DOI: 10.1534/genetics.108.092643

Local Phylogeny Mapping of Quantitative Traits: Higher Accuracy and Better Ranking Than Single-Marker Association in Genomewide Scans

Søren Besenbacher,[1] Thomas Mailund and Mikkel H. Schierup

Bioinformatics Research Center, University of Aarhus, DK-8000 Århus C, Denmark

Manuscript received June 13, 2008
Accepted for publication December 7, 2008

ABSTRACT

We present a new method, termed QBlossoc, for linkage disequilibrium (LD) mapping of genetic variants underlying a quantitative trait. The method uses principles similar to a previously published method, Blossoc, for LD mapping of case/control studies. The method builds local genealogies along the genome and looks for a significant clustering of quantitative trait values in these trees. We analyze its efficiency in terms of localization and ranking of true positives among a large number of negatives and compare the results with single-marker approaches. Simulation results of markers at densities comparable to contemporary genotype chips show that QBlossoc is more accurate in localization of true positives as expected since it uses the additional information of LD between markers simultaneously. More importantly, however, for genomewide surveys, QBlossoc places regions with true positives higher on a ranked list than single-marker approaches, again suggesting that a true signal displays itself more strongly in a set of adjacent markers than a spurious (false) signal. The method is both memory and central processing unit (CPU) efficient. It has been tested on a real data set of height data for 5000 individuals measured at ∼317,000 markers and completed analysis within 5 CPU days.

GENOMEWIDE SNP chips typing ∼500,000 markers in several thousand individuals have successfully identified many markers associated with disease (Amundadottir *et al.* 2006; Arking *et al.* 2006; Smyth *et al.* 2006; Easton *et al.* 2007; Gudmundsson *et al.* 2007a,b; Stacey *et al.* 2007; Wellcome Trust Case Control Consortium 2007). The majority of these findings are based on single-marker association tests of case/control data sets, with a few studies also including simple tests based on using pairs of markers to tag untyped markers (Gudmundsson *et al.* 2007a) or using haplotype blocks defined as the markers spanned by recombination hotspots (Gudmundsson *et al.* 2007a), or more sophisticated haplotype methods (Marchini *et al.* 2007; Wellcome Trust Case Control Consortium 2007; Browning and Browning 2008).

Whereas most recent studies have dealt with qualitative traits (*e.g.*, presence/absence of a disease), quantitative traits such as height (Gudbjartsson *et al.* 2008; Lettre *et al.* 2008; Weedon *et al.* 2008) or quantitative trait (QT) interval in the heart (Arking *et al.* 2006) will be the next traits to approach.

SNP markers on typical genotyping chips are sufficiently close to display a large level of linkage disequilibrium. Therefore, testing SNP by SNP will not necessarily be the most powerful use of data and haplotype-based methods might provide additional findings from already analyzed chip data sets (Pe'er *et al.* 2006). The fact that quantitative traits are likely to be controlled by a lot of genes each having only a small effect might make these even harder to detect than qualitative disease traits. Methodologically, a lot of effort has been devoted to devising methods that can utilize the linkage disequilibrium (LD) patterns among markers. A trade-off, however, must be made between the sophistication of the method and the computational demands of the method. We have recently developed a new multi-SNP method for case/control data (Mailund *et al.* 2006) that, although of similar accuracy, is orders of magnitude faster than the methods we have compared it with and capable of analyzing whole-genome data sets with thousands of individuals and hundreds of thousands of markers in a few central processing unit (CPU) days. The method resembles other recent methods, *e.g.*, Minichiello and Durbin (2006) and Zöllner and Pritchard (2005), in that it attempts to infer local genealogies along the genome and then test for nonrandom distribution of cases and controls on these. Our method, however, achieves significant speedups by taking a simpler approach to how local genealogies are constructed. Rather than sampling trees from the coalescent process (with or without recombination) and then averaging scores over sampled trees, we simply build a single (near) perfect phylogeny for each locus, when assuming the infinite-sites model for the underlying genetic model; see Mailund *et al.* (2006) for details.

[1] *Corresponding author:* Bioinformatics Research Center, University of Aarhus, C.F. Møllers Alle, Bldg. 1110, DK-8000 Århus C, Denmark. E-mail: besen@daimi.au.dk

Here we extend the framework to search for variants affecting quantitative traits. As for the case/control method, we construct local genealogies along the genome, but the test for association is now based on uneven distribution of quantitative trait values in this tree. We show by simulation that the approach is more accurate for localization and better at ranking true positives among many negatives in large-scale studies when compared with considering markers independently.

