“I See Dead People:” Gene Mapping Via Ancestral Inference

Paul Marjoram,[1] Lada Markovtsova[2] and Simon Tavaré[1,2,3]

[1]Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, CHP 220, Los Angeles, CA 90033.
[2]Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113.
[3]Program in Molecular Biology, SHS 172, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1340.

ABSTRACT

We introduce a Markov chain Monte Carlo [MCMC] approach to fine-scale mapping using population data. The method uses the coalescent to model the genealogy underlying a random sample of individuals, and the consequent correlation induced between SNP markers. Using this methodology we can undertake a Bayesian or full maximum likelihood analysis using multiple markers. We apply our approach to an analysis of quantitative trait Q1 and major gene MG1 in replicate 1 of the GAW12 simulated data set. Using random samples of 10 individuals we are often able to locate the region containing the functional mutation influencing Q1.

Key words: coalescent, Markov chain Monte Carlo.

INTRODUCTION

It is now routine to use pedigree-based studies for mapping of genes via an analysis of the association between the trait of interest and a set of markers. Linkage disequilibrium mapping using population-based samples provides an alternative approach for fine-scale mapping; cf. Hästbacka et al. (1992), Devlin and Risch (1995), Clayton (2000). Many of these analyses either proceed in a marginal sense, by calculating statistics for markers one at a time, for example, or make simplifying assumptions about the dependence of multiple markers. Here we present an approach that exploits the coalescent as a model for the underlying genealogy of the sample, and uses it to provide a Bayesian (or full likelihood-based) analysis of gene location based on dependent markers. In pedigree-based approaches, it is common to ignore the correlation between the founder individuals of the pedigree, or between founder individuals of different pedigrees if the analysis uses multiple pedigrees. This is not because such correlations do not impact the data, but rather because their ancestral relationships are unknown. It is precisely this correlation that can be modeled by the coalescent process. We are currently adapting the approach presented here to model this unknown ancestral information in the pedigree analysis setting, and to assess the influence this correlation may have on the results of such analyses. In this exploratory paper, we restrict ourselves to random samples of individuals drawn from replicate 1 of the simulated data. In doing so we hope to assess the potential of such methods to work on random samples, in which one can expect to see more recombinational events and thus have a greater chance of localizing the gene of interest. Furthermore, our approach allows for simultaneous estimation of other parameters of interest, such as rates of mutation or recombination.

Several recent approaches attempt to use the information present at multiple markers simultaneously, for example McPeek and Strahs (1999) and Morris et al. (2000). Their model for the correlation between loci is broadly based on population genetics theory, without making any explicit assumptions about population history or the ancestry of the sample. In this paper we describe this correlation via a generalization of the *coalescent*. The coalescent was introduced in Kingman (1982) as a mathematical description of the genealogy that underlies the evolution of a population of genes, and it has become a standard tool for the analysis of molecular population data. Several existing methods have demonstrated the efficacy of evolutionary analyses based on the coalescent

(e.g. Griffiths and Tavaré (1994), Kuhner et al. (1995), Donnelly and Stephens (2000)). The coalescent with recombination was originally proposed by Hudson (1983, 1991). It was further developed by Griffiths and Marjoram (1996), and has become known as the Ancestral Recombination Graph (ARG). It has been used as a basis for analyses of data subject to recombination (e.g. Griffiths and Marjoram (1996), Nielsen (2000), Kuhner et al. (2000)). A brief description of the ARG follows.

METHODS

An MCMC approach using Ancestral Recombination Graphs

We suppose we have randomly sampled individuals from the present day population. For simplicity we assume the population to be panmictic and of constant size, although it is relatively straightforward to adapt our approach to include population growth. We also assume that all mutations are neutral. We then look back in time at the stochastic process that describes the ancestry, or genealogy, of the sample. It is this graph that is known as the ARG and branches of this graph will be referred to here as *lines of ancestry.* An example is shown in Figure 1. As we look back in time three events may occur:

