BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 19 2007, pages 2581–2588
doi:10.1093/bioinformatics/btm386

Genetics and population analysis

Simulating association studies: a data-based resampling method for candidate regions or whole genome scans

Fred A. Wright[1,2,3,*], Hanwen Huang[1], Xiaojun Guan[4], Kevin Gamiel[4], Clark Jeffries[4,5], William T. Barry[1], Fernando Pardo-Manuel de Villena[2,6], Patrick F. Sullivan[2,6], Kirk C. Wilhelmsen[2,6] and Fei Zou[1,2,3]

[1]Department of Biostatistics, [2]Center for Genome Sciences, [3]Center for Environmental Bioinformatics, University of North Carolina, Chapel Hill, NC 27599, [4]Renaissance Computing Institute, Europa Drive, [5]School of Pharmacy and [6]Department of Genetics, UNC Chapel Hill, NC, USA

Received on August 25, 2006; revised on June 21, 2007; accepted on July 20, 2007

Advance Access publication September 4, 2007

Associate Editor: Keith Crandall

ABSTRACT

Motivation: Reductions in genotyping costs have heightened interest in performing whole genome association scans and in the fine mapping of candidate regions. Improvements in study design and analytic techniques will require the simulation of datasets with realistic patterns of linkage disequilibrium and allele frequencies for typed SNPs.

Methods: We describe a general approach to simulate genotyped datasets for standard case-control or affected child trio data, by resampling from existing phased datasets. The approach allows for considerable flexibility in disease models, potentially involving a large number of interacting loci. The method is most applicable for diseases caused by common variants that have not been under strong selection, a class specifically targeted by the International HapMap project.

Results: Using the three population Phase I/II HapMap data as a testbed for our approach, we have implemented the approach in HAP-SAMPLE, a web-based simulation tool.

Availability: The web-based tool is available at http://www.hapsample.org

Contact: fwright@bios.unc.edu; fzou@bios.unc.edu; kirk@med.unc.edu

1 INTRODUCTION

It has long been recognized (Risch and Merikangas, 1996) that large-scale genotype–phenotype association studies will have great power and precision to elucidate genetic influences in complex disease (Gibbs *et al*., 2003; Hirschhorn and Daly, 2005). However, key issues in optimal design and analysis remain unresolved. A partial list of areas of active research (Hirschhorn and Daly, 2005) includes reassessment of the relative strengths of case-control versus family based designs (Hintsanen *et al*., 2006), design of multistage association studies (Lowe *et al*., 2004; Satagopan *et al*., 2004), selection of appropriate significance thresholds (Dudbridge and Koeleman, 2004; Thomas *et al*., 2005), methods for fine-mapping and reconstructing haplotypes (De La Chapelle and Wright, 1998; Stephens and Donnelly, 2003) and approaches for handling multiple interacting susceptibility loci (Marchini *et al*., 2005).

In many cases, the best approaches depend on specifics of the disease model and polymorphism in the population (Pritchard and Cox, 2002). In order to rigorously compare competing approaches, simulation studies must be performed which provide realistic patterns of allele frequencies and linkage disequilibrium (LD) structure. Unfortunately, uncertainty of human population genetic history makes it difficult to perform such simulations. Forward simulation approaches (Dudek *et al*., 2006; Peng *et al*., 2007) can be sensitive to underlying assumptions and starting genotypes, and are typically highly variable across simulations (Calafell *et al*., 2000) for observed LD and disease outcomes. Backward coalescent approaches for multiple loci (Laval and Excoffier, 2004; Posada and Wiuf, 2003; Wang and Rannala, 2005) can be 'calibrated' to fit observed data structures (Schaffner *et al*., 2005), but remain computationally infeasible for dense SNP collections spanning large genomic regions. Moreover, coalescent methods are not well suited to handle unknown and variable selection pressures that may have affected broad genomic regions (Altshuler *et al*., 2005). These approaches involve *de novo* simulation of artificial SNPs, while the researcher may be interested in simulation tailored to a certain genomic region or the actual list of SNPs from a favored genotyping platform. Alternatively, one might simulate SNPs to fit pairwise LD measures observed in real data (Montana, 2005), but this approach may be unable to reflect higher-order haplotype structures likely crucial for evaluating haplotype reconstruction and inference (Liu and Lin, 2005). Moreover, it is not clear how to incorporate disease models into this approach.

Another possible data-based approach involves cataloguing the frequency of inferred haplotypes in real data across regions constituting haplotype blocks (Altshuler *et al*., 2005), and

*To whom correspondence should be addressed.

resampling from these haplotypes. The specification of distinct block boundaries is somewhat artificial (Schwartz *et al.*, 2003), and may not reflect longer-range LD that is apparent in real data. By sampling from haplotypes of very long range, we may avoid the problem of applying arbitrary haplotype block definitions. Recent work (de Bakker *et al.*, 2005) employed resampling data across the 500-kb HapMap ENCODE regions (Feingold *et al.*, 2004), but it is not clear how to extend these efforts to a large scale or how to flexibly specify disease models.

Many complex diseases are likely to be influenced by ancient SNP variants that are common, and maintain appreciable frequencies across continent-level populations (Altshuler *et al.*, 2005; Lohmueller *et al.*, 2003; Peng and Kimmel, 2007). Variants with low penetrance or predisposing for diseases of old age will not have undergone strong selection, and investigation of this class of diseases is among the motivations for the HapMap project (Altshuler *et al.*, 2005; Gibbs *et al.*, 2003). Under such a model, disease chromosomes may be thought of as drawn from the same population as control chromosomes, but with selection probabilities that differ from controls at causal disease loci.

