Wellcome Trust Case Control Consortium, 2007) and as the genotyping price continues to drop, we expect to see many more in the near future. With datasets of such sizes, the need for efficient, accurate association mapping methods is evident. Many studies resort to a marker-by-marker approach—e.g. a simple Fisher's exact test or χ^2-test—but, unless the trait-influencing variants are typed, its power is limited by the indirect testing through linkage disequilibrium (LD), and multi-marker approaches are generally preferred (Pe'er *et al.*, 2006). A trade-off must be made, however, between sophistication and computational tractability.

Recently, one of us has developed a new multi-SNP method called BLOSSOC (Mailund *et al.*, 2006) that, although of similar accuracy, is orders of magnitude faster than other multi-SNP methods, and is capable of analyzing whole-genome data in a few CPU hours. BLOSSOC resembles other recent methods, e.g. that of Zöllner and Pritchard (2005), Minichiello and Durbin (2006) and Clark *et al.* (2007). It constructs local tree-like genealogies along the genome and scores those genealogies according to how the cases and controls are clustered, the motivation being that near a disease-predisposing SNP, the cases will tend to cluster together in the underlying genealogy. Compared with the other methods, BLOSSOC achieves its much faster running time by taking a simpler approach to how local trees are constructed. Instead of sampling local trees from the coalescent with recombination and averaging scores over the sampled trees, it relies on a deterministic, efficient algorithm to build a single tree for each locus, assuming the infinite-sites model of mutation. Sevon *et al.* (2006) have recently proposed a method that also constructs a single tree per locus. Their approach differs from BLOSSOC in how local trees are constructed and scored. Whereas Sevon *et al.* (2006) use a time-consuming permutation test to score trees, BLOSSOC considers each tree as a decision tree and scores it with standard methods from the data mining literature (see Mailund *et al.*, 2006 for details). Tachmazidou *et al.* (2007) construct local trees using the same approach as BLOSSOC, but score them using a sophisticated MCMC algorithm that is relatively time consuming.

As a consequence of its simple approach to tree construction and scoring, BLOSSOC is very computationally efficient. Further, computer experiments shown in Mailund *et al.* (2006) indicate that this efficiency is achieved with little, if any, loss in accuracy compared with more sophisticated methods. However, a major limitation of the original version of BLOSSOC, as is also true for other methods, is its reliance on having phased haplotype data. Even with FASTPHASE (Scheet and Stephens, 2006), phasing a whole-genome dataset requires tens of days of CPU time, making this step the major bottleneck when using computationally efficient methods such as BLOSSOC for analysis.

In this article, we devise a method that eliminates the need for preanalysis phasing of genotypes into haplotypes. Our approach combines the ideas in BLOSSOC with a recently found linear-time algorithm (Ding *et al.*, 2005, 2006) for phasing genotypes on trees.

*To whom correspondence should be addressed.

This way, our new method builds local phylogenies directly from unphased data. Inferring trees from unphased genotype data is slightly slower than inferring trees from phased haplotype data, but our method is still capable of scanning the entire human genome in a few days. We also develop a new Bayesian score for the association between a local tree and the disease phenotype. Using simulated datasets, we compare the mapping results for unphased data with that for the true haplotype data and show that there is little loss in accuracy or ranking quality. As a proof of principle, we apply our method to analyze a genome-wide dataset for Parkinson disease (Fung *et al.*, 2006). We remark that our tree construction algorithm is not restricted to BLOSSOC; other association mapping methods based on scoring local genealogies, such as Sevon *et al.* (2006) and Tachmazidou *et al.* (2007), may also be generalized in the same way, enabling them to analyze unphased data.

2 ALGORITHM

At a given position in the genome, a sample of chromosomes will be related by a genealogical tree. Consider a polymorphic site in the genome and its corresponding tree. If a mutation that occurred at that site affects the probability that an individual has a disease, then it induces a non-random distribution of cases and controls at the leaves of the corresponding tree. Therefore, one approach to find a location in the genome that harbors a disease-predisposing mutation is to test for a significant clustering of affected individuals in local trees. In practice, however, local trees are not known and can only be partly inferred from molecular genetic data. Previous methods of inferring local genealogies and testing for association include Mailund *et al.* (2006); Minichiello and Durbin (2006); Wu (2007); Zöllner and Pritchard (2005). In what follows, we describe a novel, efficient algorithm that can build and score local trees directly from unphased genotype data. An outline of our algorithm is shown in Figure 1.

