BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 6 2012, pages 838–844
doi:10.1093/bioinformatics/bts033

Genetics and population analysis Advance Access publication January 17, 2012

Exact coalescent simulation of new haplotype data from existing reference haplotypes

Chul Joo Kang* and Paul Marjoram

Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: We introduce a coalescent-based method (RECOAL) for the simulation of new haplotype data from a reference population of haplotypes. A coalescent genealogy for the reference haplotype data is sampled from the appropriate posterior probability distribution, then a coalescent genealogy is simulated which extends the sampled genealogy to include new haplotype data. The new haplotype data will, therefore, contain both some of the existing polymorphic sites and new polymorphisms added based on the structure of the simulated coalescent genealogy. This allows exact coalescent simulation of new haplotype data, compared with other methods which are more approximate in nature.

Results: We demonstrate the performance of our method using a variety of data simulated under a coalescent model, before applying it to data from the 1000 Genomes project.

Availability: The source code is freely available for download at ftp://popgen.usc.edu

Contact: chulkang@usc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 17, 2011; revised on December 21, 2011; accepted on January 11, 2012

1 INTRODUCTION

We live in an era in which the genome-wide association study (GWAS) is one of the standard tools by which we interrogate the genome for polymorphisms that relate to phenotype. There is also a growing use of next-generation sequence (NGS) data. A large number of new methodological approaches are being developed and published in both areas. There is much discussion about the degree to which the GWAS approach has been successful, for example, see Eichler *et al.* (2010); Heard *et al.* (2010); Vineis and Pearce (2010), but in humans many new associations between genotype and disease status have been discovered.

An important part of the process of methods development in the GWAS era is the testing of those methods on both real and simulated test data. Frequently investigators have a single, smaller set of data from a given population, the features of which they wish to mimic when producing multiple, larger datasets for testing purposes. A common example is the HapMap data (The International HapMap 3 Consortium, 2010) or 1000 Genome Project (1000GP) data (1000 Genomes Project Consortium, 2010). In this article, we present an exact coalescent method for simulating test sets of new haplotype (or genotype) data, based upon such an existing, smaller, initial sample.

The goal is to sample new datasets, D', consisting of n (say) haplotypes, conditional upon some observed, existing data, D, which contains $n_o << n$ haplotypes. Under the assumption that the popular coalescent model is a reasonable approximation to the evolution of the data, the goal here is to produce new data, D', conditional on the unobserved genealogy, G, of the existing data, D (since it is through G that the dependence of D' on D is expressed). Note that in the presence of recombination G will be a graph, the so-called ancestral recombination graph (ARG), rather than a tree.

The problem of sampling G conditional on D is computationally challenging, and for that reason several approximate methods have been proposed, the most relevant here being HAPGEN (Spencer *et al.*, 2009) and HAP-SAMPLE (Wright *et al.*, 2007), which are based-upon the PAC-likelihood method of (Li and Stephens, 2003). The algorithm of Li and Stephens is an appealing way to reduce the computational burden involved in calculating probabilities under a full coalescent model. Instead, it uses an approximation scheme, based upon a model in which haplotypes are explained as mosaics of other existing haplotypes. This scheme is shown to (i) reduce computational burden significantly, and (ii) provide a good approximation to probabilities that would be obtained under the full coalescent model. However, it remains the case that one would prefer to sample from the exact coalescent model when such a thing is tractable. In this article, we present software that makes this practical for regions of length up to 500 kb, using as many as about $n_o = 100$ reference haplotypes.

2 METHODS

Let D denote an existing sample of haplotype data, assumed to evolve under an unstructured, neutral coalescent model. The evolutionary parameters for this model, assuming the underlying population has effective size N_e, are the per site mutation rate ($\Theta = 4N_e\mu$), per site recombination rate ($\rho = 4N_e r$) and exponential population growth rate g, where $\mu(r)$ is the mutation(recombination) rate per generation per site. We refer to these parameters collectively as Ψ. We simulate new haplotype data D' from the conditional distribution $P(D' \mid D, \Psi)$ using the coalescent model. This process can be written as

$$P(D' \mid D, \Psi) = \int_{G'} \int_{G} P(D' \mid G', D) P(G' \mid G, \Psi) P(G \mid D, \Psi)$$

where G is the (unobserved) ARG underlying D, and G' is the new ARG that underpins the newly generated haplotypes (so $G \subset G'$).