Other haplotype-based methods for association mapping in case/control studies have previously been adapted to mapping of quantitative traits, including the HapMiner method (Li *et al.* 2006) and the QHPM method (Onkamo *et al.* 2002). The latter is, however, substantially slower than QBlossoc and is thus not relevant for whole-genome analysis. The HapMiner method is faster than QHPM but still slower than QBlossoc. Our experiments show that analyzing a data set of 100 markers and 2000 individuals takes 80 sec for QBlossoc (the same parameters as used in the experiments) and 80 min for HapMiner with default parameters (without permutation tests).

The QBlossoc software is available for free at www.daimi.au.dk/~mailund/Blossoc/ and was recently used by M. C. Ledur, N. Navarro and M. Pérez-Enciso (unpublished results) for analyzing the QTL-MAS XII data sets, where one of the true causative loci was recovered only by QBlossoc.

METHODS

Building local genealogies: At any given locus in the genome, the chromosomes of our sampled individuals are related by a tree genealogy. This genealogy, however, is unknown and must be inferred from a set of markers around the locus. The inference of local genealogies we use is described in Mailund *et al.* (2006) and we give only a brief overview here and refer to Mailund *et al.* (2006) for further details.

The general approach we take is to construct a *perfect phylogeny* for a maximal region of markers for which such a tree exists (Figure 1). The perfect phylogeny can be constructed efficiently from phased haplotype data using a linear-time algorithm (Gusfield 1991). In large data sets with many individuals it is, however, very unlikely that it will be possible to build a perfect phylogeny from more than one or two markers because recombination among some individuals has created incompatibilities. In such cases we essentially ignore these incompatibilities when we build the trees by adding more than one mutation to the tree for each marker. However, we make sure to add markers to the tree on the basis of their distance from the marker we are looking at, so that we give higher weight to markers close by and many of the additional mutations are later pruned by the model selection criterion.

With a small number of markers compared to the number of chromosomes, the tree will not be fully resolved and it might not reflect the true local genealogy perfectly. This is not only because we often add incompatible markers as explained above but also because absence of incompatibility between markers does not imply absence of recombination. Many recombination events do not create incompatibilities. However, our final goal is not to infer the true local genealogy perfectly, but only to infer a topology that captures enough of the true genealogy to retain signals of association, if present in the true local genealogy.

Since this tree-building algorithm works on phased data, it is necessary to apply a phasing method to phase genotype data and infer missing data. For this purpose we recommend the Beagle program of Browning and Browning (2007), which, like our method, is fast enough to be applied to whole-genome data sets.

Scoring trees: Once a local phylogeny is inferred, we score it on the basis of how traits are distributed on the tree. Below we describe two approaches to scoring. The first approach considers the trait of each chromosome independent of other chromosomes—ignoring that chromosomes pairwise must share the phenotype of an individual—and as such takes an "allelic view" of the problem. The second approach considers individuals' genotypes explicitly as pairs of chromosomes in the tree. In the *Simulations* section, we evaluate the scores and consider different genetic models to evaluate how the more complex genotype model performs, compared to the simpler model.

Independent scoring of chromosomes: In the simpler model, each tree, T, defines a set of models for the data in the following way: Any edge, e in the tree splits the nodes of the tree into two disjunct sets S_i and S_j, where the phenotypes of leaves in S_i have a different distribution than the phenotypes of the leaves in S_j. The biological interpretation is that a mutation that affects the phenotypes of the descendants has occurred on the edge e. Since we do not expect to infer local phylogenies perfectly, this edge is not actually likely to exist in the *true* underlying genealogy, but as long as the edges in T group the majority of leaves correctly, signals from the true phylogeny should also be reflected in the inferred phylogeny.

If we look at another edge, e', then it will split either S_i or S_j into two new disjunct sets. More generally a set of k edges will split the leaves of the tree into $k + 1$ disjunct subsets. Such a model with more subsets can be better if several mutations have occurred in a gene that affects the phenotype. For each subset of nodes in the tree, $\{S_i\}$, we can assign a distribution mean, μ_i, such that the phenotype of a leaf in this subtree is normally distributed with mean μ_i and variance σ^2. To limit the number of parameters in the models we use the same variance for all subtrees.