2.1 Building local phylogenies

We infer local trees in two different ways, depending on whether a perfect phylogeny—i.e. a genealogy consistent with the infinite-sites mutation model without recombination—exists for a sufficiently wide region around the local site we wish to score. When building trees, we require at least *m* markers be used in the inference, where *m* is an option to the BLOSSOC program. In general, we recommend using larger (respectively, smaller) values of *m* in regions with high (respectively, low) LD. When *m* markers are compatible with the infinite-sites and no recombination assumption, we say that a *perfect phylogeny* exists, and we can construct trees directly from unphased data. When the *m* markers are incompatible with the assumptions, we perform a local phase inference for the *m* markers only and construct a tree using heuristics from Mailund *et al.* (2006).

Given a set *G* of *n* genotypes of the same length, the Perfect Phylogeny Haplotyping (PPH) problem is to find *n* pairs of haplotypes that explain *G* and fit a perfect phylogeny. Vijayasatya and Mukherjee (2005) and Ding *et al.* (2005, 2006) have independently developed linear-time algorithms for this problem. In this article, we utilize the latter algorithm, called LPPH.

```
Input: Set of genotypes $G$ and their disease
status, user specified number $m$
Output: Likelihood scores for each marker

For each marker $i$
  Find the largest interval $I$ around
     marker $i$ such that genotypes in $I$ have
     a PPH solution
  If the size of $I$ is at least $m$, then
     build a local tree $T$ for genotypes in
     $I$ using the LPPH algorithm
  If the size of $I$ is less than $m$
     Add neighboring markers to $I$ until
       its size equals $m$
     Use the entropy minimization algorithm
       to infer the phase of genotypes in $I$
     Build a local tree $T$ for the
       haplotypes in $I$
  Score $T$ and output the score as the
     score for marker $i$
```

Fig. 1. An outline of our algorithm. PPH, 'Perfect Phylogeny Haplotyping'; LPPH, 'Linear time Perfect Phylogeny Haplotyping'. See the main text for details.

Our algorithm constructs a local tree for each marker, incorporating neighboring markers as follows. To construct a local tree for a marker *x*, initialize *X* to be the set containing only *x*. Then, alternate the following two steps until neither is possible:

(1) If *X* and the next marker immediately to the *left* together admit a PPH solution, then add that marker to *X*.

(2) If *X* and the next marker immediately to the *right* together admit a PPH solution, then add that marker to *X*.

Selecting a set of markers in this way keeps the focal marker *x* near the center of the region.

If the resulting *X* contains at least *m* markers, where *m* is specified by the user, we score the tree as described in the next section. However, for a large sample size and high recombination rate, the size of regions for which PPH solutions exist tends to be very small. Such a region will contain only few markers and consequently the corresponding inferred tree will contain only few edges, making it difficult to infer reliably the clustering of cases and controls. A region with a set of *m* markers that does not admit a PPH solution is called an *incompatible region*. For such a region, we use an entropy minimization algorithm—previously considered by Halperin and Karp (2005) and Gusev *et al.* (2007)—to locally infer the phase of input genotypes and use the tree building algorithm described in Mailund *et al.* (2006) to build local phylogenies from the inferred haplotypes. The entropy minimization algorithm has phasing accuracy slightly worse than that of other widely-used methods such as FASTPHASE (Scheet and Stephens, 2006) and HAP (Halperin and Eskin, 2004), while being several orders of magnitude faster (Gusev *et al.*, 2007).

Given a haplotype *h* and a phasing solution ϕ for a set of *n* genotypes in *G*, define the *coverage of h under* ϕ, denoted by COV(h,ϕ), as the number of genotypes in *G* that are phased by *h* and some other haplotype in ϕ, plus twice the number of genotypes in *G* that are phased by the haplotype pair (h,h). For a fixed phasing solution ϕ, the sum of $\mathrm{COV}(h,\phi)$ over all haplotypes in ϕ is equal to $2n$.

As in Halperin and Karp (2005) and Gusev *et al.* (2007), we define the *entropy* of a phasing solution ϕ as

$$H(\phi)=\sum_{h:\mathrm{COV}(h,\phi)\neq 0} -\frac{\mathrm{COV}(h,\phi)}{2n}\log\frac{\mathrm{COV}(h,\phi)}{2n}.$$

Halperin and Karp (2005) first introduced the problem of finding a phasing solution ϕ of G with the minimum entropy, and Gusev *et al.* (2007) later developed an accurate and highly efficient algorithm to solve the problem. Intuitively, the entropy of a phasing solution ϕ is a measure of haplotype diversity in ϕ. As is shown in Gusev *et al.* (2007), if the probability of a haplotype h is estimated by counting the number of times that h appears in a phasing ϕ, then maximizing the log-likelihood of phasing ϕ is equivalent to minimizing the entropy of phasing ϕ.