*To whom correspondence should be addressed.

Our method, RECOAL (REference haplotype simulation using a COALescent approach), proceeds in several steps, each of which we detail below. First we sample an unobserved G, underlying D. Then, using a coalescent prior, we generate a new G' that contains G but adds n new tips, each of which will correspond to a new haplotype. The n_o original haplotypes are then discarded. The general scheme of RECOAL is represented in Figure 1. Specifically, we proceed as follows.

2.1 Sampling of coalescent genealogy for reference haplotypes

An ancestral recombination graph G for n_0 reference haplotypes is sampled from the conditional distribution $P(G \mid D, \Psi)$. We use the Metropolis-Hastings Markov chain Monte Carlo method of C.J.Kang and J.Felsenstein (submitted for publication) to sample graphs from this distribution.

2.2 Simulation of coalescent genealogy for new haplotypes

After sampling G, we propose a larger ancestral recombination graph G', containing G and also including n new tips that will lead to the new data D'. We let G^* denote the newly simulated ancestry, so that $G' = G \cup G^*$. G^* is sampled using the usual coalescent prior, conditional on G and Ψ.

2.3 Simulation of data for new haplotypes

For each site, i, types for the new haplotype data at those sites, denoted by D'_i, are simulated conditional on the types at the same sites in the reference haplotypes, denoted by D_i, as well as on G' and the mutation rate. We let G'_i denote the genealogy (a tree in this case) defined by G' at site i. We do this using the following procedure.

First, the type of the node I_i^a, corresponding to the root of G'_i, is sampled from $P(I_i^a \mid G'_i, D_i)$. We then work our way from top-to-bottom (i.e. root-to-leaf) through G'_i, generating the type of each internal node. Let b denote such an internal node and denote its type by I_i^b. We sample I_i^b from $P(I_i^b \mid d_i^b, I_i^c, G'_i, \Psi)$, where I_i^c is the (already sampled) type of the parental (higher) node, c, directly connected to b, and d_i^b is the types of the subset of haplotypes that are descendants of node b. After sampling the types of the all internal nodes, the types for the new haplotype data are sampled. The type of a new haplotype d_i^x is sampled from $P(d_i^x \mid I_i^c, G'_i, \Psi)$, where x is the tree tip representing this new haplotype and c is the internal node directly connected to x (and which has type I_i^c). Note that RECOAL uses the finite sites model for mutations—allowing for multiple mutation events at each site. In particular, we used the Hasegawa-Kishino-Yano model (HKY85) (Hasegawa *et al.*, 1985) as the DNA substitution model.

2.4 Ascertained SNPs

Many SNP datasets are collected using an ascertainment process in which some kind of bias is likely to be present. An example is the use of a so-called SNP-chip, in which SNPs are often selected to be included on the platform only if they are common. Another example is data from NGS technologies, in which, under a variety of calling schemes, the more copies of a mutant allele that are present, the easier the polymorphism is to call (See Supplementary Material for a full discussion.) This leads to an ascertainment bias in that rarer SNPs are preferentially missing from the reference haplotypes, and the number of polymorphisms is thereby decreased.

Applying our algorithm without respecting this ascertainment bias would lead to incorrect performance. For example, we would tend to sample graphs that were shorter, on average, than is correct (since the method would be ignorant of the ascertainment process). Ascertainment is a difficult problem to deal with, but we present an ascertainment-corrected version of our algorithm that allows for ascertainment schemes in which the probability of detecting an SNP is dependent upon its frequency. We give theoretical details of the ascertainment correction in the Supplementary Material.

Fig. 1. The simulated coalescent genealogy G' for one site. The solid line represents the sampled coalescent genealogy (G) for the reference haplotypes. (For simplicity sake, we illustrate this as a tree, but it will in general be a graph.) The dotted line represents the newly simulated component G^*, which shows the ancestry of the new haplotypes (indicated in italics). G' is the union of G and G^*.

3 RESULTS

We demonstrate performance of the proposed method using both data simulated under the coalescent model and data drawn from the 1000 Genomes Project. We begin with examples using simulated data.