Given a tree with n leaves numbered $1, \ldots, n$, with corresponding phenotypes $Y = y_1, \ldots, y_n$ and a model, Θ, dividing the leaves into k subsets $S_1, \ldots, S_k$, with means $\mu_1, \ldots, \mu_k$, the conditional likelihood of the data becomes

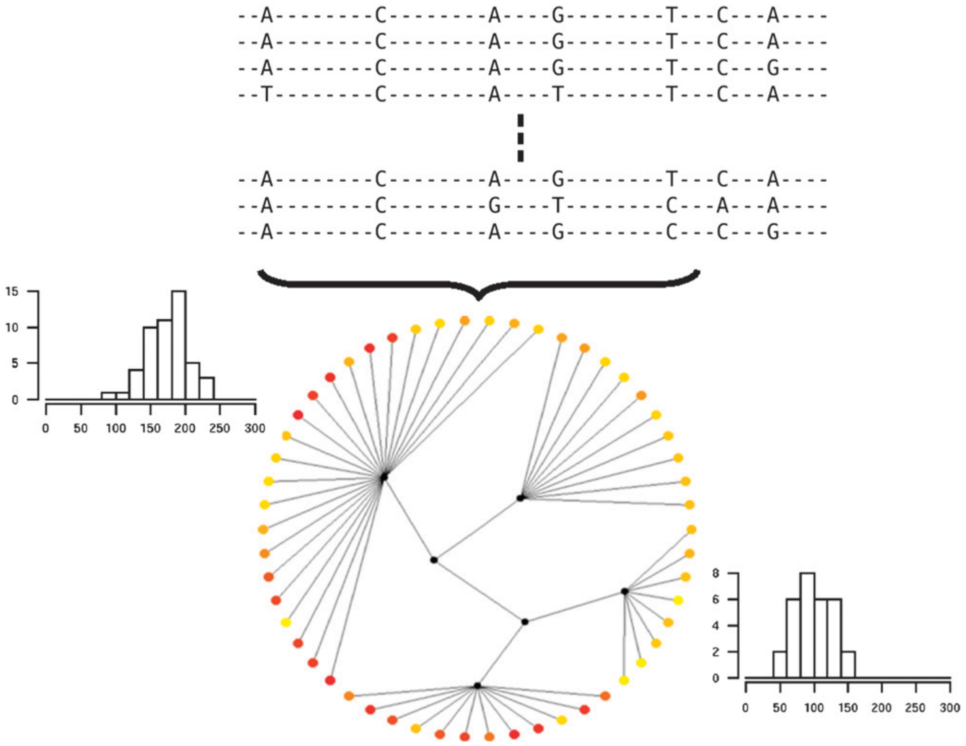
--A--------C--------A---G--------T---C---A----
--A--------C--------A---G--------T---C---A----
--A--------C--------A---G--------T---C---G----
--T--------C--------A---T--------T---C---A----
--A--------C--------A---G--------T---C---A----
--A--------C--------G---T--------C---A---A----
--A--------C--------A---G--------C---C---G----
15
10
5
0
0 50 100 150 200 250 300
8
6
4
2
0
0 50 100 150 200 250 300

FIGURE 1.—A tree built from five SNP markers. The nodes in the tree are colored after their phenotype values; high values are red and low values are yellow. The two histograms show the distribution of phenotypes in two of the subtrees.

$$\Pr(Y \mid \Theta, \vec{\mu}, \sigma) = \prod_{i=1}^{k} \prod_{x \in S_i} \mathcal{N}(y_x \mid \mu_i, \sigma^2), \qquad (1)$$

assuming the phenotypes of the leaves are independent. To score the tree, we use the maximum likelihood for the set of models compatible with this tree, but use the Bayesian information criterion (BIC) to avoid overfitting when selecting the set of clusters, Θ:

$$\text{score}(T) = \max_{\Theta,\vec{\mu},\sigma}\{-2 \ln \Pr(\mathrm{Y} \mid \Theta, \vec{\mu}, \sigma) + k \cdot \ln(n)\}. \qquad (2)$$

When applying the scoring, we start out by testing all the models with one mutation; then we test all combinations of two mutations and then on up to some maximum number of mutations $k_{\max}$.