The phasing algorithm in Gusev *et al.* (2007) employs a set of short overlapping windows to phase long genotypes. In our work, we typically deal with short genotypes and hence use the following simple version of that algorithm:

```
(1) Generate a random phasing solution $\phi$ for genotypes $G$.
(2) Repeat the following:
    (a) Find the pair $(g, (h_1, h_2))$ such that $H(\phi')$ is minimized,
        where $\phi'$ is obtained from $\phi$ by re-explaining $g \in G$ with $(h_1, h_2)$.
    (b) If $H(\phi') < H(\phi)$, then let $\phi = \phi'$, else exit the loop.
(3) Output phasing solution $\phi$.
```

Missing data are handled as follows. When using our algorithms to obtain a phasing solution, we first phase those genotypes without any missing entries and calculate the frequency of each distinct haplotype in a solution ϕ' to that subproblem. Then, for each genotype g with one or more missing entries, we test whether the non-missing entries in g can be phased using a haplotype in ϕ' and some other haplotype. If more than one haplotype in ϕ' satisfies this criterion, we choose the one with a higher frequency. (If there is a tie, we choose an arbitrary one among them.) Missing entries in g are then imputed according to that phasing. If none of the haplotypes in ϕ' can be used to phase g in this way, then g is resolved arbitrarily.

2.2 Scoring local phylogenies

Once a local phylogeny is constructed, it is scored according to how well the tree helps explain the phenotype. We consider the tree a hierarchical clustering of chromosomes: each partition of the tree into subtrees defines a clustering. To each cluster c, we assign a disease risk θ_c—assumed to be independent of other clusters—and the disease status of a chromosome in cluster c is modeled as being affected with probability θ_c and unaffected with probability $1-\theta_c$.

Given a clustering $\mathcal{C}=\{c_1,\ldots,c_n\}$ and corresponding disease risks $\Theta=\{\theta_1,\ldots,\theta_n\}$, the likelihood of the observed disease status is given by

$$L(\mathcal{C},\Theta)=\prod_{i=1}^{n}\theta_i^{A_i}(1-\theta_i)^{U_i},$$

where A_i denotes the number of affected leaves in cluster c_i and U_i the number of unaffected leaves in cluster c_i. Assigning a risk per cluster ignores that phenotype risk may be a function of genotypes, and thus a function of two clusters rather than one. It is a straightforward extension to model phenotype as a function of genotypes, although the implementation of the score functions would then get somewhat more involved.

When scoring a tree, the clustering and risks are nuisance parameters that we integrate out in a Bayesian approach. For the cluster specific risks, we follow the approach in Waldron *et al.* (2006) and choose independent uninformative β-priors $\pi(\theta)=1$ and obtain

$$L(\mathcal{C})=\prod_{i=1}^{n}\int_0^1\theta^{A_i}(1-\theta)^{U_i}\pi(\theta)\mathrm{d}\theta=\prod_{i=1}^{n}B(A_i+1,U_i+1),$$

where B is the β function. To integrate out $\mathcal{C}$, we again choose a uniform prior on clusters, obtaining the following final score:

$$L=\frac{1}{|\mathcal{C}|}\sum_{\mathcal{C}}L(\mathcal{C}),$$

where $|\mathcal{C}|$ denotes the total number of clusterings possible in the tree. Summing over all clusterings is computationally prohibitive, so in our implementation we only sum over all sub-trees and cluster the chromosomes into only two clusters, those that are leaves in the sub-tree and those that are outside. When there are more than one perfect phylogeny consistent with a region, we average the scores over all trees for that region. This corresponds to integrating over the unknown tree with a flat prior over topologies.

As a Bayes factor, any number greater than 1 can be taken as evidence for association, while any number smaller than 1 can be taken as evidence against. However, some scores higher than 1 can still occur by chance, so we recommend doing permutation tests to judge the true significance of a score.

3 RESULTS

In this section, we evaluate our new algorithm in comparison with previous methods. We first consider simulated data in which the true disease-predisposing locus is known. When comparing with methods that require phased data, we use the true phased data obtained from simulation, thus yielding the best performance quality that those methods can achieve. We then analyze the genome-wide dataset from a Parkinson disease study described in Fung *et al.* (2006). This is admittedly a small genome-wide dataset, with only 267 cases and 270 controls. As a proof of concept, however, it demonstrates the scalability of our method.

3.1 Ranking experiments

In genome-wide association studies, false positives are a major problem, meaning that potentially many leads may need to be followed before a true hit can be found. Hence, the quality with which a method ranks the true hits compared to spurious hits is one of the most important measures of performance.

