Bioinformatics Advance Access published June 8, 2011

HAPGEN2: simulation of multiple disease SNPs.

Zhan Su [1]*, Jonathan Marchini[2,1]† and Peter Donnelly[1,2]†

[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK.
[2]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.

Associate Editor: Dr. Jeffrey Barrett

ABSTRACT

Motivation: Performing experiments with simulated data is an inexpensive approach to evaluating competing experimental designs and analysis methods in genome-wide association studies. Simulation based on resampling known haplotypes is fast and efficient and can produce samples with patterns of linkage disequilibrium (LD), which mimic those in real data. However, the inability of current methods to simulate multiple nearby disease SNPs on the same chromosome can limit their application.

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.

Contact: zhan@well.ox.ac.uk

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 ignore 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 HapMap3 and 1000G haplotype data as reference panels, HAPGEN is able to simulate data for many populations. In addition, it is fast and can simulate a single disease SNP under a general disease model, allowing the user to specify the risk allele and heterozygote and homozygote relative risks. Other resampling methods also exist (Wright *et al*., 2007; Li and Li, 2008), but they and HAPGEN can only simulate a single disease SNP on the same haplotype. There are many complex diseases with multiple associated loci on the same chromosome, some of them in close proximity (e.g. Strange *et al*. (2010)), so the ability to simulate multiple disease SNPs on the same chromosome would be desirable. To address this issue, we have devised a new approach, extending HAPGEN, to simulate multiple nearby disease SNPs on the same chromosome.

*to whom correspondence should be addressed

†both authors equally directed this work

2 METHODS

The HAPGEN2 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, $H^R = \{h_1, \ldots, h_r\}$ as input, where each haplotype is typed at L bi-allelic sites, that is $h_i = (h_{(i,1)}, \ldots, h_{(i,L)})$ and $h_{(i,j)} \in \{0, 1\}$, the LS model models each newly simulated haplotype as an imperfect mosiac of the haplotypes in H^R 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, $D = \{d_k : d_k \in \{1, \ldots, L\}, k = 1, \ldots, K\}$ with effect sizes $RR = \{(rr_k^1, rr_k^2)\}$, where rr_k^1 and rr_k^2 are the disease risks of carrying one and two copies of the 1 allele relative to carrying two copies of the 0 allele at d_k, which combine multiplicatively across the K disease SNPs. The haplotypes, $H^P = \{h_{r+1}, \ldots, h_p\}$, for the control individuals are simulated first, followed by the haplotypes, $H^Q = \{h_{p+1}, \ldots, h_q\}$, for the case individuals.

Simulating control data

We simulate the control data as population controls (so that some of them may be cases) and simulate each additional haplotype, $h_{i+1} \in H^P$, sequentially under the LS model. We use the copying states, $z_{(i+1,j)} \in \{1, \ldots, i\}$, which evolve in a Markov manner, to indicate the haplotype that $h_{(i+1,j)}$ copies at site j. We simulate each haplotype in three stages. First, the crossover events, which are locations where $z_{(i+1,j)} \neq z_{(i+1,j-1)}$, are simulated according to the transition probabilities

$$P(z_{(i+1,j)} = z | z_{(i+1,j-1)}) = \frac{(1 - \exp(-\frac{\rho_j}{i}))}{i} + \exp(-\frac{\rho_j}{i}) I_z, \quad (1)$$

where I_z is 1 if $z = z_{(i+1,j-1)}$ and 0 otherwise, and ρ_j is genetic distance between SNPs $(j-1)$ and j. Conceptually, the cross-over events mimicks the effect of recombination and breaks up h_{i+1} into independent segments, $\{h_{(i+1,s_1)}, \ldots, h_{(i+1,s_n)}\}$, where each segment is a haplotype of SNPs between two crossover events. Second, the copying state for each segment is sampled uniformly from $\{1, \ldots, i\}$. Finally, the allele at each SNP is simulated conditional on the copying state and a mutation parameter μ_i:

$$p(h_{(i+1,j)} = h_{(z,j)} | z_{(i+1,j)} = z) = 1 - \mu_i. \quad (2)$$

Spencer *et al*. (2009) found that $\mu_i = \frac{\theta}{2(i+\theta)}$, where $\theta = \frac{1}{\sum_{i=1}^{m} \frac{1}{n}}$, simulated amounts of novel haplotype variation similar to data simulated under the coalescent model.