With these considerations, we developed a method to simulate realistic human autosomal SNP data for disease association studies, by resampling chromosome-length haplotypes derived from real data. The simulated data follows observed linkage disequilibrium structure and allele frequencies at actual SNP loci, and thus is well suited for power analyses and investigations of competing techniques for study design and analysis. We started by assuming that phased SNP data are available at a series of loci from a sample of individuals. Assuming random mating, the individual typed chromosomes form the relevant pool from which we draw in order to simulate SNP alleles for new individuals. We further implemented an artificial 'crossover' process that allows recombination of chromosomes at simulated crossovers. This crossover process mimics meiosis, but is arguably not necessary, as the original chromosome sample is already reflective of the population. However, we reasoned that a modest crossover process would increase novelty and avoid long-range allelic association produced by chance variation or subtle population substructure. As described below, we take care to simulate crossovers in a manner that preserves haplotype block structure, and the crossover rate is controlled by the user.

Figure 1 shows a schematic of the approach. The user identifies one or more 'disease' SNPs in the pool data (currently limited to one per autosome), for which ascertainment of cases determines the genotype probabilities. Although each pool chromosome is a high-density haplotype of allele values, only the disease SNP is shown in the figure. We denote the genotype at the jth disease locus by g_j and the joint genotypes for the L disease loci by $\boldsymbol{g} = \{g_1, \ldots, g_L\}$. We use $D = 1$ to denote a case individual, $D = 0$ to denote control. At the direct level of sampling from the chromosome pool, we must specify $P(\boldsymbol{g}|D = 1)$ for each $\boldsymbol{g}$. For our software, we also offer alternate means of specifying the disease model, in terms of genotype relative risks or absolute disease risks as described in the Methods Section. A random $\boldsymbol{g}$ is drawn from $P(\boldsymbol{g}|D = 1)$, and for each disease locus j two case chromosomes are drawn from the pool to achieve genotype g_j. Sampling is performed with

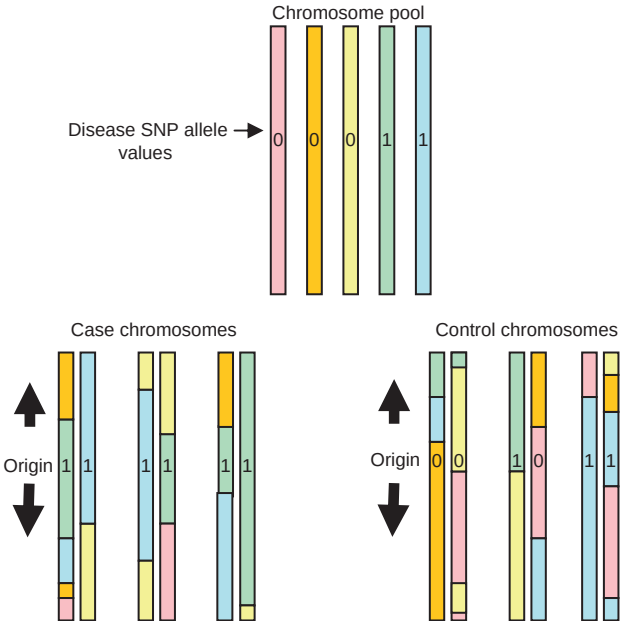

Fig. 1. Simulation of case and control chromosomes. *Chromosome pool*. These are derived from HapMap or other source as chromosome-length haplotypes. *Case chromosomes*. Genotypes at the disease SNP are determined according to $P(\boldsymbol{g}|D = 1)$, and pool chromosomes are chosen to be compatible with the genotype. Then the artificial crossover process is simulated, following the process described in the text, using the disease SNP location as the origin. The example depicted here shows a recessive disease, for which two '1' alleles are required. *Control chromosomes*. Genotypes at the disease SNP are determined according to $P(\boldsymbol{g}|D = 0)$, and otherwise the process is the same as with case chromosomes. Simulation of affected-child trio data proceeds similarly, with transmitted chromosomes simulated in the same manner as case chromosomes. Non-transmitted chromosomes in the trios are simulated in a similar manner to control chromosomes, but follow the unconditional genotype frequencies $P(\boldsymbol{g})$ at the disease loci.

replacement, as homozygosity of short-range haplotypes may arise in real data from a single shared ancestry. For low penetrance diseases, this conditional sampling scheme is far more efficient than unconditionally generating large numbers of genotypes and retaining only the small fraction with disease. The remaining allele values are generated by following an artificial crossover process using the disease locus as the origin (see Methods Section). At each generated crossover, a random chromosome from the pool is used to continue extending the haplotype. Final genotypes are created by combining the two haplotypes for each individual.

Control chromosomes are generated similarly, following $P(\boldsymbol{g}|D = 0)$ at the origin. These values differ from the unconditional $P(\boldsymbol{g})$ for diseases with high prevalence, appropriately reflecting that control individuals exhibit an excess of low-risk alleles at disease loci. Our approach automatically computes the necessary $P(\boldsymbol{g}|D = 0)$ values from the disease model specification and the disease prevalence (see Methods Section). Autosomes not containing disease loci are simulated for both case and control by randomly sampling from the pool. Any crossover origin may be used for such chromosomes, and we start at the p-terminus.