In this section, we describe how well BLOSSOC performs in ranking. We simulated 100 case/control datasets, each with 1000 cases and 1000 controls, under an additive disease model with varying genetic relative risk. With A denoting the disease-predisposing allele and a the wild-type, the genotype relative risk of the heterozygote Aa is denoted by GRR. Since we assumed an additive model in our simulations, the GRR of the mutant homozygote AA was $2\times\mathrm{GRR}-1$. Experiments with dominant or

Table 1. Ranking experiment results showing the average number of top-10 ranked markers within a given distance of the true disease-predisposing locus

ρ	WR	GRR	1 kb			10 kb			100 kb		
			χ^2	P	U	χ^2	P	U	χ^2	P	U
100	5%	1.2	0.08	0.11	**0.12**	0.63	0.76	**0.86**	4.88	**4.89**	4.58
100	5%	1.4	0.06	**0.09**	0.08	0.66	0.75	**0.82**	4.97	5.36	**5.47**
100	5%	1.6	0.13	0.13	0.13	0.97	**1.43**	1.24	5.34	**6.14**	5.99
100	5%	1.8	0.10	0.21	**0.22**	1.14	1.75	**1.78**	6.04	**6.51**	6.40
100	5%	2.0	0.20	**0.21**	0.19	1.26	**1.59**	1.56	5.82	6.11	**6.20**
100	10%	1.2	**0.12**	0.11	0.09	0.70	**0.71**	0.69	4.42	4.50	**4.72**
100	10%	1.4	**0.09**	0.05	0.06	0.85	0.81	**0.90**	**5.37**	5.29	5.20
100	10%	1.6	0.12	0.16	**0.18**	1.25	1.83	**1.84**	6.02	**6.90**	6.84
100	10%	1.8	0.09	**0.14**	0.12	1.03	**1.64**	1.62	5.50	**6.58**	6.52
100	10%	2.0	0.07	0.16	**0.17**	1.06	**1.80**	1.70	5.82	**6.62**	6.54
400	5%	1.2	0.01	0.01	**0.03**	0.37	0.40	**0.43**	2.25	2.46	**2.67**
400	5%	1.4	0.04	**0.05**	0.03	0.40	**0.47**	0.36	2.45	2.78	**2.81**
400	5%	1.6	0.01	**0.03**	0.02	0.55	**0.73**	0.69	2.78	**3.54**	3.26
400	5%	1.8	0.03	0.05	**0.06**	0.50	**0.80**	0.63	2.89	**4.02**	3.74
400	5%	2.0	0.07	**0.09**	**0.09**	0.64	**0.95**	0.86	3.24	**4.61**	4.27
400	10%	1.2	**0.07**	0.03	0.03	0.24	**0.30**	0.27	1.77	2.23	**2.27**
400	10%	1.4	0.13	0.15	**0.16**	0.44	**0.66**	0.61	2.75	**3.46**	3.29
400	10%	1.6	0.07	**0.12**	0.11	0.68	0.82	**0.83**	2.89	3.26	**3.42**
400	10%	1.8	**0.10**	0.08	0.07	0.60	**0.72**	0.61	3.32	**4.18**	4.10
400	10%	2.0	0.10	**0.14**	**0.14**	0.81	1.19	**1.20**	3.58	**4.84**	4.70

Based on 100 simulated datasets each with 1000 case and 1000 control individuals. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to BLOSSOC results using $m=5$ for phased (unphased, respectively) data. GRR, 'Genetic Relative Risk'; WR,'Wildtype Risk'.

recessive (instead of additive) disease models produce similar results (results not shown). The wild-type risk (denoted WR in Tables 1 and 2) of being diseased was varied between 5% and 10%, while GRR $=$ 1.2, 1.4, 1.6, 1.8 and 2.0 were used. The SNP density was also varied; we used 100 SNP markers with the population-scaled recombination rates $\rho=4{,}N_ec=100$ and 400, where N_e denotes the effective population size and c the recombination rate per generation per sequence. Assuming $N_e=10000$ and a recombination rate of 10^{-8} per adjacent pair of sites per generation, $\rho=100$ and 400 correspond to 250 kb and 1 Mb, respectively. Sequences were simulated using the coalescence based simulator COASIM (Mailund *et al.*, 2005) with the infinite-sites mutation model, and were then paired up randomly to construct diploid individuals. After assigning disease status, the disease-predisposing SNP was removed from the data.

We performed analysis both using the original phase-known version of BLOSSOC (Mailund *et al.*, 2006) on the true phased data, and using our new algorithm on unphased data. We first studied how the performance of BLOSSOC depends on the minimum number m of markers to include when building a tree. In general, larger values of m should be used for regions with high LD and smaller values for regions with low LD. In our experiment, we tried using $m=3,5$ and 10, and found $m=5$ to be slightly superior to $m=3$ and $m=10$, for both $\rho=100$ and $\rho=400$ (results not shown). In what follows, we therefore use only $m=5$.