Scoring chromosomes in pairs: This approach for scoring also uses the Bayesian information criterion to score the trees. The difference is that now every individual has two leaves, which means that in a model where the leaves are split into k different subsets there are $k(k + 1)/2$ different possible combinations that the two leaves from an individual can have. As before, we model the trait for individuals as normally distributed with independent means for each class, μ_{ij}, but with shared variance, σ: Given a data set with n individuals we denote by $h_{11}, h_{12}, \ldots, h_{n1}, h_{n2}$ the haplotypes/leaves, with h_{i1} and h_{i2} coming from the same individual, and by $Y = y_1, \ldots, y_n$ the phenotypes. The model, Θ, divides the leaves into k subsets $S_1, \ldots, S_k$ and the mean value of individuals with one leaf in S_i and one in S_j is μ_{ij}. The likelihood of the data becomes

$$\Pr(Y \mid \Theta, \vec{\mu}, \sigma) = \prod_{i=1}^{k} \prod_{j=1}^{i} \prod_{\substack{(h_{x1}\in S_i \wedge h_{x2}\in S_j)\vee \\ (h_{x2}\in S_i \wedge h_{x1}\in S_j)}} \mathcal{N}(y_x \mid \mu_{ij}, \sigma^2), \qquad (3)$$

where the innermost product is over all individuals with one chromosome in S_i and one in S_j. The score for the tree becomes

$$\text{score}(T) = \max_{\Theta,\vec{\mu},\sigma}\{-2 \ln \Pr(\mathrm{Y} \mid \Theta, \vec{\mu}, \sigma) + K \cdot \ln(n)\}, \qquad (4)$$

where K is the number of nonempty pairs of classes, a number between 1 and $k(k + 1)/2$.

Simulations: We tested accuracy, power, and ranking on simulated data sets where the causative variants were known. We used the CoaSim (MAILUND *et al.* 2005) simulation software to simulate a population of chromosomes containing SNP markers (with a minor allele frequency >0.05) and a quantitative trait nucleotide that affects the phenotype with a frequency within a specified range. The chromosomes were paired to produce diploid individuals, and we then assigned phenotype values to the individuals on the basis of their genotype at the quantitative trait nucleotide. As in LONG and LANGLEY (1999), we use a phenotype distribution where a fixed amount of the variation is attributable to the quantitative trait nucleotide. The quantitative trait nucleotide was then removed from the simulated data set such that localization is only through linkage disequilibrium with other markers. We have simulated data under both an additive model and a recessive model. For the additive model we used the formula

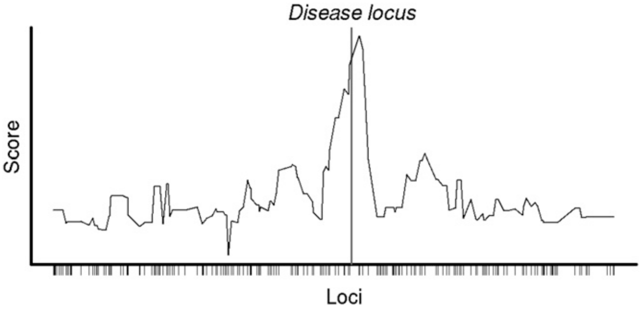

Figure 2.—The QBlossoc score along a simulated data set with 200 markers (marked on the *x*-axis). The position of the quantitative trait mutation is shown as a vertical line.

$$y_i = \sqrt{1-\pi}N(0,\,1) + Q_i\sqrt{\frac{\pi}{2p(1-p)}}, \qquad (5)$$

where π is the percentage of the variation attributable to the quantitative trait nucleotide, p is the allele frequency, $N(1,\,0)$ is standard normally distributed, and Q_1 takes the value of -1, 0, or 1 depending on whether the ith individual is homozygous wild type, heterozygous, or homozygous mutant at the trait locus. For the recessive model, where only individuals that were homozygous mutants at the trait locus had a different distribution of the trait in question, the formula is

$$y_i = \sqrt{1-\pi}N(0,\,1) + Q_i'\sqrt{\frac{\pi}{p'(1-p')}}, \qquad (6)$$

where p' is the fraction of individuals that are homozygous mutants and Q_i' is 1 if the ith individual is homozygous mutant and 0 otherwise.