Using the above mechanism, we may similarly generate data for affected-child trios. In doing so, we simulate 'transmitted' versus 'non-transmitted' chromosomes (Falk and Rubinstein, 1987), obviating the need for explicit simulation of the meiotic events in the trio. Assuming that trio ascertainment is based on the child, the child's (transmitted) chromosomes are simulated in the same manner as case chromosomes above. Non-transmitted chromosomes are simulated using the unconditional probabilities $P(\boldsymbol{g})$ at the disease loci, because the selection mechanism (case status of child) does not influence the genotype probabilities for non-transmitted chromosomes. Transmitted and non-transmitted chromosomes are then used to obtain genotypes for the trio.

2 METHODS

2.1 A HapMap-based pool

We currently use three separate chromosome pools derived from the HapMap population samples. These consist of (i) 30 parent–child trios from Utah, USA, with ancestry from northern and western Europe (CEU); (ii) data from 45 unrelated Japanese in Tokyo, Japan (JPT), and 45 unrelated Han Chinese in Beijing, China (CHB), combined as JPT + CHB and (iii) 30 parent-child trios of Yoruba people from Ibadan, Nigeria (YRI). These samples provide an ideal initial dataset for our approach. The samples were typed at ~1 million autosomal SNPs for the Phase I HapMap freeze and ~3.6 million SNPs for Phase II (containing Phase I as a subset). The data contain mostly common SNP variants (Altshuler *et al.*, 2005), and most SNPs from major genotyping platforms are represented (Barrett and Cardon, 2006; Matsuzaki *et al.*, 2004). Our software uses haplotypes as released by the HapMap consortium, phased using the PHASE software (Stephens and Donnelly, 2003).

2.2 Simulation of case haplotypes

For the joint disease genotypes $\boldsymbol{g} = \{g_1, \ldots, g_L\}$, we currently assume that the L disease loci reside on separate chromosomes (handling multiple disease loci per chromosome is more complicated, and is the subject of future research). We estimate the unconditional population genotype probabilities $P(\boldsymbol{g})$, assuming Hardy–Weinberg equilibrium in the chromosome pool. Specifically, let i index the H haploid genomes in the pool. For the L disease loci, suppose the observed allele value for the ith haploid genome at the jth disease locus is $a_{ij} \in \{0,1\}$. Here '1' always denotes the minor allele. The observed control minor allele frequency is $p_j = \Sigma_i a_{ij}/H$. The genotype for an individual at the jth disease locus is denoted by the number of minor alleles, $g_j \in \{0,1,2\}$. The Hardy–Weinberg assumption is that $P(g_j = 0) = (1-p_j)^2$, $P(g_j = 1) = 2p_j\,(1-p_j)$ and $P(g_j = 2) = p_j^2$, and finally

$$P(\mathbf{g}) = \prod_{j=1}^{L} P(g_j).$$

Our approach is meaningful only for disease loci with minor allele frequency (MAF) > 0 in the original pool, and gives positive probabilities for every possible *joint* genotype for the L loci. In contrast, the original HapMap data may not contain all possible joint genotypes, due to table sparseness if L is large.

2.2.1 Absolute genotype (AG) specification Sampling from the disease model is ultimately performed using $P(\boldsymbol{g}|D=1)$. Using the input type we refer to as absolute genotype (AG) specification, the user directly specifies these probabilities, along with the disease prevalence $P(D=1)$. Although the AG format is conceptually straightforward, it is often more convenient to specify the model in terms of genotype relative risks (GRR) or absolute disease risks (AR). Our approach also accepts these latter two input types, which are automatically converted to the AG probabilities as described subsequently. For any of these input types there are a range of possible risk values, and values outside of this range can result in genotype probabilities outside the range (0,1). Our web tool described below automatically flags any such errors in risk specification.

2.2.2 Genotype relative risk (GRR) specification The user specifies the disease prevalence $P(D=1)$ and relative risks $\mathrm{RR}_g = P(D=1|\boldsymbol{g})/P(D=1|\boldsymbol{g}_0)$ for all $\boldsymbol{g}$ compared to a referent $\boldsymbol{g}_0$ (which has $\mathrm{RR}_{g0} = 1$).

We have

$$P(\mathbf{g}|D=1) = \frac{P(\mathbf{g})P(D=1|\mathbf{g})}{P(D=1)} = \frac{P(\mathbf{g})\mathrm{RR}\mathbf{g}}{P(D=1)} \times P(D=1|\mathbf{g}_0)$$

and the last term may be computed using

$$P(D=1|\mathbf{g}_0) = \frac{\sum_g P(\mathbf{g}|D=1)}{\sum_g P(\mathbf{g})\mathrm{RR}_g/P(D=1)} = \frac{P(D=1)}{\sum_g P(\mathbf{g})\mathrm{RR}_g}.$$

2.2.3 Absolute risk (AR) specification The user specifies $P(D=1|\boldsymbol{g})$ for all $\boldsymbol{g}$. We have

$$P(\mathbf{g}|D=1) = \frac{P(\mathbf{g})P(D=1|\mathbf{g})}{P(D=1)}$$

where

$$P(D=1) = \sum_{\mathbf{g}} P(D=1|\mathbf{g})P(\mathbf{g})$$

is calculated directly.

Note that any of the input types requires specifying the risks associated with each of the 3^L joint genotypes. This level of detail enables complete flexibility in specifying various penetrances and disease locus interactions, although for many purposes users may wish to specify only a single disease locus.