3.1 Coalescent data

We used Hudson's coalescent-based simulation program (called ms) (Hudson, 2002) to simulate 1040 haplotypes for a given parameter combination Ψ. We then randomly selected 40 haplotypes from those 1040 haplotypes. Using the selected 40 samples as the reference haplotypes, referred to as H_{ref}, we applied our method to simulate 1000 new haplotypes using the same Ψ. We refer to these as H_{new}. We then compare properties of H_{new} to the 1000 unsampled haplotypes in the initial coalescent data (referred to as H_{coal}).

3.1.1 SNP frequency and allele frequency spectrum Figure 2 plots the the MAF of each polymorphic locus in H_{coal} against the frequency of the same locus in H_{ref}. We then make the same comparison between H_{new} and H_{ref}. The correlation between MAFs is $r^2 = 0.905$ in the former case and $r^2 = 0.903$ in the latter, indicating very similar behavior in the actual and generated 1000 haplotypes. We show aggregate results from 100 replicate analyses.

In Table 1, we present the results of applying our algorithm to data generated under a range of different mutation rates, and no recombination, across 100 replicates. We show the average number of SNPs over a 10 kb region in H_{ref}, H_{coal} and H_{new}. We also show the results of applying Watterson's estimate Θ of the mutation rate to each dataset (Watterson, 1975). This shows that the ability of RECOAL to both reproduce SNPs in the reference haplotypes, as well as add new SNPs, leads to a good agreement between the overall polymorphism level in H_{new} and H_{coal}. Tables 2 presents the results of similar analyses, but with recombination included, for

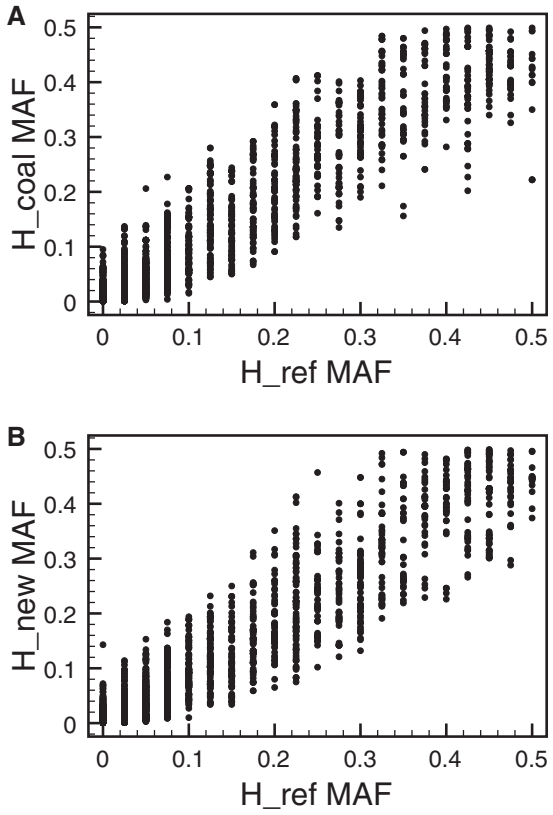

Fig. 2. Comparison of minor allele frequencies (MAFs) between H_{coal} and H_{ref} (**A**) and between H_{new} and H_{ref} (**B**).

Table 1. The average number of SNPs (with 95% confidence interval), and Watterson's estimate of mutation rate ($\hat{\Theta}$), for H_{coal}, H_{ref} and H_{new}

Θ		10^{-4}	10^{-3}	10^{-2}
H_{coal}	SNPs	7.51 (6.97–8.06)	73.34 (70.69–76.07)	737.6 (706.5–770.4)
	$\hat{\Theta}$	1.003×10^{-4}	0.9799×10^{-3}	0.9856×10^{-2}
H_{ref}	SNPs	4.25 (3.80–4.71)	39.87 (37.55–42.26)	419.4 (390.1–452.2)
	$\hat{\Theta}$	0.9992×10^{-4}	0.9373×10^{-3}	0.9861×10^{-2}
H_{new}	SNPs	7.78 (7.18–8.38)	74.40 (72.01–76.88)	733.4 (703.7–764.2)
	$\hat{\Theta}$	1.039×10^{-4}	0.9941×10^{-3}	0.9799×10^{-2}