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, $(h_{i+1}, h_{i+2}) \in H^Q$, proceeds in four stages. First, the crossover 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_D^1, h_D^2) be the subset of (h_{i+1}, h_{i+2}) that consist of the alleles at the disease SNPs, so that $h_D^j = (h_{(i+j,d_1)}, \ldots, h_{(i+j,d_k)})$ for $j = 1, 2$. The crossover events separate h_D^1 and h_D^2 into segments, $\{h^1_{s_1^1}, \ldots, h^1_{s_{n_1}^1}\}$ and $\{h^2_{s_1^2}, \ldots, h^2_{s_{n_2}^2}\}$. We simulate (h_D^1, h_D^2) from its joint distribution, which is calculated from the relative risks and the marginal frequencies of each segment in H^P and H^R, using Bayes Theorem:

$$\begin{aligned} p((h_D^1, h_D^2)|\text{case}) &\propto p(\text{case}|(h_D^1, h_D^2)) * p(h_D^1, h_D^2) \\ &= \prod_{k=1}^{K} p(\text{case}|g_{d_k}) * p(h_D^1) * p(h_D^2) \\ &\propto \Big(\prod_{k=1}^{K} rr_k^{g_{d_k}}\Big) * \prod_{i=1}^{n_1} p(h^1_{s_i^1}) * \prod_{j=1}^{n_2} p(h^2_{s_j^2}), \end{aligned}$$

where $g_{d_k} = h^1_{d_k} + h^2_{d_k}$ is the genotype at d_k, and $p(h_s)$ is the frequency of the haplotype segment h_s in H^R and H^P. Third, the copying state for each segment, $h_{(i+1,s)}$, is simulated independently and is drawn uniformly from $\{1, \ldots, i\}$, like we do for the controls, if s does not include any disease SNPs; or else it is drawn from

$$P(z_{(i+1,j)} = z) \propto \prod_{d_k : d_k \in s} \mu_{(i+1)}^{(1-I_{d_k})} * (1 - \mu_{(i+1)})^{I_{d_k}} \quad \forall j \in s,$$

where I_{d_k} is 1 if $h_{(i+1,d_k)} = h_{(z,d_k)}$ and 0 otherwise. Finally, each allele for $h_{(i+1,s)}$ 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 700kb 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 less than 10 seconds on a 2.93 GHz processor laptop, and will increase linearly with the number of SNPs and individuals.

Figure 1, produced by HAPLOVIEW (Barrett, 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-values), for the log-additive test, across the region, illustrating the signal of association at the disease SNPs; subsequent plots show the p-values conditioned on the genotypes at d_1, at d_1 and d_2, and at d_1, d_2 and d_3 respectively, confirming that there are indeed 3 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.

Fig. 1. LD patterns, in terms of r^2, in the HapMap reference haplotypes (top) and the simulated haplotypes (bottom).

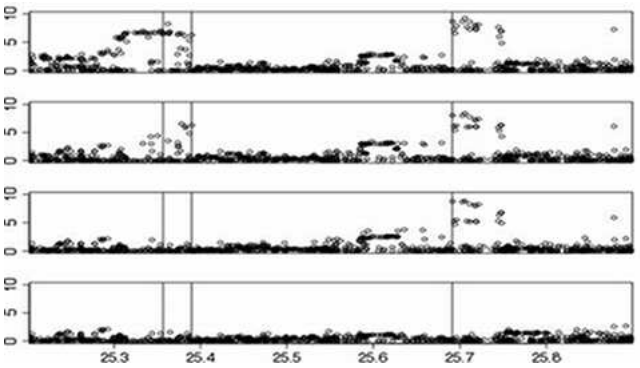

Fig. 2. Top plot shows the $-\log_{10}$(p-values) under the log-additive test at each SNP in the simulated data. The location of the disease SNPs, d_1, d_2, d_3, are indicated (from left to right) by the vertical lines. Subsequent plots (from the top) show the p-values conditioned on the genotypes at d_1, at d_1 and d_2, and at d_1, d_2 and d_3.

The model described here can be easily extended to simulate interacting disease SNPs (we currently provide an R package that does this) and admixture (using reference panels from multiple populations), which we hope to implement in the future.

REFERENCES

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*, **21**, 263-265.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model, *Bioinformatics*, **18**, 337-338.

Lambert, B. W., Terwilliger, J. D. and Weiss, K.M. (2002) ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth, *Bioinformatics*, **24**, 1821-1822.

Li, C. and Li, M. (2008) GWAsimulator: a rapid whole-genome simulation program, *Bioinformatics*, **24(1)**, 140-142.

Li, N. and Stephens, M. (2003) Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data,*Genetics*, **165**, 2213-2233.

Spencer, C. C. A., Su, Z., Donnelly, P. and Marchini, J. (2009) Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip, *PLoS Genetics*, **5(5)**.

Su, Z., Cardin, N., WTCCC, Donnelly, P. and Marchini, J. (2009) A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies, *Statistical Science*, **24(4)**, 430-450.

Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., (...), Zou, F. (2007) Simulating association studies: a data-based resampling method for candidate regions or whole genome scans, *Bioinformatics*, **23(19)**, 2581-2588.

Strange, A., Capon, F., Spencer, C.C.A., Knight, J., (...), Trembath, R.C. (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1, *Nature Genetics*, **42**, 985-990.