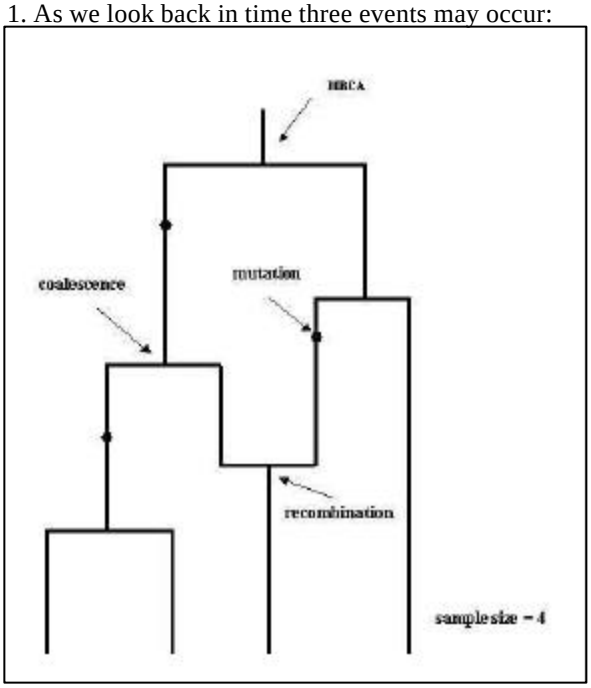

Figure 1: An example Ancestral Recombination Graph.

i. Two lines of ancestry will *coalesce* to form a single line of ancestry. This corresponds to two ancestral alleles being inherited, *in their entirety*, from a single common gene in the previous generation.

ii. A *mutation* will occur to a line of ancestry, changing the type of a gene.

iii. A single line of ancestry will split into two lines, A and B say, such that some of the markers are inherited from A, while the remaining markers are inherited from B. This corresponds to a *recombination* event.

The process continues until there is a single line of ancestry. This point is commonly referred to as the most recent common ancestor (MRCA) of the sample.

A detailed description of the ARG can be found in Griffiths and Marjoram (1996). We use this process, a model of the evolution of the population of interest, as the basis of our MCMC approach. This approach explicitly allows for the non-independence of multiple loci. Furthermore, the method does not assume the mutation that leads to the phenotype of interest occurred only once, or at a fixed point in time, and also allows for recurrent mutation.

Our analysis uses SNP marker data. Since our theory is neutral, this was generated by including all polymorphic sites in the non-coding regions of the GENE6 sequence. We suppose there is a single functional mutation affecting Q1 and that it has three possible genotypes: 00, 01, and 11. We further assume that, conditional on its genotype g at the trait locus, the value of Q1 is Normally distributed. We model mutation by assuming all markers, and the site influencing the trait, mutate according to a symmetric mutation model. In brief, this means mutation from type 0 to type 1 occurs at the same rate as mutation from type 1 to type 0. (This is for simplicity only, and is not a crucial restriction of the method.) We allow the trait locus to mutate at a different rate than that for the markers and we label the wild-type at all markers as 0.

Our analysis assumes the data represent a random sample of individuals. In order to construct such a sample we extract all 133 founders and marry-ins that could be unambiguously haplotyped from the first replicate of the

simulated data. We chose to analyze MG1, one of the two genes that influence Q1. In order to remove extraneous factors we fit a linear regression to allow for the effects of Age, Sex and E1. We then use the residuals, *R*, for Q1 after the regression has been fitted, as well as *M*, the SNP marker data for the n individuals in our sample. Time constraints prohibit an analysis of the entire data set, so samples of reduced size were chosen. In this paper we present results of an analysis of multiple samples of 10 individuals each. These samples were drawn from the set of 133 individuals that were haplotyped earlier. In each case we extract all marker loci that are polymorphic in the sample, and we *exclude* the data at the functional site. Other relevant parameters are r, the rate of recombination; n, the rate of mutation; *N*, the population size; $m = (m_{00},m_{01},m_{11})$, the means of the Normal distributions determining the values of the quantitative trait residuals R; and *L*, the location of the mutation that influences Q1. In order to ensure identifiability, we order the means such that $m_{00} \leq m_{01} \leq m_{11}$. Our method proceeds by exploring the space of all possible ARGs and parameter values that may have led to the data. Due to the enormous dimension of the state-space we use an MCMC approach. In order to reflect a situation in which there might be uncertainty about the rates of recombination and mutation, or the exact nature of the dependence of the value of Q1 on MG1, we obtain $f(L, A, g, r, n, m \mid R, M, N)$, the posterior distribution of the location of MG1, as well as the ARG *A* and other parameters of interest, given the marker data and residuals for Q1. The posterior distribution of the location of MG1 can be obtained from the marginal of this distribution. We adopt a Bayesian approach, since coalescent theory gives us a natural prior for the distribution of the ARG. By taking Uniform prior distributions we can calculate maximum likelihood estimates of *L* (or r,n, and m).