2.3 Control haplotypes

Using any of the input types, the AG values and the prevalence $P(D=1)$ are available from the previous subsection. We obtain the control genotype frequencies as follows. We have

$$P(\mathbf{g}) = P(\mathbf{g}|D=0)P(D=0) + P(\mathbf{g}|D=1)P(D=1)$$

and rearranging the terms gives

$$P(\mathbf{g}|D=0) = \frac{P(\mathbf{g}) - P(\mathbf{g}|D=1)P(D=1)}{P(D=0)},$$

for which all terms on the right-hand side are known. From these probabilities, the control chromosomes can be simulated in the same manner as case chromosomes. For rare diseases, the $P(\boldsymbol{g}|D=0)$ values will be very close to the unselected genotype probabilities $P(\boldsymbol{g})$.

2.4 Extreme-phenotype designs

Finally, we note that similar reasoning can be used to specify appropriate genotype probabilities for designs in which individuals with extreme phenotypes are chosen for genotyping. We assume that the investigator has a model $p(\delta|\boldsymbol{g})$, which is the density of a quantitative trait δ depending on the joint disease genotype. Then

$$P(\mathbf{g}|\delta > \delta_{\mathrm{upper}}) = \frac{P(\mathbf{g})P(\delta > \delta_{\mathrm{upper}}|\mathbf{g})}{P(\delta > \delta_{\mathrm{upper}})}$$

where

$$P(\delta > \delta_{\mathrm{upper}}) = \sum_{\mathbf{g}} P(\mathbf{g})P(\delta > \delta_{\mathrm{upper}}|\mathbf{g})$$

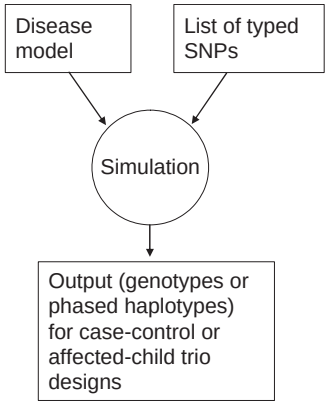

Fig. 2. Input/output schematic for HAP-SAMPLE. A disease model file specifies the disease SNPs and their associated risks, while another file lists the SNPs for which data are simulated. Output for case-control or affected-child trio designs is presented as genotypes or phased haplotypes.

for some critical upper phenotype value δ_{upper} that determines the sampling of individuals with 'high' trait values. The genotype probabilities can be used in an AG specification to simulate 'case high' individuals (and any simulated controls are discarded). Similarly, values $P(\boldsymbol{g}|\delta < \delta_{\text{lower}})$ for a lower threshold δ_{lower} can be used to simulate 'case low' individuals.

2.5 The HAP-SAMPLE tool

We have implemented our approach in a program called HAP-SAMPLE, which has a web-based interface for ease of use by the research community (Fig. 2, www.hapsample.org). Inputs consist of the disease model file (the various types of specifications are given above) with specific rs#'s for the disease SNPs, and a file listing SNPs to be 'genotyped'. The disease SNPs need not be among the genotyped SNPs, but both disease SNPs and typed SNPs must be available in the chromosome pool. Output consists of SNP genotypes, reflecting current typing technologies. In addition, the output contains the simulated haplotypes, which are useful in evaluating the success of haplotype reconstruction approaches. Simulation times for a sample of 1000 cases and 1000 controls ranges from a few seconds for a few dozen SNPs to a few minutes for 100 000 SNPs.

The default limit of SNPXgenotype observations has been set to 10^8 for each of the case and control groups. Thus, e.g. 1 million cases and 1 million controls can be simulated for 100 markers in a genomic region. These simulated individuals can then be split to form 1000 independent simulations of 1000 cases versus 1000 controls. Similarly, simulation of whole genome scans can also be performed individually and repeatedly, although investigators should contact the authors when planning numerous simulated whole genome scans.

2.6 The crossover process

Placing simulated crossovers in regions of high LD will tend to reduce LD in the simulated data compared to observed data. We guard against this possibility by mimicking meiosis down to fine scales, where the presence of haplotype block structure reflects variation in local meiotic recombination rates (Altshuler *et al.*, 2005; Myers *et al.*, 2005). We applied the LDMAP software (Maniatis *et al.*, 2002) to each phased HapMap population to produce three maps of linkage disequilibrium units (LDU). The approach has some similarities to studies of recombination hotspots (Li and Stephens, 2003; Myers *et al.*, 2005), and is straightforward for frequent updating. At large scales, the LDU map reflects the sex-averaged meiotic map (Kong *et al.*, 2002; Tapper *et al.*, 2005), but provides an interpolation at much finer scales using LD patterns observed in the HapMap data. HAP-SAMPLE users specify a desired average number of crossovers per centiMorgan, which are simulated according to a Poisson process (i.e. a Haldane map function). The crossovers are then converted from genetic location to physical location by the LDU map. The value 0.01 corresponds to the rate of a single meiosis, and in terms of crossovers is similar to randomly selecting pool chromosomes and performing mating for a single generation. We recommend using a higher crossover rate of 1.0 per cM, which reduces long-range LD, but for small distances does not reduce LD much below that observed in the pool.