For each of the localization experiments in the RESULTS section, we simulated 500 data sets with 2000 diploid individuals where the trait locus had a minor allele frequency between 10 and 20%. Data sets contain 100 SNP markers simulated with a scaled recombination rate $\rho = 4N_e$ of 200, which in a human setting means that we are looking at intervals of ~0.5 cM, or 500,000 bp, *i.e.*, a density of 1 marker per 5 kb that is similar to the most commonly used SNP typing chips. For the power experiments the data sets consisted of only 10 markers but the densities of the markers were the same.

RESULTS

Localization: A typical run of QBlossoc on a simulated data set is shown in Figure 2. When comparing QBlossoc with single-marker methods for ability to localize the quantitative trait nucleotide, we used the highest-scoring marker as our estimate of the trait locus and compared this estimate with the true locus.

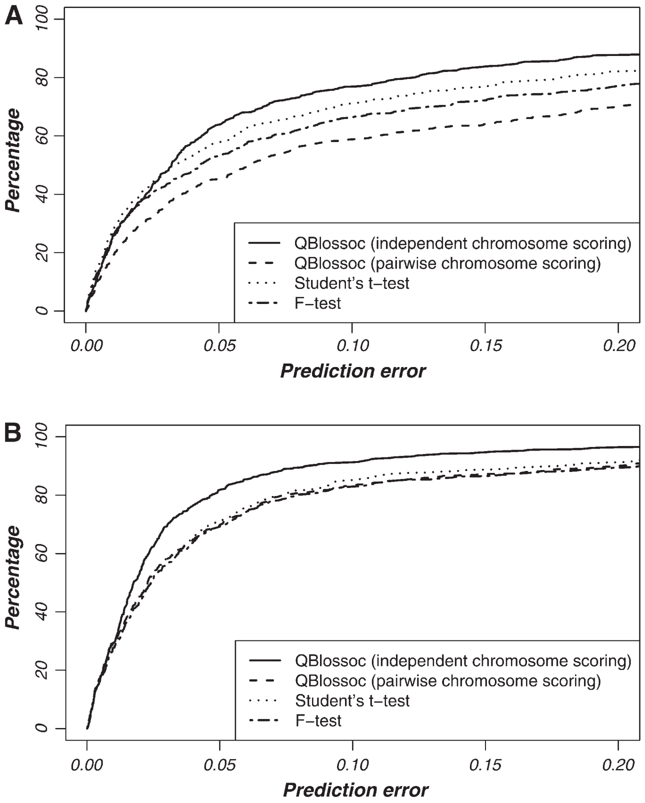

Figure 3.—Accuracy of localization, additive effect. Each plot shows summary results for 500 data sets. In the left plot the trait-affecting nucleotide explains 1% of the variation and in the right it explains 2%. The *x*-axis shows the relative distance between the highest-scoring marker and the position affecting the quantitative trait, and for a given *x* the *y*-axis shows which percentage of the data sets had a relative distance less than *x*.

Figures 3 and 4 compare the accuracy of localization, measured as the distance between the highest-scoring marker for the different methods and the known position of the true locus, for the additive and recessive models, respectively. We compare our methods with two standard statistical tests on single markers: a Student's *t*-test that tests if the trait distributions of the two different alleles are the same and an *F*-test that tests if the trait distributions of the three different genotypes have the same mean. Figures 3 and 4 show that QBlossoc generally leads to more accurate localization than the single-marker test despite the fact that the quantitative trait nucleotides have a frequency similar to that of the markers. If a quantitative trait locus is expected to be of a recessive nature, the scoring of chromosomes in pairs should be used, but without such expectation about the nature of the quantitative trait locus the independent scoring of chromosomes is probably preferable.

Power: Power of detection for the different methods was assessed by comparing the highest score in 500 replicated data sets for a given effect of the causative locus to the null distribution of scores. The null distribution was empirically determined by simulating 5000 data sets with no effect of the quantitative trait

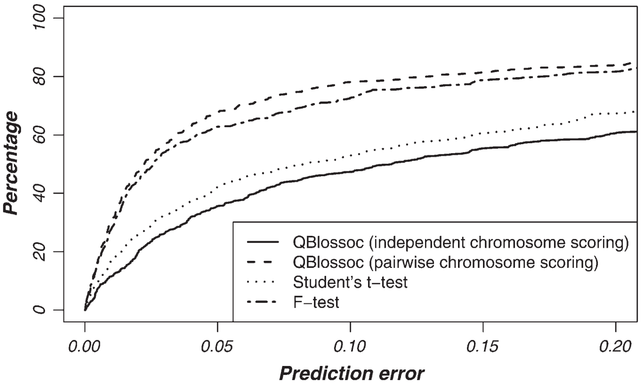

FIGURE 4.—Accuracy of localization, recessive effect: summary results for 500 data sets. The *x*-axis shows the relative distance between the highest-scoring marker and the position affecting the quantitative trait, and for a given *x* the *y*-axis shows which percentage of the data sets had a relative distance less than *x*. The amount of variation attributable to the quantitative trait nucleotide (π) was 0.02.