As a way of summarizing the ranking results, we considered the mean fraction of top-10 ranked markers found within 1, 10 or 100 kb of the disease-predisposing marker location. See Table 1 for results. The results for an 2×2 allelic χ^2-test, applicable to an additive disease model, are also shown there. For each parameter setting, we have highlighted the best performing method in bold print. Table 2 shows the fraction of datasets with at least one top-10 ranked marker within 1, 10 or 100 kb of the disease-predisposing marker. Measured this way, we see that, close to the true disease-predisposing marker, the phase-known BLOSSOC method performs the best, followed by our new phase-unknown BLOSSOC method, and then the χ^2-method. Farther, away from the disease-predisposing marker (the 100 kb columns), however, the χ^2-method generally outperforms both BLOSSOC methods. An explanation for this is the higher correlation between neighboring markers in the BLOSSOC method. One spurious signal will tend to give spurious signals at neighboring markers, and a high scoring marker far from the true locus can therefore move most (or all) of top-10 away from the true locus. Consequently, considering a raw ranking—ignoring the correlation between markers—is probably not optimal for BLOSSOC. Future work will include ways of ranking BLOSSOC scores in a more meaningful way.

3.2 Localization experiments

Another important criterion for assessing the performance of a fine-mapping method is localization; that is, how accurately the position of an untyped disease-predisposing SNP can be estimated. In our localization study, we simulated case-control datasets using COASIM for a region corresponding to 100 kb physical distance (or $\rho=40$). We assumed that the region contains exactly one disease-predisposing SNP. We used GRR $=1.4, 1.6, 1.8$ and 2.0, and assumed that the wild-type risk is 5%. We simulated 100 datasets with 500 case and 500 control individuals, as well as 100 datasets with 2000 case and 2000 control individuals.

Table 2. Ranking experiment results showing the fraction of data sets with at least one top-10 marker within a given distance of the true disease-predisposing locus

ρ	WR	GRR	1 kb			10 kb			100 kb		
			χ^2	P	U	χ^2	P	U	χ^2	P	U
100	5%	1.2	**0.13**	**0.13**	0.12	0.42	**0.46**	0.39	**0.98**	0.90	0.86
100	5%	1.4	0.15	**0.22**	0.20	0.71	**0.74**	0.70	**0.99**	0.95	0.98
100	5%	1.6	0.27	**0.29**	0.28	0.73	**0.82**	0.80	0.99	0.99	0.99
100	5%	1.8	0.19	**0.30**	0.28	0.79	**0.83**	0.80	**1.00**	0.98	0.98
100	5%	2.0	0.19	**0.30**	0.28	0.81	**0.85**	0.84	1.00	1.00	1.00
100	10%	1.2	0.11	0.11	**0.14**	**0.59**	0.52	0.51	**0.98**	0.94	0.96
100	10%	1.4	0.24	**0.27**	**0.27**	**0.73**	0.66	0.63	**0.98**	0.96	0.95
100	10%	1.6	0.16	**0.26**	0.24	0.74	**0.82**	0.73	**1.00**	0.98	0.99
100	10%	1.8	0.23	**0.30**	**0.30**	**0.85**	**0.85**	0.83	1.00	1.00	1.00
100	10%	2.0	0.28	0.42	**0.43**	0.85	0.86	**0.91**	1.00	1.00	1.00
400	5%	1.2	**0.05**	0.04	0.04	**0.39**	0.27	0.28	**0.91**	0.75	0.80
400	5%	1.4	0.08	0.07	**0.09**	0.46	**0.51**	0.46	**0.95**	0.88	0.86
400	5%	1.6	0.08	**0.13**	0.11	0.45	**0.55**	0.53	**0.96**	0.88	0.94
400	5%	1.8	0.12	**0.16**	0.15	0.70	**0.78**	0.77	**0.99**	**0.99**	0.98
400	5%	2.0	0.16	0.19	**0.20**	0.69	**0.73**	**0.73**	**1.00**	0.99	**1.00**
400	10%	1.2	**0.04**	0.03	**0.04**	0.27	**0.32**	0.23	**0.89**	0.75	0.73
400	10%	1.4	0.03	**0.06**	0.05	0.47	**0.51**	0.48	**0.93**	0.88	0.91
400	10%	1.6	0.12	**0.14**	0.13	0.68	0.73	**0.74**	**0.98**	0.96	0.97
400	10%	1.8	0.14	0.16	**0.17**	0.69	0.76	**0.77**	0.99	0.99	0.99
400	10%	2.0	**0.13**	**0.13**	0.12	0.69	**0.75**	0.74	**1.00**	0.98	0.98

Based on 100 simulated datasets each with 1000 case and 1000 control individuals. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to Blossoc results using $m = 5$ for phased (unphased, respectively) data. GRR, 'Genetic Relative Risk'; WR, 'Wildtype Risk'.

