HAPGEN2: simulation of multiple disease SNPs
Zhan Su1,*, Jonathan Marchini1,2,† and Peter Donnelly1,2,†
1Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN and 2Department of Statistics, University of Oxford, Oxford OX1 3TG, UK
Associate Editor: Jeffrey Barrett

1 INTRODUCTION

Genome-wide association studies have become a powerful approach for uncovering the genetic variants that impact human phenotypes. Simulation studies are a popular and inexpensive approach to evaluate new methods for statistical analysis (Su et al., 2009) and to examine the power of different experimental designs (Spencer et al., 2009).

The traditional approach of simulating a population forwards (Lambert, 2008) or backwards (Hudson, 2002) in time ignores the large amount of observed genetic data that are available, can be computationally intensive and can struggle to match real LD patterns. To overcome these problems, Spencer et al. (2009) introduced a novel simulation approach, HAPGEN, which uses an alternative resampling approach. Given a reference panel of haplotypes, this method produces a sample of haplotypes with patterns of LD similar to those in the reference panel. Using the haplotypes, this method produces a sample of haplotypes with

Results: We introduce a new simulation algorithm based on a successful resampling method, HAPGEN, that can simulate multiple nearby disease SNPs on the same chromosome. The new method, HAPGEN2, retains many advantages of resampling methods and expands the range of disease models that current simulators offer.

Availability: HAPGEN2 is freely available from http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html.

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 31, 2011; revised on May 22, 2011; accepted on May 24, 2011

2 METHODS

The HAPGEN simulation approach is similar to that of HAPGEN and is based on the Li and Stephens (LS) model (Li and Stephens, 2003) of LD. Briefly, given a reference panel of haplotypes, \( h^0 \) (\( h_1, \ldots, h_n \)) as input, each haplotype is typed at \( d \) biallelic sites, that is \( h_i = (h_{i1}, \ldots, h_{in}) \) and \( h_{ij} \in \{0,1\} \), the LS model models each newly simulated haplotype as an imperfect mosaic of the haplotypes in \( h^0 \) and the haplotypes that have already been simulated (see below for more details). Simulation of case–control data is based on a set of disease SNPs, \( D = \{d_1, d_2 \in \{1, \ldots, l_1 \}, l_1 = 1, \ldots, K \} \) with effect sizes and \( \rho = \{\rho_1, \rho_2\} \), where \( \rho_1 \) and \( \rho_2 \) are the disease risks of carrying one and two copies of the 1 allele relative to carrying

The haplotypes, \( H^K = \{h_{11}, \ldots, h_{1k}\} \), for the control individuals are simulated first, followed by the haplotypes, \( H^K = \{h_{21}, \ldots, h_{2k}\} \), for the case individuals.

2.1 Simulating control data

We simulate the control data as population controls (so that some of them may be cases) and simulate each additional haplotype, \( h_{ij \in K} \), sequentially under the LS model. We use the copying states, \( z_{ij} \in \{0,1\} \) which evolve in a Markov manner, to indicate the haplotype that \( h_{ij} \) copies at site \( j \). We simulate each haplotype in three stages. First, the allele at each SNP is simulated according to the transition probabilities

\[
\text{Pr}(z_{ij+1} = 0 | z_{ij+1} = 1) = \left(1 - \exp\left(-\frac{\rho_1}{\theta}\right)\right), \quad \text{Pr}(z_{ij+1} = 1 | z_{ij+1} = 0) = \exp\left(-\frac{\rho_2}{\theta}\right),
\]

where \( \theta = \frac{1}{1 - z} \) and \( \rho_1 \) is genetic distance between SNPs \( (j-1) \) and \( j \). Conceptually, the cross-over events mimics the effect of recombination and breaks up \( h_{ij} \) into independent segments, \( (h_{i1}, h_{i2}, \ldots, h_{i(j-1)}) \) and \( (h_{i(j+1)}, \ldots, h_{ik}) \), where each segment is a haplotype of SNPs between two cross-over events. Second, the copying state for each segment is sampled uniformly from \( \{1, \ldots, l_1\} \). Finally, the allele at each SNP is simulated conditional on the copying state and a mutation parameter \( \mu \),

\[
p(h_{ij} = 1 | h_{ij} = 0, z_{ij}) = \exp(-\mu z_{ij}), \quad p(h_{ij} = 0 | h_{ij} = 1, z_{ij}) = \exp(-\mu (1 - z_{ij})).
\]

Spencer et al. (2009) found that \( \mu = \frac{\rho_1}{\theta} \), simulated amounts of novel haplotype variation similar to data simulated under the coalescent model.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.
2.2 Simulating case data

We simulate the case haplotypes in a similar way, but we simulate them sequentially in pairs (with each pair corresponding to a case individual) and oversample haplotypes carrying the risk alleles based on the relative risks.