nucleotide. Figure 5 shows the percentage of data sets containing a maximum score that was higher than the 99% quantile of the null distribution, as a function of the effect π and the frequency of the quantitative trait nucleotide. The trait variant frequency is simulated in ranges varying from 2–4% (top left corner) to 8–10%, while the effect varies in the range 0.005–0.02. For all frequency ranges, we see that the power of QBlossoc grows faster than the power of the *t*-test, as a function of π, but that the power of the two converges as π decreases. The difference in power between QBlossoc and the *t*-test increases as the variant frequency decreases.

Ranking: In genomewide association studies it is common practice to select the highest-scoring markers and seek replication of an effect in a different cohort. Thus, the ranking of true positive findings among the many negative ones is at least as important as initial genomewide significance, because it is important to ensure that as many of the true findings as possible are included among the markers that are tested in the replication cohort. We tested the ability of QBlossoc to rank the true positive markers against single-marker methods in the following way. We define the rank of a marker as its position in a sorted list of all the marker values. It is possible for two markers to get exactly the same score and in that case we randomly decide which of them gets the lowest rank. The histograms in Figure 6 show the distribution of the highest-ranking marker that is not >20 kb away from the phenotype-affecting variant for the 500 simulated data sets from Figure 3. The results in Figure 6 show that we are more likely to have a high scoring marker closer to the causative locus using our method than with a single-marker Student's *t*-test.