Table 3. Percentage of datasets with the highest scoring marker within distance ϵ (in kb) from the disease-predisposing SNP, which is untyped

ϵ	GRR = 1.4			GRR = 1.6			GRR = 1.8			GRR = 2.0		
	χ^2	P	U	χ^2	P	U	χ^2	P	U	χ^2	P	U
500 case and 500 control individuals												
10	38	23	19	44	41	33	41	47	45	42	51	45
20	50	47	42	64	62	57	61	70	64	62	72	69
30	59	64	54	77	69	68	69	82	79	73	82	84
40	69	71	69	82	79	80	77	86	86	80	88	91
50	78	81	82	89	85	85	81	89	91	85	91	93
60	85	86	86	90	88	89	88	94	93	94	94	97
70	92	92	90	94	92	94	93	95	94	97	96	98
80	98	96	97	98	97	99	98	97	97	99	97	99
2000 case and 2000 control individuals												
10	44	48	45	44	52	52	49	65	68	57	72	74
20	68	63	66	62	72	69	72	78	79	77	88	88
30	76	81	80	74	88	82	79	88	88	89	95	94
40	84	87	90	84	92	87	86	92	89	95	96	97
50	88	93	97	94	96	91	91	94	92	96	98	99
60	95	97	98	97	99	96	95	98	97	97	99	100

Based on 100 simulated datasets for each setting, with 5% WR and $\rho = 40$. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to Blossoc results using $m = 5$ for phased (unphased, respectively) data.

If a unique marker had the highest score, then it was chosen as our estimate of causal SNP position. Otherwise, if $k > 1$ consecutive markers $i_1, \ldots, i_k$ had the highest score, we took the midpoint between markers i_1 and i_k as our estimate. We did not encounter any case in which two markers with the highest score were separated by a marker with a lower score.

Empirical cumulative distributions of localization error ϵ (in kb) are shown in Table 3 for 500 case and 500 control individuals, and 2000 case and 2000 control individuals. Three different methods were used to analyze the data: single-marker χ^2-test, Blossoc on phased data and Blossoc on unphased data; we used $m = 5$ when using Blossoc. The faster the cumulative distribution approaches 1 as ϵ increases, the better the method. Several things are worthwhile noting. First, the localization results of Blossoc on phased data and that of the method developed here for unphased data are comparable. Second, both versions of Blossoc (for phased and unphased data) perform better than χ^2 as GRR increases. Third, the amount of improvement increases as the number of case/control individuals increases.

We also compared Blossoc with Margarita, the method developed by Minichiello and Durbin (2006). Figure 2E of Minichiello and Durbin (2006) shows a plot of empirical cumulative distribution of localization error for 1000 case and 1000 control individuals, under an additive model with GRR = 2.0 and the frequency of the disease-predisposing allele equal to 0.04. In their simulation, a region of size 1 Mb with $\rho = 440$ was used and 300 tagging SNPs were selected. We used Blossoc to analyze the same simulated datasets. Table 4 compares the localization results of the χ^2-test, Blossoc and Margarita. Clearly, both Blossoc and Margarita are more accurate than the χ^2-test in terms

Table 4. Percentage of datasets with the highest scoring marker within distance ϵ (in kb) from the disease-predisposing SNP, for 1000 case and 1000 control individuals, and GRR = 2.0

ϵ	χ^2	BLOSSOC results		MARGARITA results	
		P	U	P	U
50	34	42	46	58	64
100	56	62	70	80	80
150	68	70	78	84	88
200	74	78	84	86	90
250	76	80	86	90	92
300	78	82	88	92	94
350	82	84	90	94	96

Based on simulated datasets from Minichiello and Durbin (2006); each dataset was for a 1 Mb region with $\rho = 440$ and 300 tagging SNPs. Columns denoted χ^2 correspond to single-marker χ^2-test results, while columns denoted 'P' ('U', respectively) correspond to results for phased (unphased, respectively) data. We used $m = 5$ in running BLOSSOC. The results for MARGARITA correspond to Figure 2E of Minichiello and Durbin (2006).

of localization. Further, BLOSSOC is more accurate on unphased data than on phased data; the same behavior was observed in MARGARITA. In general, MARGARITA is more accurate than BLOSSOC, but this increase in accuracy is obtained at the expense of significantly longer running time; for the simulation study shown in Table 4, BLOSSOC took about 3 s per phased dataset and 905 s per unphased dataset, while MARGARITA took 118 620 s per phased dataset and 300 512 s per unphased dataset. That is, MARGARITA is slower than BLOSSOC by a factor of 40 000 for phased data and a factor 300 for unphased data. Such increases in running time can make the difference between a feasible and an infeasible analysis. More results on running time are discussed in Section 3.4.