We use a Metropolis-Hastings algorithm (cf. Gilks et al. (1996)) to simulate from the required conditional distribution. Informally, the method explores the state-space corresponding to $L=(L, A, g, r, n, m)$ by proposing a series of small changes to the values of one or more of these parameters. The changes are proposed according to a transition kernel $Q(L \to L')$ and are accepted or rejected according to a probability known as the Hastings ratio. By iterating this process until the chain of current states becomes stationary, and then sampling from this chain, one can approximate independent realizations from the distribution of interest. For details in a related application see Marjoram et al. (2000). Our transition kernel proposes changes to the ARG and to the recombination and mutation rates in an identical way to Marjoram et al. (2000). In the current context, we need also to explore the space of possible gene locations, genotypes at MG1, and (m_{00},m_{01},m_{11}). Suppose the functional mutation is currently in location *x*. We propose a new location $y \sim U(0,l)$, where *l* is the length of the sequence. We propose changes to the genotypes at the functional mutation by picking a sequence at random and flipping its type from 0 to 1, or vice-versa. Changes to (m_{00},m_{01},m_{11}) are made by picking one of them at random and proposing a new value according to a Uniform(C, D) distribution. We do this in such a way that the means retain their ordering. (*C* and *D* are arbitrary lower and upper bounds for the means, and do not impact our results.)

RESULTS

In general, analyses in which we tried to estimate the genotypes at the functional mutation, as well as its location, were unstable and had poor sampling properties. This may in part be because of the small sample size. We suspect that a more efficient proposal distribution will solve this problem, but in the current paper we propose a simpler solution. The initial values of (m_{00},m_{01},m_{11}) are set to be the 25th, 50th and 75th percentiles of the distribution of the residuals for Q1 in the entire sample of 133. Each individual in a given sub-sample is then assigned a fixed genotype *g*, where *g* is the genotype corresponding to the value of m_g closest to the residual trait value for that individual. These genotypes are kept fixed, while the means of the distributions for Q1 are still allowed to vary. These analyses have much better behavior. We present representative output for four such analyses in Figure 2. Each graph shows the posterior distribution for the location of the functional mutation. Approximately 50% of the samples lead to a posterior distribution that accurately reflects the location of MG1. Nonetheless, these results are encouraging, especially considering the small sample size. The width of the peak present in most of the distributions is indicative of the presence of several markers that segregate across the same individuals in this area; cf. Thomas et al. (2000). We note that, due to the small size of the samples and the prevalence of the functional mutation, not all samples will contain a signal as to the location of MG1.