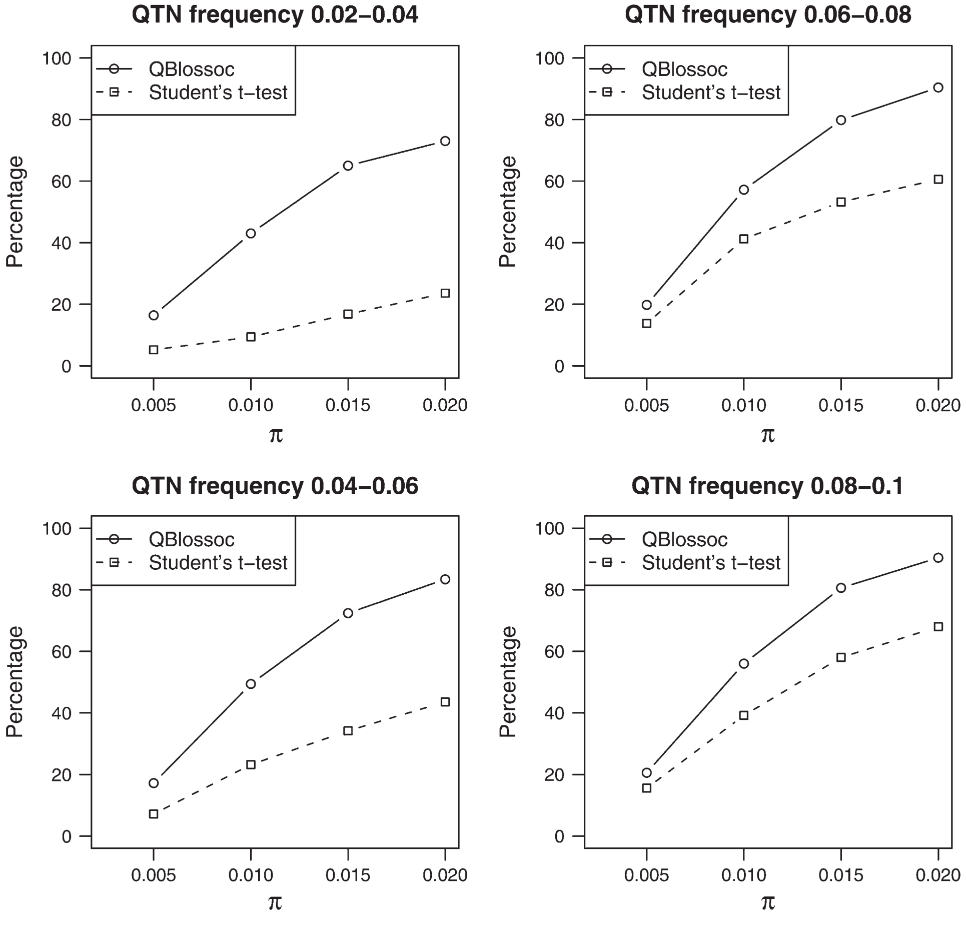

FIGURE 5.—The power to detect a phenotype-affecting variant for different levels of heritability (π) and causal variant (quantitative trait nucleotide, QTN) frequency, using QBlossoc or a Student's *t*-test.

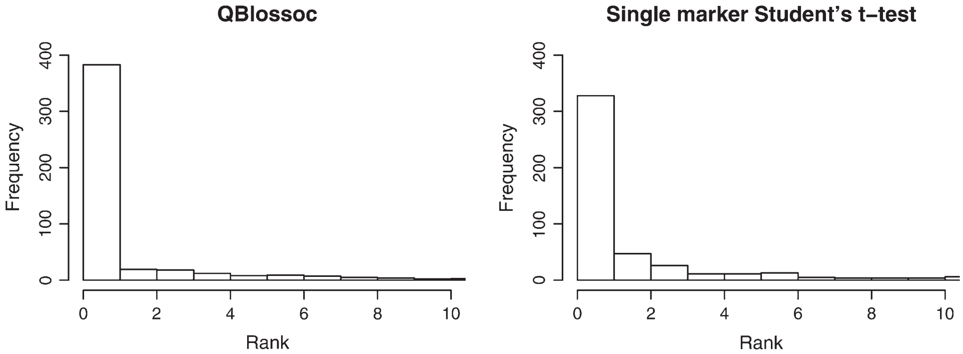

Figure 6.—Distribution of the highest-ranking marker within a 20-kb radius of a quantitative trait nucleotide in 500 data sets with an additive disease model. The value of π was 0.02 and the quantitative trait nucleotide had a frequency between 0.1 and 0.2.

Proof of principle on a genomewide data set: We tested the performance of QBlossoc on a proprietary data set from deCode Genetics on height in 5000 individuals typed for ~317,000 markers using the Illumina human Hap chip. The genotype data were first phased using Beagle. A QBlossoc analysis was performed on the residuals from a linear model of height with the effect of age cohort and sex removed. The analysis was completed in 55 hr of CPU time with $m = 10$ and $k_{\max} = 1$ (or 175 hr with $m = 5$ and $k_{\max} = 2$) and provided a ranking of markers that differs from the ranking from a Student's *t*-test single-marker analysis (SMA). We determined the 20 regions with the highest SMA and QBlossoc scores and compared these with the published findings, which include 64 markers in 51 regions (Weedon *et al.* 2007, 2008; Gudbjartsson *et al.* 2008; Lettre *et al.* 2008; Sanna *et al.* 2008). Three of the published findings are in the top 20 SMA regions ($D' > 0.5$) and a different finding is in the top 20 QBlossoc regions in the present study. Thus, in this test, SMA analysis detects more of the known height loci but Blossoc finds a height locus that SMA does not detect. We note that the published height loci were indeed initially found by SMA analysis in larger cohorts than the one used here (>10,000 individuals). It is promising that QBlossoc can find additional susceptibility loci missed by single-marker analysis. Thus, it will be of interest to determine if other top 20 QBlossoc findings can be replicated. This is presently being done by adding more individuals to the Icelandic cohort and replicating the best findings from this analysis.

DISCUSSION

We have presented QBlossoc as a new fast, accurate, and powerful approach to haplotype-based analysis of genomewide association mapping studies where the trait measured is quantitative.

The method constructs a rough estimate of the local phylogenies along the genome and scores these with respect to the clustering of phenotypes on the tree. When scoring trees, we model phenotypes as normally distributed with a mean that depends on the local tree. This implicitly restricts the method to data where individuals have not been sampled conditional on their phenotype. For studies where, *e.g.*, extreme trait values have been selected, it would be more appropriate to translate the continuous trait into a binary trait and use our previous method.