Table 2. The average number of SNPs (with 95% confidence interval) and Watterson's estimate of mutation rate ($\hat{\Theta}$), for H_{coal}, H_{ref} and H_{new}

Region size		10 kb	50 kb	100 kb
H_{coal}	SNPs	74.13 (71.26–77.17)	374.7 (357.8–392.7)	783.6 (739.0–832.1)
	$\hat{\Theta}$	0.9177×10^{-3}	$1.001.\times10^{-3}$	1.047×10^{-3}
H_{ref}	SNPs	39.87 (37.55–42.26)	214.8 (199.4–231.8)	456.2 (412.6–503.2)
	$\hat{\Theta}$	0.9373×10^{-3}	1.010×10^{-3}	1.073×10^{-3}
H_{new}	SNPs	74.40 (72.01–76.88)	385.8 (368.1–405.1)	787.8 (71.14–76.72)
	$\hat{\Theta}$	0.9941×10^{-3}	1.031×10^{-3}	1.053×10^{-3}

Here $\Theta=0.001$ and $\rho=0.0005$.

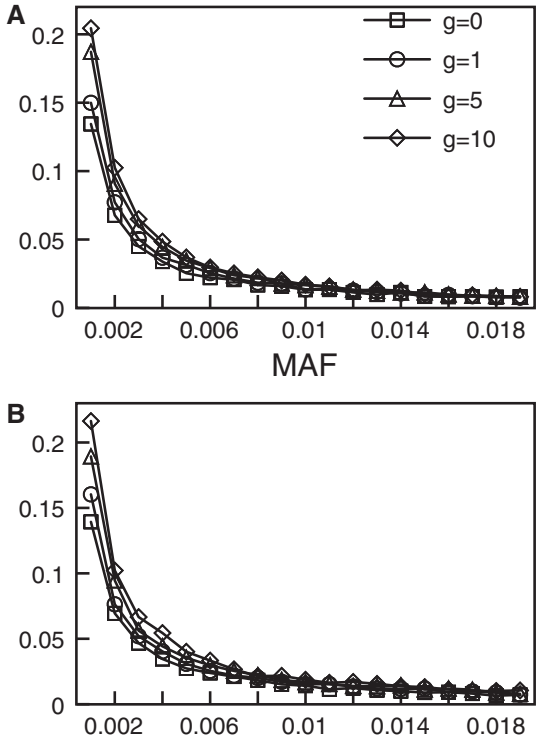

Fig. 3. The MAF spectrum for datasets simulated by ms (**A**) and RECOAL (**B**) with different growth rates.

a number of different region lengths. In all cases, we see an excellent agreement between the degree of polymorphism in H_{new} and H_{coal}.

In Figure 3, we compare the MAF distribution at polymorphic loci in H_{coal} and H_{new} across 100 datasets. We show this for a variety of population growth rates. We see a good agreement between the MAF spectrum patterns simulated by ms and by RECOAL.

3.1.2 Pattern of linkage disequilibrium As well as preserving the distribution of the number of polymorphic loci and their MAFs, it is also important to preserve patterns of linkage disequilibrium (LD). In Figure 4, we show the pairwise r^2 values between adjacent markers in H_{ref}, H_{coal} and H_{new} across 100 replicate analyses in which data were simulated over a 10 kb region with $\Theta=0.001$, $\rho=0.0005$. In Figure 4A, the comparison is between each adjacent pair of markers in H_{ref} and the same pair of markers in H_{coal}, and in Figure 4B the comparison is between pairs in H_{ref} and H_{new}. Again, we see a good agreement between the figures. RECOAL reproduces the LD values from the reference haplotypes with similar variation