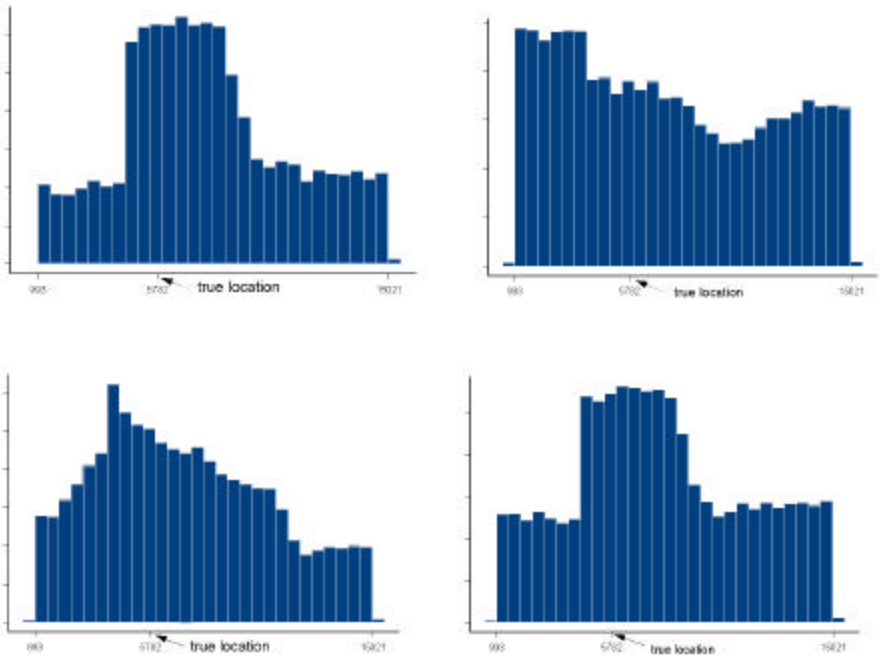

Figure 2: Posterior distribution of the location of the functional mutation for random samples of 10 individuals.

DISCUSSION

We have presented results of a pilot study designed to assess the potential of methods based upon ancestral inference using the coalescent applied to gene mapping. While we regard the results as encouraging, it should be stressed that this is a preliminary study. We feel the results demonstrate the potential of coalescent methods based upon multiple markers for random samples of individuals, and note that even with samples sizes as small as ten, we often still see informative posterior distributions. Of course, this may be a consequence of a large signal present in the entire data set. In the longer term we wish to perform a more comprehensive analyses using larger samples. (As has been previously demonstrated in the coalescent literature, we see that it is, loosely speaking, more informative to use extra markers rather than sequence more individuals.) We conditioned on the fact that the markers were segregating in each sample, but we ignored the fact that there was no variation between the markers. The effects of this ascertainment problem remain to be investigated. In some sense our analysis replaces assumptions common to other methods with a new set of model-based assumptions. The relative merits of different approaches may depend on the particular application or population demographics. Furthermore, we have not included the effects of gene-conversion. In the present context the effects of gene-conversion and recombination are similar, so the omission is likely to be of little consequence. In principle, it is straightforward to generalize our approach to include gene-conversion, or multiple functional mutations, but this was not possible in the time available.

ACKNOWLEDGEMENTS

The authors were supported in part by R01 grant GM 58897 from the National Institutes of Health.
ST was also supported in part by grant DBI 9504393 from the National Science Foundation.

REFERENCES

Clayton D (2000): Linkage disequilibrium mapping of disease susceptibility genes in human populations. International Statistical Review 68:23-43.
Devlin B, Risch N (1995): A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-322.
Gilks WR, Richardson S, Spiegelhalter DJ (eds.) (1996): Markov chain Monte Carlo in practice. Chapman and Hall.
Griffiths R, Marjoram P (1996): Ancestral inference from samples of DNA sequences with recombination. J Comp Biol 3:479-502.
Griffiths RC, Tavaré S (1994): Ancestral inference in population genetics. Statistical Science 9:307-319.
Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander ES (1992): Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genetics 2:204-211
Hudson RR (1983): Properties of a neutral allele model with intragenic recombination. Theoret. Popul.Biol. 23:183-201.
Hudson RR (1991): Gene genealogies and the coalescent process. In Oxford Surveys in Evolutionary Biology, ed. Futuyma D, Antonovics J, 7:1-44. Oxford University Press.
Kingman JFC (1982): On the genealogy of large populations. J Appl Prob 19A:27-43.
Kuhner MK, Yamato J, Felsenstein J (1995): Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140:1421-1430.
Kuhner MK, Yamato J, Felsenstein J (2000): Maximum likelihood estimation of recombination rates from population data. Genetics, *submitted*.
Marjoram P, Markovtsova L, Tavaré S (2000): Estimation in ancestral recombination graphs using Markov chain Monte Carlo. *Research Report* (available at http://hto-e.usc.edu)
McPeek MS, Strahs A (1999): Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858-875.
Morris AP, Whittaker JC, Balding DJ (2000): Bayesian fine-scale mapping of disease loci, by hidden Markov models. Am J Hum Genet 67:155-169.
Nielsen R (2000): Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931-942.
Stephens M, Donnelly P (2000): Inference in molecular population genetics. J. Royal Statist. Soc. B., in press.
Thomas DC, Morrison J, Clayton D (2000): Bayes estimates of haplotype effects. (This volume.)