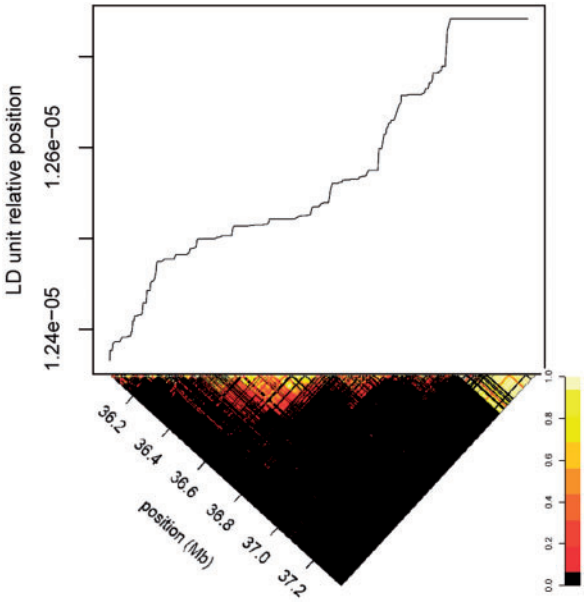

Fig. 3. The LDU map versus physical position for an illustrative region of chromosome 22. The lower triangle depicts the corresponding r^2 LD values from the CEU dataset. LD units increase little across regions/blocks with high LD, punctuated by rapid increases at the block boundaries.

Figure 3 shows the LDU map location (expressed as a fraction of the chromosome) for an illustrative 1.2 mb region of chromosome 22q, matched with the pairwise r^2 LD measure for CEU (Hudson, 1985). The LDU map is largely flat in regions of high LD, and increases rapidly across the boundary of regions that are in low LD. The conversion from LDU to physical location ensures that simulated crossovers are unlikely to occur in regions of high LD.

3 RESULTS

3.1 The HapMap samples

Phased HapMap data are available at www.hapmap.org, computed using a version of the PHASE software (Stephens and Donnelly, 2003). For the trios, the phasing has a very low error rate, in the range of 0.05–0.10%, while error among unrelateds is estimated to be ~5% (Marchini *et al.*, 2006). The 30 trios in each of CEU and YRI produce 120 phased chromosome-length haplotypes for each chromosome, while the 90 individuals in the JPT + CHB dataset produces 180 phased haplotypes. Many simulations of interest will produce far more haplotypes than the size of the pools, and a reasonable

question arises: can the finite pool support such simulations? The answer depends on the type of simulation, and a full investigation cannot be given here. However, the questions can be broadly addressed in terms of (i) representativeness of the samples, and (ii) sampling variation resulting from treating the finite sample as a population. The HapMap individuals were not randomly sampled as representative of larger populations, but the CEU samples have been presented as reflective of Europeans (Altshuler *et al.*, 2005), who may be relatively homogeneous in LD patterns (Nejentsev *et al.*, 2004). Moreover, direct comparisons of CEU data to independent Finnish (Willer *et al.*, 2006) and Spanish (Ribas *et al.*, 2006) samples reveals strong concordance of allele frequencies and LD patterns. Similarly, the CHB, JPT and YRI samples have been used as representative in continental-level investigations of population genetics (Lin *et al.*, 2006; Tenesa and Dunlop, 2006; Tenesa *et al.*, 2007; Tian *et al.*, 2006). Thus, we believe it is reasonable to use the HapMap samples until more extensive and representative data are available.

The effects of sampling variation, in contrast, can be determined from the data using well-understood statistical principles. Using a particular SNP as a candidate disease locus, the sampling variation can modestly affect inference. For 120 alleles (the smallest of our pools), standard errors in minor allele frequencies range from 0.041, for an allele with an MAF of 0.5, to 0.020, for an allele with an MAF of 0.05. We do not recommend using HAP-SAMPLE to simulate disease SNPs with smaller MAF values until the available pools are larger.

In practice, the specified disease loci are often not true candidates, but are merely intended to be representative of potential disease loci. Thus, e.g. a researcher interested in the power to detect a disease locus which has a population MAF in controls of 0.2 may choose as 'causal' a SNP with an observed MAF of 0.2, and the sampling error is immaterial. For many purposes, it is the population *distribution* of allele frequencies that matters, and this is estimated highly accurately in a sample of 120 haploid genomes. Similarly, local LD structure shows little sampling variation for these sample sizes. Figure 4 shows a 'heat map' of r^2 LD values using the Phase I CEU samples for an illustrative region of chromosome 10p. Sampling variation can be reflected by calculating r^2 values in bootstrap resamples of the 120 haplotypes spanning the region. The decay of linkage disequilibrium is extremely similar across the bootstrapped samples (95% interval boundaries shown in Fig. 4c), indicating that sampling variation for the LD patterns is minor.

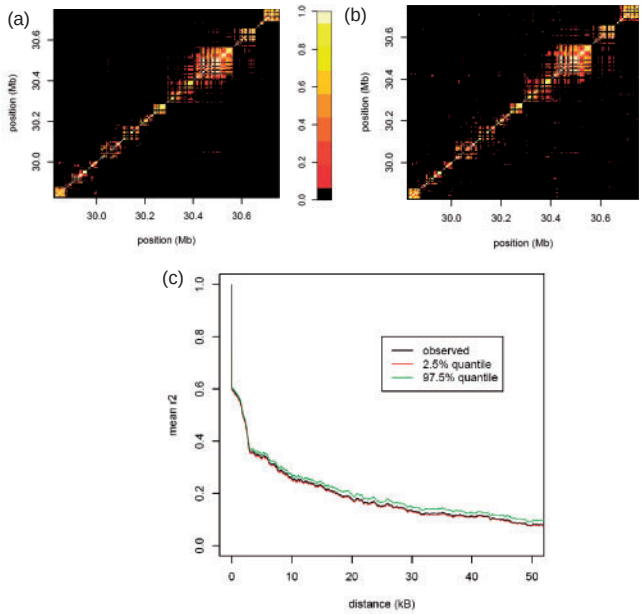

Fig. 4. r^2 linkage disequilibrium patterns in an illustrative region of chromosome 10p11, using Phase I HapMap CEU data. (**a**) Results obtained after phase estimation, with allele values estimated to be 99.95% accurate. (**b**) Results for a single bootstrap resample of the data from a, which is nearly identical. (**c**) mean r^2 as a function of distance for the region. Upper and lower quantiles for 1000 bootstrap resamples shows little sampling variation in pairwise LD patterns.