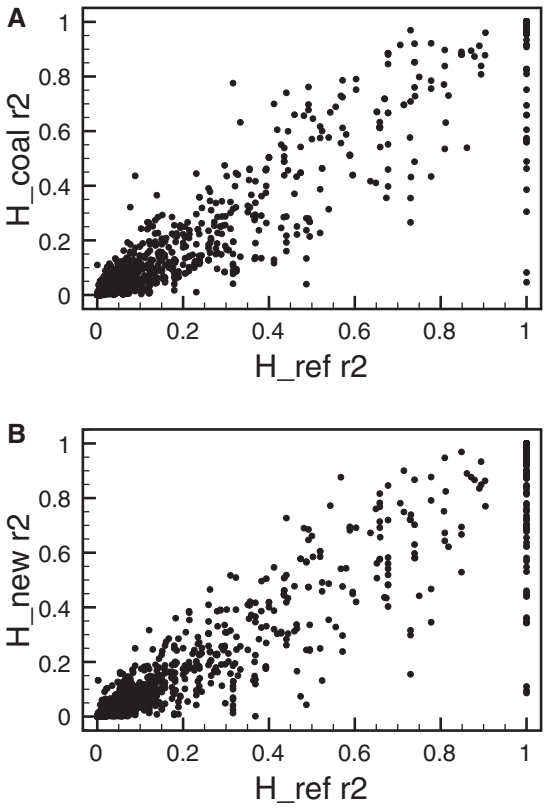

Fig. 4. Changes of pair-wise r^2 values between pairs of adjacent loci in H_{ref} and H_{coal} (**A**) and in H_{ref} and H_{new} (**B**).

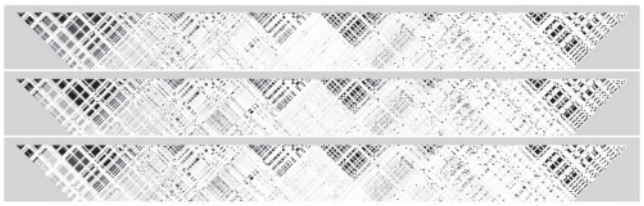

Fig. 5. LD patterns in a single, representative example dataset for H_{coal} (top), H_{ref} (middle) and H_{new} (bottom).

(an r^2 of 0.93 compared with $r^2 = 0.95$ for the actual coalescent data), suggesting that data produced by RECOAL are good proxies for the unobserved data in this context.

Next, to illustrate the behavior in a single dataset, Figure 5 shows an example of LD patterns, plotted by Haploview (Barrett *et al.*, 2005), for data simulated with ms using an overall recombination rate of $\rho = 0.0005$ but with the presence of a single hotspot in the middle at which the rate of recombination is 100 × greater than in the rest of the region. The LD patterns are well preserved.

3.1.3 Application to ascertained SNPs We now give an example of applying RECOAL to ascertained data. Here, we simulate a situation in which the probability of an SNP being reported as polymorphic is a function of the frequency of the mutant allele in the sample at that position, such as might be appropriate for data from NGS. We denote the number of SNPs present in H_{new} that were also observed in H_{ref} by S_O. S_H denotes new SNPs present in H_{new} that will also be present in H_{ref}—these are the SNPs added to reflect the ascertainment scheme (Supplementary Material). Finally, S_N denotes SNPs which are present in H_{new} as the result of new mutation in G^* and which are non-polymorphic in H_{ref}. Table 3 shows the number of SNPs in the newly simulated haplotypes from the reference haplotypes ascertained with the different SNP calling threshold λ, in each of these categories, both with and without the ascertainment correction scheme. Without application of the ascertainment correction, the new haplotypes have fewer SNPs than they should. (The results for $\lambda = 0$ give the correct null behavior, since there the ascertainment has no effect.) When ascertainment correction is used, the new haplotypes have a number of SNPs that is close to the correct value, although we see a slight tendency to introduce too many new SNPs as λ increases.

Table 3. The number of observed SNPs in H_{ref} and H_{new} without ascertainment correction (no asc.) and with ascertainment correction (asc)

λ		0	0.1	0.2
H_{ref}		41.16 (37.42–45.28)	23.20 (19.68–27.24)	16.96 (13.44–21.00)
no asc	S_O	40.70 (36.75–44.65)	23.20 (19.68–27.24)	16.96 (13.44–21.00)
	S_H	0 (0–0)	0 (0–0)	0 (0–0)
	S_N	35.42 (33.51–37.33)	40.66 (37.86–43.74)	45.08 (41.84–48.32)
	H_{new}	76.12 (72.20–80.26)	63.86 (59.54–68.68)	61.04 (56.87–66.42)
asc	S_O	40.70 (36.75–44.65)	23.20 (19.68–27.24)	16.96 (13.44–21.00)
	S_H	0 (0–0)	17.92 (15.84–20.00)	24.76 (22.22–27.30)
	S_O+S_H	40.70 (36.75–44.65)	41.12 (37.01–45.31)	41.72 (37.48–46.32)
	S_N	35.42 (33.51–37.33)	36.60 (34.44–38.76)	37.16 (35.38–39.02)
	H_{new}	76.12 (72.20–80.26)	77.72 (73.41–82.61)	78.88 (73.98–84.20)