We have shown that QBlossoc is sufficiently fast to analyze the results of state of the art experiments in a reasonable amount of time and that it possesses advantages over simpler single-marker approaches. Simulations show that the approach under reasonable parameter choices is better at ranking true positives and better at localizing them than single-marker approaches. More importantly, when the quantitative trait nucleotide frequency is low, *i.e.*, 0.01–0.05, QBlossoc performs even better, relative to a single-marker approach, since haplotype approaches can combine sets of common markers to define a rare haplotype in strong LD with a rare causative variant. QBlossoc should thus be increasingly useful for the next generation of association mapping studies ahead that will employ >20,000 individuals and search for trait-affecting variants of lower frequency.

LITERATURE CITED

Amundadottir, L. T., P. Sulem, J. Gudmundsson, A. Helgason, A. Baker *et al.*, 2006 A common variant associated with prostate cancer in European and African populations. Nat. Genet. **38:** 652–658.

Arking, D., A. Pfeufer, W. Post, W. Kao, C. Newton-Cheh *et al.*, 2006 A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat. Genet. **38:** 644–651.

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

Browning, B. L., and S. R. Browning, 2008 Haplotypic analysis of Wellcome Trust Case Control Consortium data. Hum. Genet. **123:** 273–280.

Easton, D. F., K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson *et al.*, 2007 Genome-wide association study identifies novel breast cancer susceptibility loci. Nature **447:** 1087–1093.

Gudbjartsson, D. F., G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson *et al.*, 2008 Many sequence variants affecting diversity of adult human height. Nat. Genet. **40:** 609–615.

Gudmundsson, J., P. Sulem, A. Manolescu, L. T. Amundadottir, D. Gudbjartsson *et al.*, 2007a Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat. Genet. **39:** 631–637.

Gudmundsson, J., P. Sulem, V. Steinthorsdottir, J. T. Bergthorsson, G. Thorleifsson *et al.*, 2007b Two variants on chromosome 17 confer prostate cancer risk, and the one in tcf2 protects against type 2 diabetes. Nat. Genet. **39:** 977–983.

Gusfield, D., 1991 Efficient algorithms for inferring evolutionary trees. Networks **21:** 19–28.

Lettre, G., A. U. Jackson, C. Gieger, F. R. Schumacher, S. I. Berndt *et al.*, 2008 Identification of ten loci associated with height highlights new biological pathways in human growth. Nat. Genet. **40:** 584–591.

Li, J., Y. Zhou and R. Elston, 2006 Haplotype-based quantitative trait mapping using a clustering algorithm. BMC Bioinformatics **7:** 258.

Long, A. D., and C. H. Langley, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. **9:** 720–731.

Mailund, T., M. Schierup, C. Pedersen, P. Mechlenborg, J. Madsen *et al.*, 2005 CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics **6:** 252.

Mailund, T., S. Besenbacher and M. Schierup, 2006 Whole genome association mapping by incompatibilities and local perfect phylogenies. BMC Bioinformatics **7:** 454.

Marchini, J., B. Howie, S. Myers, G. McVean and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. **39:** 906–913.

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

Onkamo, P., V. Ollikainen, P. Sevon, H. T. T. Toivonen, H. Mannila *et al.*, 2002 Association analysis for quantitative traits by data mining: Qhpm. Ann. Hum. Genet. **66:** 419–429.

Pe'er, I., P. de Bakker, J. Maller, R. Yelensky, D. Altshuler *et al.*, 2006 Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat. Genet. **38:** 663–667.

Sanna, S., A. U. Jackson, R. Nagaraja, C. J. Willer, W.-M. Chen *et al.*, 2008 Common variants in the gdf5-uqcc region are associated with variation in human height. Nat. Genet. **40:** 198–203.

Smyth, D., J. Cooper, R. Bailey, S. Field, O. Burren *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.

Stacey, S. N., A. Manolescu, P. Sulem, T. Rafnar, J. Gudmundsson *et al.*, 2007 Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat. Genet. **39:** 865–869.

Weedon, M. N., H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans *et al.*, 2008 Genome-wide association analysis identifies 20 loci that influence adult height. Nat. Genet. **40:** 575–583.

Weedon, M. N., G. Lettre, R. M. Freathy, C. M. Lindgren, B. F. Voight *et al.*, 2007 A common variant of hmga2 is associated with adult and childhood height in the general population. Nat. Genet. **39:** 1245–1250.

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.

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

Communicating editor: G. Gibson