3.2 Simulation examples

HAP-SAMPLE is a general simulation tool that is immediately available to the research community. We provide three case-control examples here: a simulation of a candidate disease region with a single causal SNP, a whole genome scan for two causal SNPs and a three-population simulation of a region under recent selection. The examples serve to illustrate the tool, which can be used for a much wider variety of purposes.

For the region of 10p11 depicted in Figure 5, we supposed that SNP rs#11007734 upstream of the gene *SVIL* (representative of a promoter polymorphism) would increase in allele frequency from 0.2 in controls ($n = 300$) to 0.35 in cases ($n = 300$). We assumed Hardy–Weinberg equilibrium among the cases, although this is not a requirement of HAP-SAMPLE. After a single simulation using the CEU population and all Phase I SNPs in the region, *P*-values for Fisher's exact test were computed, and are depicted in the Figure 5a (*left panel*) along with a hypothetical genome-wide threshold of $P = 2 \times 10^{-7}$. Several interesting and realistic features emerge. The causal SNP is indeed significant, although it is not the most significant, and several nearby markers (e.g. due to low MAF) show little evidence of association. The leftmost marker shows $P < 10^{-4}$, even though it is over 200 kb distal. This sort of observation is often puzzling for investigators, raising suspicions of a second mutation. However, here the observation can be explained only in terms of LD with the sole causal SNP. We also depict in Figure 5a (*left panel*) the *P*-values for those SNPs in the region that are on the Affymetrix 100K SNP array. We use this array for illustration, arrays of higher density are available and (using the array SNP list) can be simulated via HAP-SAMPLE. Using the array, the association is essentially overlooked (the minimum array *P*-value in the region is about 0.004), illustrating that a higher-density scan may be necessary to detect the association. Using the same disease model and all the Phase I SNPs in the region, we performed 1000 simulations of 500 cases versus 500 controls (Figure 5a, *right panel*) to compute the power to detect association at the genome-wide threshold. The figure illustrates the relative power for nearby SNPs, as well as the very low

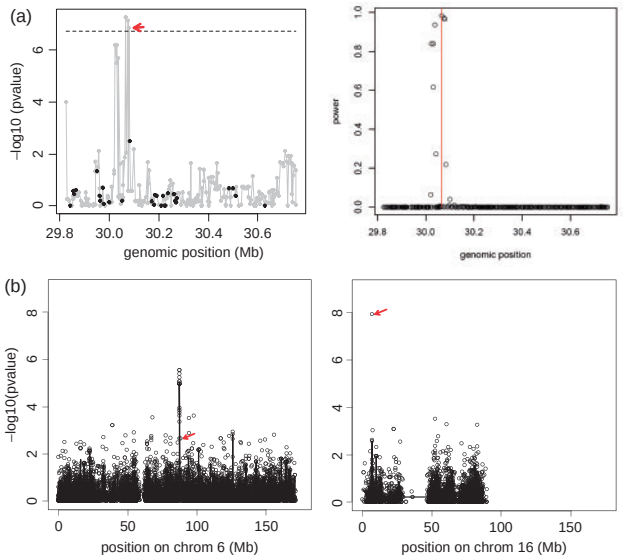

Fig. 5. (**a**) Left panel: analysis of a single CEU simulated dataset with 'causal' SNP rs#11007734 (upstream of *SVIL*, 10p11.2) using the HAP-SAMPLE approach. Each point represents the *P*-value from a Fisher's exact test of genotype counts in cases ($n=300$) versus controls ($n=300$). Controls were assumed to have a population frequency for the minor allele equal to the observed HapMap frequency of 20%, while cases were assumed to have frequency 35%. The arrow indicates the causative SNP, while black points correspond to those SNPs available on the Affymetrix 100K SNP platform. The dashed line indicates an example genome-wide significance threshold of $P=2\times10^{-7}$. Right panel: statistical power to detect an increase in allele frequency for the same disease model and SNPs as depicted in the left panel, using Fisher's exact test at each SNP for 500 cases and 500 controls. The location of the causal SNP is indicated in red. (**b**) Results from a simulated genome scan with the Affy 100K platform in CEU (300 cases, 300 controls). Chromosomes 6 (*left panel*) and 16 (*right panel*) are shown, and causal SNPs rs#2268994 and rs#10500350 (on the platform, indicated in red) with the genotype relative risk model described in text. Open circles show $-\log_{10}$(*P*-values) for individual SNPs, while the black lines show the result of median smoothing the values with SNP window width 3.