The 95% confidence intervals are also shown. We simulated 100 replicates of 1000 new haplotypes over a 10 kb region with $\Theta = 0.001$ and $\rho = 0.0005$ using 40 reference haplotypes.

3.2 Comparison with other methods

Next, we compare RECOAL to another haplotype simulation method that creates new haplotype data by resampling existing reference haplotypes. There are two such methods: HAPGEN (Spencer *et al.*, 2009) and HAP-SAMPLE, (Wright *et al.*, 2007). Both these methods simulate new haplotypes using schemes based upon Li and Stephen's elegant PAC likelihood algorithm (Li and Stephens, 2003). Both methods also perform well in terms of reproducing the pattern of LD in the region considered (results not shown). However, the models do differ in how they treat the mutation process. There is no mechanism for introducing new polymorphic sites in HAP-SAMPLE. As such, HAP-SAMPLE will necessarily under-represent the degree of polymorphism in the newly generated

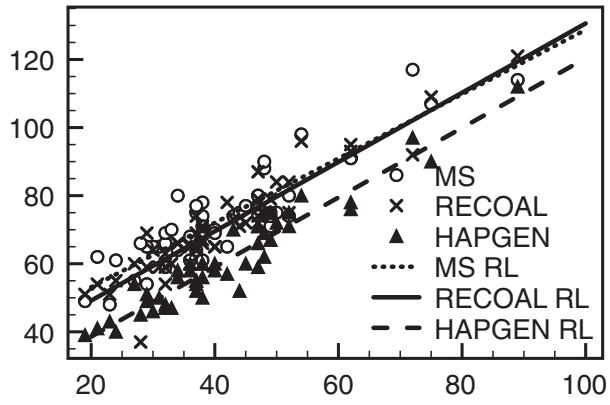

Fig. 6. The relationship between the number of SNPs in the reference haplotypes (x-axis) and the number of SNPs in simulated haplotypes (y-axis). We simulated 1000 haplotypes from 40 reference haplotypes with RECOAL (cross) and HAPGEN (filled triangle) using $\Theta = 0.001$. For comparison, we plot the number of SNPs in the reference and original populations (y-axis) in the initial data [simulated by ms (circle)]. We also drew linear regression lines (RL) for three cases.

haplotypes. For this reason, we focus our comparison on HAPGEN, which introduces new mutations using a mutation rate that decreases as each new haplotype is produced (reflecting the intuition that as more haplotypes are sampled, the probability of detecting new polymorphism will decrease).

As before, we examine the number of SNPs that are observed on the haplotypes generated by each algorithm. We show results for a range of sizes for H_{ref}. Figure 6 shows the relationship between the number of SNPs in H_{ref} and the number of SNPs in H_{new}, using both methods. For reference, we also include a comparison between H_{ref} and H_{coal} in the initial data, which were simulated using ms, with $\Theta = 0.001$. We show results across 50 replicates. As a summary, we fit a regression line in each case. We see a good agreement between H_{coal} and H_{new} when using RECOAL, and a good agreement between the regression line for results from RECOAL and those from ms. The agreement is also quite good when using HAPGEN, but there appears to be a systematic tendency to add slightly too few SNPs when using HAPGEN.

Tables 4 and 5 show the number of SNPs simulated by HAPGEN and RECOAL for a variety of sizes of H_{ref}. While the 95% confidence interval for RECOAL includes the mean number of SNPs in the simulated population data in each case, and appears to show no systematic dependency upon the size of H_{ref}, we observe that the average number of SNPs in haplotypes simulated by HAPGEN appears to be dependent on the size of both H_{ref} and H_{new}. In defense of the latter algorithm, we recall that it is an approximation designed to improve computational tractability, thereby allowing application of such a resampling scheme in a range of contexts that is broader than is permissible using our own method. Our point in this article is to present an exact coalescent method that can be used when the region of interest is reasonably small (less than ~500 kb).