3.3 Coriell Parkinson's disease genome-wide dataset

Parkinson's disease is a progressive neurodegenerative disorder, affecting more than one per thousand individuals (Kuopio *et al.*, 1999). Fung *et al.* (2006) carried out a genome-wide SNP genotyping assay of publicly available samples from a cohort of 267 Parkinson's disease patients and 270 neurologically normal controls. A total of 408 803 unique SNPs were used from the Illumina Infinium I and HumanHap300 assays. We only considered markers with less than 5% missing data, but applied no other filters. We analyzed the unphased version of this dataset using BLOSSOC with option $m = 5$. The entire analysis took about 4 h on an Intel(R) Pentium(R) 4 CPU 3.00 GHz, 512 Mb RAM. In comparison, it took FASTPHASE 2 days to phase only chromosome 21 of the dataset (containing 6612 SNPs). Beagle (Browning and Browning, 2007), a much faster phase inference tool, is capable of phasing the entire dataset in a reasonable time, but still takes about 8 h, while Blossoc takes 4 h including the association test.

The top-10 highest scoring SNPs found by BLOSSOC are shown in Table 5. Fung *et al.* (2006) performed a single-marker genotypic χ^2 association test and found 26 SNPs with uncorrected p values less than 10^{-4}. However, none of the SNPs showed significant association after Bonferroni correction. One of the 26 SNPs was in the set of top-100 SNPs ranked by BLOSSOC, and other 13 of the 26 SNPs were within 100 kb from at least one SNP in BLOSSOC's top-100.

Table 5. Top-10 highest scoring SNPs in the Parkinson disease dataset (Fung *et al.*, 2006) analyzed using BLOSSOC

Chromosome	dbSNP ID	Location	BLOSSOC score
10p12	rs792456	22214547	29.9567
10p12	rs792455	22233428	29.4246
10p12	rs2666781	22245682	29.4223
10p12	rs2807982	22255866	25.9754
10p12	rs2666750	22259562	25.9754
7p15	rs7793103	21920080	16.1051
7p15	rs7798144	21920802	16.1051
7p15	rs11760455	21921256	16.1051
7p15	rs3829757	21921944	16.1051
8p22	rs7824519	14267167	13.1250

Locations correspond to that of NCBI Build 36.1.

Table 6. Running times of BLOSSOC

Number of Individuals	m	Number of SNPs	Running time (h)
500	5	100,000	0.81
500	5	200,000	1.67
1000	5	100,000	9.17
1000	5	200,000	19.15
2000	5	100,000	50.83
4000	5	50,000	114.17
500	10	100,000	2.93
500	10	200,000	5.58
1000	10	100,000	33.83
1000	10	200,000	64.04
2000	10	100,000	149.17

See the main text for the computer spec.

We also compared the SNPs in BLOSSOC's top-100 with the set of top-100 loci ranked by single-marker χ^2-test. One third of BLOSSOC's top-100 were within 20 kb of at least one top-100 χ^2-SNP while 58 where farther than 1 Mb away from any χ^2 top 100. Such an overlap seems reasonable between multi-marker and single-marker association methods.

3.4 Running time

We examined the running time of BLOSSOC on datasets simulated using FREGENE (Hoggart *et al.*, 2007), software capable of simulating large-scale sequence data. Table 6 shows a summary of running time results. In light of the fact that FASTPHASE takes 3 days to phase a dataset with 500 individuals and 10 000 SNPs, speed clearly is a notable advantageous feature of BLOSSOC. Note that the running time of BLOSSOC is roughly linear in the number of SNPs, and it depends more on the number of individuals than on the number of SNPs.

4 DISCUSSION

We have presented a multi-locus association mapping method that builds local phylogenies directly from unphased genotype data, thus avoiding the time-consuming step of phasing the genotype data before analysis. We have shown that, for both ranking quality and localization accuracy, the performance of our method on unphased data is comparable to that for the case in which the true phase of the data is known.