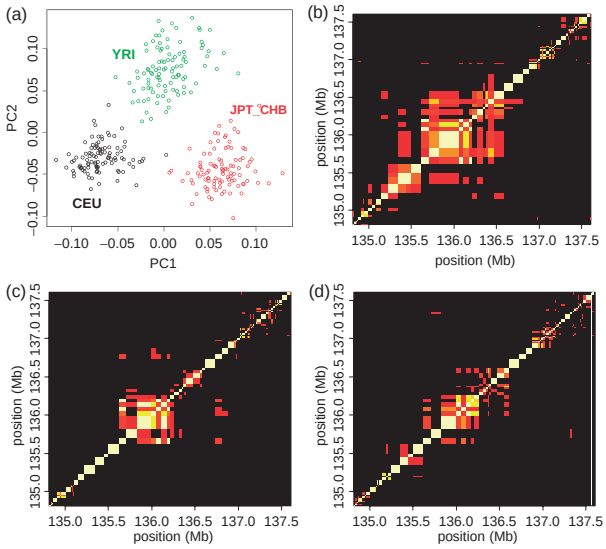

Fig. 6. One Hundred individuals from each of the CEU, JPT + CHB and YRI populations were simulated for the Affymetrix 100K array. (**a**) Using only the 97 informative SNPs in a 2.5 Mb region contain the lactase gene *LCT*, the three populations can be distinguished via principal components. (**b**) r^2 heatmap of the CEU samples shows markedly longer region of high LD than for (**c**) the combined Asian samples JPT + CHB, or (**d**) the Yoruba samples YRI.

power of numerous less-informative SNPs in the immediate vicinity of the causal SNP.

As another example, we consider a disease model involving both SNPs rs#2268994 (chromosome 6, intron 1 of *SLC35A1*, MAF = 0.49) and rs#10500350 (chromosome 16, intron 3 of *A2BP1*, MAF = 0.10). We used HAP-SAMPLE to specify genotype relative risks of 1.0, 2.2 and 3.3 for joint disease genotypes {0,0}, {0,1}, {0,2}, respectively, and GRR = 5.0 for the remaining joint genotypes. Figure 5b shows the Fisher's exact test *P*-values for the Affymetrix 100K array SNPs on the two chromosomes containing the disease SNPs, which are both on the array. SNPs at or near the causal SNPs are the most significant, although not necessarily achieving genome-wide significance. The figure also shows the result of a median smoothing of the *P*-values, which also identifies the causal SNP regions. Although the disease model involves an interaction between the causal SNPs, a logistic model (data not shown) with interaction terms for the two SNPs is not more highly significant than the individual SNP tests.

Finally, we offer a simple illustration of how the three populations exhibit different characteristics, which are readily demonstrated using our approach. We used HAP-SAMPLE to simulate 100 individuals from each of the three populations CEU, JPT + CHB and YRI, with no disease gene, using the SNPs on the Affymetrix 100K array. For 97 informative SNPs in a 2.5 Mb region contain the lactase gene *LCT*, the origin of the three samples is apparent from the first two principal components (Fig. 6a) (Price *et al.*, 2006). The region is thought to have undergone a recent selective sweep in Europeans (Bersaglieri *et al.*, 2004), producing a signature long region of high LD in the CEU samples (Fig. 6b). A similar selective sweep is thought to have occurred in East African pastoral populations (Tishkoff *et al.*, 2007), which does not include the Yoruba. Accordingly, the region of high LD is much shorter in the JPT + CHB or Yoruba samples (Fig. 6c and d). Although the lactase region is exceptional, principal component analyses using 1000 random SNPs (data not shown) easily distinguishes among the three source populations. In the current HAP-SAMPLE implementation, extreme population stratification (at the level of the three HapMap populations) can be simulated directly by sampling separately from the three pools and combining the datasets. The incorporation of more subtle stratification is the subject of future work.

4 DISCUSSION

Despite much recent work on whole genome scan designs, there is little consensus on a number of key analytic issues. To our knowledge, HAP-SAMPLE is the first tool that can simulate realistic association data reflective of *actual* whole genome

genotyping platforms. Pure model-based simulation of such data is particularly difficult, as the choice of typed SNPs has resulted from additional criteria that may not be fully described by the model, including tag-SNP selection criteria (de Bakker *et al*., 2005) and biological phenomena such as restriction enzyme sites (Matsuzaki *et al*., 2004). The general specification of disease models in HAP-SAMPLE will be useful to evaluate methods for detecting multiple risk loci (Becker *et al*., 2005; Marchini *et al*., 2005). Moreover, our approach preserves observed local LD structure for realistic finer scale examinations of candidate genes or regions.

We anticipate that HAP-SAMPLE will be particularly useful for investigations of haplotype–phenotype association methods, the power of which depends greatly on the length and specificity of associated haplotypes and risk allele frequencies (de Bakker *et al*., 2005). Similarly, *de novo* simulation approaches may not be able to realistically reflect the utility of phenomena such as Hardy–Weinberg disequilibrium to detect association in case-control (Nielsen *et al*., 1998), or case-only (Lee, 2003) studies.

As currently implemented, the random sampling in our approach largely eliminates substructure within each population, which may influence association results (Marchini *et al*., 2004). Our approach might be extended to include these effects by introducing dependencies in the pairing of pool chromosomes, although further work is necessary to ensure that the results reflect observed substructure parameter estimates (Altshuler *et al*., 2005). Similarly, extensions might include artificial output for admixture mapping (Smith and O'Brien, 2005) by sampling individuals with differing genomic proportions from the separate chromosome pools/populations (Altshuler *et al*., 2005). However, the precise manner of simulating stratification/admixture deserves careful study, and such procedures will require extensive comparisons to true admixed populations. Although HAP-SAMPLE is currently limited to assuming random mating within the pool, it may viewed as representative of situations where stratification has been appropriately controlled. Extensions to the X chromosome will be straightforward, but will require specifying the sex of the simulated individuals and more complicated disease models.

The HapMap samples provide limited opportunity for specifying rare disease variants, because of selection bias in HapMap markers (Clark *et al*., 2005) and sample size constraints. Careful analysis of the HapMap ENCODE regions (Feingold *et al*., 2004) may provide the basis for adding artificial low-frequency allelic variation to the existing data for increased novelty of haplotypes. Our approach also might be modified to simulate one or more recent disease mutations by selecting pool chromosomes to harbor the mutations. Then, forward simulation or single-locus coalescent approaches could be used to simulate an entire set of 'case' chromosomes reflecting the disease SNP ancestry.