3.3 Simulation with 1000 Genome Project data

We close by showing an example application of RECOAL aimed at simulating new haplotypes from the low-coverage 1000GP data. We simulated 1000 haplotypes for a 500 kb region (21–21.5 MB) of Chromosome 21 using haplotypes constructed from the 4× coverage 1000GP CEU data. Since the data are low coverage, some SNPs are not called. The rate at which this occurs (i.e. the ascertainment bias), is reported as a function of allele frequency (in the sample) in (1000 Genomes Project Consortium, 2010) (Fig. 2). When applying RECOAL, we used fine-scale recombination rates estimated by the HapMap consortium (The International HapMap 3 Consortium, 2010). Θ for that region is ~0.001 per site (Kang and Marjoram, 2011).

We summarize the simulation results in Table 6. Again, we report SNPs in three classes: S_O, SNPs which are also observed in the reference haplotypes; S_N, SNPs which appear only on the simulated haplotypes; and S_H, SNPs present in H_{new} that will also be present, but were not observed due to ascertainment, in H_{ref}. We see that the ascertainment correction adds 98.4 SNPs on average to correct for the estimated ascertainment probabilities.

Table 4. The average number of SNPs (with 95% confidence interval) when simulating 1000 haplotypes using RECOAL (RC) and HAPGEN (HG) from reference haplotype sets of differing sizes

Ref. size	10	40	100
H_{ref}	29.75 (26.00–34.08)	41.83 (37.30–46.70)	51.55 (47.20–56.40)
RC	72.98 (68.72–77.53)	73.05 (67.20–78.90)	69.88 (65.10–75.15)
HG	56.15 (52.18–60.58)	61.68 (57.02–66.58)	78.53 (74.00–83.67)

For sake of comparison, an average of 72.07 SNPs (69.54–74.88) were simulated by ms when generating 1000 haplotypes with the same parameter values.

Table 5. The average number of SNPs (with 95% confidence interval) when simulating datasets of new haplotypes of differing sizes using RECOAL (RC) and HAPGEN (HG)

Sample size	100	500	1000	5000
MS	50.45 (46.75–54.30)	66.55 (62.25–70.92)	72.07 (69.54–74.88)	93.3 (88.60–98.05)
RC	53.75 (48.83–58.70)	64.45 (60.23–68.92)	72.43 (66.54–73.31)	92.23 (87.67–97.22)
HG	51.43 (46.58–57.90)	58.85 (53.75–62.12)	61.68 (57.02–66.58)	72.65 (67.60–78.88)

Forty reference haplotypes were used. For comparison, we also show results obtained when using ms to simulate a population of the same size in each case (MS).

Table 6. The number of SNPs and mean MAFs of those SNPs in data simulated using reference haplotypes drawn from 1000GP

		Total	S_O	S_H	S_N
1000 GP	SNPs	3536.5	1669.78	98.36	1767.98
	MAF	0.055	0.106	0.033	0.009

Fig. 7. The LD patterns of the 1000 Genome Project reference haplotypes (top) and of new haplotypes simulated by RECOAL (bottom).

We also show an example of the LD pattern observed in H_{new} and H_{ref} in Figure 7. The pattern of LD of the simulated haplotypes shows high agreement.

4 DISCUSSION

In this article, we introduce an exact coalescent method for the simulation of new haplotype data from an existing set of reference haplotypes. Our method, RECOAL, simulates data from the distribution $P(D' \mid D, \Psi)$ using the full coalescent model rather than using an approximation to that model. In addition to the replication of SNPs present in the reference haplotypes, RECOAL also simulates new SNPs using a coalescent prior. It preserves LD structure well, and we demonstrate that new polymorphism is added at an appropriate level.

Because of its exact nature, RECOAL is less computationally tractable than the approximation underpinning HAPGEN, but it is produces data under the full coalescent model rather than the approximation to the coalescent model exploited by HAPGEN. As such, we argue that it is appropriate to use RECOAL when computational considerations permit. Nonetheless, we believe HAPGEN remains an excellent alternative approach.