The current algorithm deals with incompatible regions by using an entropy minimization algorithm to infer the phase of input genotypes and uses the tree building algorithm described in Mailund *et al.* (2006) to build local phylogenies from the inferred haplotypes. In our algorithm, using larger values of the parameter m results in more incompatible regions, and it is not yet clear how much of the false-positive associations and localization errors are caused by the entropy minimization algorithm. In our experiments, we found $m=5$ to be slightly superior to $m=3$ and $m=10$, for both $\rho=100$ and $\rho=400$ (results not shown). Future work will include ways of finding the optimal setting for m for different datasets. One possible way is through analyzing the differences in LD patterns within and between datasets.

We found in our experiments that most of the running time of our method was spent on phasing incompatible regions. As a result, one observes a significant increase in the running time when one increases the value of m or the number of individuals. Hence, developing more efficient algorithms to deal with incompatible regions will be worthwhile.

Our approach builds local phylogenies to test for a significant clustering of affected individuals in local trees. As a consequence, population structure within a sample can result in false-positive associations. We would advise caution when applying our algorithm to datasets from heavily isolated populations.

A potential problem with the BLOSSOC approach is selecting candidate markers for replication. For single marker analysis, highly ranked SNPs are candidates to be tested in a replication cohort to distinguish true signals from spurious ones. In BLOSSOC, scores are based on local phylogenies inferred from contiguous sets of markers, and retyping large regions for potential replication can be costly. We are currently considering ways of finding minimal haplotypes around top loci ranked by BLOSSOC that can tag a disease-predisposing SNP better than can any single marker.

ACKNOWLEDGEMENTS

T.M. is grateful to Mikkel Schierup and Søren Besenbacher for fruitful discussions of the scoring approach and the software design of BLOSSOC. We thank Mark Minichiello for providing us with the simulation data used in Minichiello and Durbin (2006).

Funding: National Science Foundation (NSF) (IIS-0513910 to Z.D.); Danish Research Agency – Forskningsrådet for Natur og Univers (FNU) (272-07-0380 to T.M.) National Institute of Health (1K99-GM080099 and 4R00-GM080099 to Y.S.S.).

Conflict of Interest: none declared.

REFERENCES

Amundadottir,L.F. *et al.* (2006) A common variant associated with prostate cancer in European and African populations. *Nat. Genet.*, **38**, 652–658.

Arking,D.E. *et al.* (2006) A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat. Genet.*, **38**, 644–651.

Barrett,J.C. and Cardon,L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Gen.* **81**, 1084–1097.

Clark,T.G. *et al*. (2007) Bayesian logistic regression using a perfect phylogeny. *Biostatistics*, **8**, 32–52.

Ding,Z. *et al*. (2005) A linear-time algorithm for the perfect phylogeny haplotyping. In *Proceedinds of the 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, Vol. 3500 of *Lecture Notes in Bioinformatics*. Springer-Verlag, Berlin, Germany, pp. 585–600.

Ding,Z. *et al*. (2006) A linear-time algorithm for the perfect phylogeny haplotyping (PPH) problem. *J. Comput. Biol.*, **13**, 522–553.

Fung,H.C. *et al.* (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.

Gusev,A. *et al*. (2007) Highly scalable genotype phasing by entropy minimization. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **5**, 252–261.

Halperin,E. and Eskin,E. (2004) Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, **20**, 1842–1849.

Halperin,E. and Karp,R. (2005) The minimum-entropy set cover problem. *Theor. Comput. Sci.*, **348**, 240–250.

Hoggart,C.J. *et al*. (2007) Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725–1731.

Kuopio,A.-M. *et al*. (1999) Changing epidemiology of Parkinson's disease in southwestern Finland. *Neurology*, **52**, 302–308.

Mailund,T. *et al*. (2005) CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, **6**, e6.

Mailund,T. *et al*. (2006) Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, **7**, 454.

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

Pe'er,I. *et al*. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663–667.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Sevon,P. *et al*. (2006) TreeDT: tree pattern mining for gene mapping. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 174–185.

Smyth,D. *et al*. (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet.*, **38**, 617–619.

Tachmazidou,I. *et al.* (2007) Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genet.*, **3**, e15.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Vijayasatya,R. and Mukherjee,A. (2005) An efficient algorithm for perfect phylogeny haplotyping. In *Proceedings of the IEEE Computer System Bioinformatics Conference*. IEEE Computer Society, Los Alamitos, USA, pp. 103–110.

Waldron,E.R.B. *et al*. (2006) Fine mapping of disease genes via haplotype clustering. *Genet. Epidemiol.*, **30**, 170–179.

Wu,Y. (2007) Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms. In *Proceedings of the 11th Annual Internationnal Conference on Research in Computational Molecular Biology (RECOMB)*, Springer-Verlag, Berlin, Germany, pp. 488–502.

Zöllner,S. and Pritchard,J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, **169**, 1071–1092.