ACKNOWLEDGEMENTS

Supported in part by the Carolina Center for Exploratory Genetic Analysis (P20 RR020751), the Carolina Environmental Bioinformatics Research Center (EPA RD-83272001), NIH grants P30ES10126, P50 GM076468, R01 GM074175 and R01 HL068890, and CF Foundation Zou05P0.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al*. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

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

Becker,T. *et al*. (2005) Haplotype interaction analysis of unlinked regions. *Genet. Epidemiol*., **29**, 313–322.

Bersaglieri,T. *et al*. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet*., **74**, 1111–1120.

Calafell,F. *et al*. (2000) Haplotype evolution and linkage disequilibrium: A simulation study. *Hum. Hered*., **51**, 85–96.

Clark,A.G. *et al*. (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*., **15**, 1496–1502.

de Bakker,P.I.W. *et al*. (2005) Efficiency and power in genetic association studies. *Nat. Genet*., **37**, 1217–1223.

De La Chapelle,A. and Wright,F.A. (1998) Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl Acad. Sci. USA*, **95**, 12416–12423.

Dudbridge,F. and Koeleman,B.P.C. (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am. J. Hum. Genet*., **75**, 424–435.

Dudek,S.M. *et al*. (2006) Data simulation software for whole-genome association and other studies in human genetics. *Proc. Pac. Symp. Biocomput*., **11**, 499–510.

Falk,C.T. and Rubinstein,P. (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet*., **51**, 227–233.

Feingold,E.A. *et al*. (2004) The ENCODE (ENCyclopedia of DNA elements) Project. *Science*, **306**, 636–640.

Gibbs,R.A. *et al*. (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Hintsanen,P. *et al*. (2006) An empirical comparison of case-control and trio-based study designs in high-throughput association mapping. *J. Med. Genet*., **43**, 617–624.

Hirschhorn,J.N. and Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet*., **6**, 95–108.

Hudson,R.R. (1985) The sampling distribution of linkage disequilibrium under an infinite Allele model without selection. *Genetics*, **109**, 611–631.

Kong,A. *et al*. (2002) A high-resolution recombination map of the human genome. *Nat. Genet*., **31**, 241–247.

Laval,G. and Excoffier,L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.

Lee,W.C. (2003) Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. *Am. J. Epidemiol*., **158**, 397–400.

Li,N. and Stephens,M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233.

Lin,E. *et al*. (2006) A case study of the utility of the HapMap database for pharmacogenomic haplotype analysis in the Taiwanese population. *Mol. Diagn. Ther*., **10**, 367–370.

Liu,Z.Q. and Lin,S.L. (2005) Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet. Epidemiol*., **29**, 353–364.

Lohmueller,K.E. *et al*. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet*., **33**, 177–182.

Lowe,C.E. *et al*. (2004) Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun*., **5**, 301–305.

Maniatis,N. *et al*. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.

Marchini,J. *et al*. (2004) The effects of human population structure on large genetic association studies. *Nat. Genet*., **36**, 512–517.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Marchini,J. *et al.* (2006) A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, **78**, 437–450.

Matsuzaki,H. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.

Montana,G. (2005) HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. *Bioinformatics*, **21**, 4309–4311.

Myers,S. *et al.* (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**, 321–324.

Nejentsev,S. *et al.* (2004) Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum. Mol. Genet.*, **13**, 1633–1639.

Nielsen,D.M. *et al.* (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am. J. Hum. Genet.*, **63**, 1531–1540.

Peng,B. and Kimmel,M. (2007) Simulations provide support for the common disease-common variant hypothesis. *Genetics*, **175**, 763–776.

Peng,B. *et al.* (2007) Forward-time simulations of human populations with complex diseases. *PLoS Genet.*, **3**, e47.

Posada,D. and Wiuf,C. (2003) Simulating haplotype blocks in the human genome. *Bioinformatics*, **19**, 289–290.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Pritchard,J.K. and Cox,N.J. (2002) The allelic architecture of human disease genes: common disease – common variant . . . or not? *Hum. Mol. Genet.*, **11**, 2417–2423.

Ribas,G. *et al.* (2006) Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum. Genet.*, **118**, 669–679.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Satagopan,J.M. *et al.* (2004) Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*, **60**, 589–597.

Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.

Schwartz,R. *et al.* (2003) Robustness of inference of haplotype block structure. *J. Comput. Biol.*, **10**, 13–19.

Smith,M.W. and O'Brien,S.J. (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.*, **6**, 623–626.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Tapper,W. *et al.* (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl Acad. Sci. USA*, **102**, 11835–11839.

Tenesa,A. and Dunlop,M.G. (2006) Validity of tagging SNPs across populations for association studies. *Eur. J. Hum. Genet.*, **14**, 357–363.

Tenesa,A. *et al.* (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, **17**, 520–526.

Thomas,D.C. *et al.* (2005) Recent developments in genomewide association scans: A workshop summary and review. *Am. J. Hum. Genet.*, **77**, 337–345.

Tian,C. *et al.* (2006) A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. *Am. J. Hum. Genet.*, **79**, 640–649.

Tishkoff,S.A. *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, **39**, 31–40.

Wang,Y. and Rannala,B. (2005) In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. *Am. J. Hum. Genet.*, **76**, 1066–1073.

Willer,C.J. *et al.* (2006) Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet. Epidemiol.*, **30**, 180–190.