It should be noted that Step 1 of our approach is based on an MCMC method. As such the user must allow the MCMC chain to converge to the stationary distribution of $P(G \mid D, \Psi)$. The time taken to reach convergence increases with both overall recombination rate and number of reference haplotypes. Thus, the combination of recombination rate and the number of reference haplotypes determine the limits of computational tractability of RECOAL. As an example, the simulation used in this article, which includes 40 reference haplotypes over a 100 kb region with human population parameters, takes ~4 h to simulate 1000 new haplotypes on a typical desktop machine. As such, RECOAL is intractable for simulation of regions larger than ~500 kb. However, in the context of follow-up of a GWAS, for example, in which the area being investigated may be quite short, our software allows for the exact coalescent simulation of new haplotypes conditional on the existing data, under the coalescent model, rather than using a more approximate method. When larger regions are desired, or much larger reference samples exist, we recommend the approximate method HAPGEN still be used.

Of course, our method is model-based and, as such, is exact in the sense that it samples from exactly the correct distributions *conditional* on that model. This, then, is a good time to remind the reader of a quote attributed to George Box: 'All models are wrong; some are useful'. The coalescent has proven itself to be widely useful, but it is still, of course, 'wrong'. One example of its 'wrong-ness' in the present context is that we use a neutral model of molecular evolution, ignoring natural selection. There is a literature about the coalescent with selection included (Kaplan *et al.*, 1988; Neuhauser and Krone, 1997), and some programs (Spencer and Coop, 2004; Teshima and Innan, 2009) can simulate data with the presence of the selective mutation. But those models are not tractable when one wishes to condition upon the existing data, so, like most other coalescent simulators, RECOAL simulates data under neutrality. Another area in which our model is wrong is that it is based upon an unstructured coalescent. In other words, we assume random mating. While this has proven to be a robust approximation to the behavior of single populations, it has proven necessary to introduce extensions to the coalescent when population structure is present. As such we would recommend caution when applying our method to datasets in which multiple populations are present, or selection is expected to have occurred. In the former context, for example, it might be more appropriate to produce new haplotypes for each subpopulation separately, by dividing the data into its subpopulations and applying RECOAL to each subpopulation in turn.

The existing version of RECOAL allows for fine-scale variation of recombination rates and for exponential growth of population size. Extensions to other more complex demographic scenarios such as bottlenecks and population admixture are also possible, and are currently under development.

Finally, we also presented an extension of our method to ascertained data in which the probability of ascertainment is a function of allele frequency. Admixture is an extremely challenging problem in coalescent methods and we note in passing that our correction illustrates how sampling of an underlying coalescent genealogy, condition upon observed polymorphism data, can be adapted to respect such an ascertainment scheme.

Funding: National Institutes of Health (MH084678) and National Science Foundation (HG005927), as well as improvements due to comments received from the reviewers.

Conflict of Interest: none declared.

REFERENCES

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Barrett,J.C. *et al.* (2005) Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, **21**, 263–265.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol.*, **22**, 160–174.

Heard,E. *et al.* (2010) Ten years of genetics and genomics: what have we achieved and where are we heading? *Nat. Rev. Genet.*, **11**, 723–733.

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

Kang,C.J. and Marjoram,P. (2011) Inference of population mutation rate and detection of segregating sites from next-generation sequence data. *Genetics*, **189**, 595–605.

Kaplan,N.L. *et al.* (1988) The coalescent process in models with selection. *Genetics*, **120**, 819–829.

Li,N. and Stephens,M. (2003) Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics*, **165**, 2213–2233.

Neuhauser,C. and Krone,S.M. (1997) The genealogy of samples in models with selection. *Genetics*, **145**, 519–534.

Spencer,C.C.A. and Coop,G. (2004) Selsim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, **20**, 3673–3675.

Spencer,C.C.A. *et al.* (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.*, **5**, e1000477.

Teshima,K. and Innan,H. (2009) mbs: modifying hudson's ms software to generate samples of dna sequences with a biallelic site under selection. *BMC Bioinformatics*, **10**, 166.

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Vineis,P. and Pearce,N. (2010) Missing heritability in genome-wide association study research. *Nat. Rev. Genet.*, **11**, 589.

Watterson,G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popn. Biol.*, **7**, 256–276.

Wright,F.A. *et al.* (2007) Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, **23**, 2581–2588.