Simulation of each haplotype pair, \((h_{i+1}, h_{i+2})\) in \(H^n\), proceeds in four stages. First, the cross-over events are simulated in the same way as for the controls, according to (1). Second, the alleles at the disease SNPs are simulated. Let \((h_{i+1}', h_{i+2}')\) be the subset of \((h_{i+1}, h_{i+2})\) that consist of the alleles at the disease SNPs, so that \(h_{j}' = (h_{i+1+1}', \ldots, h_{i+1+2j}')\) for \(j = 1, 2, \ldots\). The cross-over events separate \(h_{i+1}'\) and \(h_{i+2}'\) into segments, \(h_{1}'\), \(h_{2}'\), \(h_{3}'\), \(h_{4}'\), \(h_{5}'\), \(h_{6}'\), \(h_{7}'\), \(h_{8}'\), \(h_{9}'\), and \(h_{10}'\). We simulate \((h_{1}', h_{2}')\) from its joint distribution, which is calculated from the relative risks and the marginal frequencies of each segment in \(H^n\) and \(H^0\), using Bayes’ Theorem:

\[
p(h_{1}', h_{2}') \propto p(\text{case}|h_{1}, h_{2}) \times p(h_{1}', h_{2}')/p(h_{1}, h_{2})
\]

where \(s_{i} = h_{1} + h_{2}\) is the genotype at \(d_{i}\), and \(p(h_{1}, h_{2})\) is the frequency of the haplotype segment \(h_{i}\) in \(H^n\) and \(H^0\). Third, the copying state for each segment, \(h_{i+1,j}\), is simulated independently and is drawn uniformly from \([1, \ldots, k]\), like we do for the controls, if \(s\) does not include any disease SNPs; or else it is drawn from

\[
P(h_{i+1,j} = 1) \propto \prod_{k=1}^{k} \left(1 - p_{i+1+k}\right) + \prod_{k=1}^{k} p_{i+1+k} \times \prod_{k=1}^{k} p(h_{2}', h_{3}),
\]

\[
\alpha \propto \prod_{k=1}^{k} \left(1 - p_{i+1+k}\right) + \prod_{k=1}^{k} p_{i+1+k} \times \prod_{k=1}^{k} p(h_{2}', h_{3}),
\]

where \(d_{i,1} = h_{i+1,j}\) if \(h_{i+1,j} = h_{i+1,j}'\) and 0 otherwise. Finally, each allele for \(h_{i+1,j}\) is simulated according to (2). Copying states and alleles for \(h_{i+2}\) are simulated in the same way.

3 RESULTS

To demonstrate HAPGEN2, we have simulated, using HapMap2 CEU as the reference panel, 2000 cases and 2000 controls at 880 SNPs across a 700 kb region on chromosome 21, with 3 disease SNPs, at positions \(d_1 = 25356790, d_2 = 25390071\) and \(d_3 = 25691378\), each under a log-additive disease model with a heterozygote relative risk of 1.3. The simulation process took \(<10s\) on a 2.93 GHz processor laptop, and will increase linearly with the number of SNPs and individuals.

Figure 1, produced by HAPLOVIEW (Barrett et al., 2005), shows the similarity between the LD patterns of the reference panel (top) and the simulated haplotypes (bottom). The top plot in Figure 2 shows the --\(\log_{10}(P\text{-values})\), for the log-additive test, across the region, indicating the signal of association at the disease SNPs; subsequent plots show the \(P\)-values conditioned on the genotypes at \(d_1\), \(d_2\) and at \(d_1\) and \(d_2\), respectively, confirming that there are indeed three independent disease SNPs.

4 DISCUSSION

We have introduced a new resampling method that can simulate multiple disease SNPs on the same haplotype, which will be particularly useful for investigating disease models involving multiple disease SNPs within close proximity. HAPGEN2 is fast, simple to use and available as a C++ package from http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html, along with instructions and supporting resources, such as recombination rates, HapMap and 1000G reference panels.

REFERENCES


