

## Letter to the Editor

## Haplotyping Using SIMPLE: Caution on Ignoring Interference

Shili Lin,<sup>1\*</sup> Zachary Skrivanek,<sup>2</sup> and Mark Irwin<sup>1</sup><sup>1</sup>Department of Statistics, Ohio State University, Columbus, Ohio<sup>2</sup>Eli Lilly, Indianapolis, Indiana*To the Editor:*

Accurately inferred haplotypes can provide valuable information in genetic studies. In a recent article in *Genetic Epidemiology*, Skrivanek et al. [2003] describe a software package, SIMPLE, for multipoint linkage analysis. Although the focus of the report was on nonparametric linkage analysis using allele sharing statistics, haplotype analysis is also within the capability of SIMPLE. One particular feature of SIMPLE is its ability of accounting for interference. As we demonstrate below, through the analysis of a published Episodic Ataxia (EA) dataset [Litt et al., 1994], this is an important feature as ignoring interference may lead to obligatory multiple (double or more) recombinant haplotypes.

Likelihood-based and rule-based methods are two types of approaches frequently employed to infer haplotypes of pedigree members based on their (partially) observed phase-unknown genotype data. Rule-based methods, such as the algorithm proposed by Qian and Beckmann [2002] for finding Minimum-Recombinant Haplotype Configurations (MRHCs), are fast. However, likelihood-based methods have their advantages that cannot be matched by rule-based methods. By examining the entire distribution ( $P(h|g)$ ) of haplotype configurations ( $h$ ) conditional on the observed genotypes ( $g$ ), one can offer multiple haplotypes with probabilities larger than a certain investigator-set threshold for further consideration. This set of probable haplotypes may include only configurations each having a total number of

recombinations exceeding that of any of the MRHCs, if all the MRHCs involve multiple recombinant haplotypes with smaller probabilities.

Lin and Speed's work [1997] is an example of a likelihood-based approach. Their algorithm uses the Markov chain Monte Carlo (MCMC) methodology to obtain (dependent) samples from the conditional distribution  $P(h|g)$ . This algorithm is, in principle, applicable to pedigrees of arbitrary sizes and complexities. "An exploratory and experimental approach was taken" by Lin and Speed [1997], with runs consisting of only 100 iterations, for the analysis of the EA dataset. Even with such short runs, the two configurations estimated to have the highest probabilities are indeed among the two most, and the two second most, probable configurations of the EA dataset, respectively, as confirmed by Qian and Beckmann [2002]. The EA dataset was also analyzed using a different MCMC method, with simulated annealing, by Sobel et al. [1996], which produced a single haplotype configuration that is confirmed, also by Qian and Beckmann [2002], to be among the two second most probable configuration, which is about 1/4 as probable as the most probable one.

In addition to the above-mentioned flexibility of likelihood-based approaches, realistic interference models, such as the chi-square model that has been demonstrated to fit human data adequately [Broman and Weber, 2000; Lin et al., 2001], can be incorporated into this type of approach. The set of chi-square models is indexed by an interference parameter,  $m$ , that specifies the strength of

\*Correspondence to: Shili Lin, PhD, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247.

E-mail: shili@stat.ohio-state.edu.

Contract grant sponsor: NSF; Contract grant number: DMS-9971770.

Received 2 July 2003; Accepted 7 July 2003

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.10275

interference. It has been found that  $m=4$  models interference in the human genome well, through using either inferred meioses [Broman and Weber, 2000] or pedigree data [Lin et al., 2001]. Also note that  $m=0$  corresponds to Haldane's no interference model. Although interference models can be incorporated into MCMC methods, neither Sobel et al. [1996] nor Lin and Speed [1997] took that into consideration; their analyses used Haldane's no interference model.

The EA dataset consists of a pedigree of 29 individuals, with 27 of them genotyped at 9 markers spanning an 11-cM segment on chromosome 12. Due to its simple pedigree structure and the fact that there is little missing data, SIMPLE, also a likelihood-based approach, can be much more efficient. SIMPLE, a Monte Carlo importance sampling method using sequential imputation, produces (independent) samples from a distribution,  $P'(h|g)$ , that differs from, but relates to, the target distribution,  $P(h|g)$ , as follows:

$$P'(h|g) = P(h|g)P(g)/w(h),$$

where  $w(h)$  is a weight function. The conditional probability of a given haplotype,  $h^*$ , can be estimated by

$$\hat{P}(h^*|g) = \sum_{j=1}^N I(h^{(j)} = h^*)w'(h^{(j)}),$$

where  $j$  indexes the realizations from SIMPLE, and  $w'$  is the normalized importance sampling weight:  $w'(h^{(j)}) = w(h^{(j)}) / \sum_{k=1}^N w(h^{(k)})$ . Interference, based on the chi-square model, is accounted for in SIMPLE.

Since Qian and Beckmann's MRH algorithm does not account for interference either, their results, as well as those from Lin and Speed [1997] and Sobel et al. [1996], are comparable with ours for  $m=0$ . The results with 100,000 iterations from SIMPLE yielded only four configurations, A, B, C, D (A-D), having estimated probabilities greater than 1%, matching the four MRHCs found by Qian and Beckmann. These four configurations are defined by the haplotypes of four individuals, 1006, 1007, 113, and 114 (Table 1). The haplotypes of the remaining individuals in the pedigree are common among A-D (see Qian and Beckmann's fig. 2 for their haplotype assignments). The relative probabilities of A-D, obtained from the estimated conditional probabilities, are 0.10, 0.10, 0.41, and 0.39

(Table II), matching those from Qian and Beckmann closely.

To draw the 100,000 realizations from SIMPLE, it took 4 min on a linux machine with an AMD Athlon 1800+MP processor running at 1.533 GHz and 3GB of RAM. Although it would generally take longer to get results from SIMPLE than from running the MRH algorithm, 4 min vs., say, 1 min, of computing time would hardly be an important factor of consideration, for most investigators, when choosing a better program for haplotype analysis. Also from Table II, we can see that, even with just 100 iterations, the estimated relative probabilities for A-D are 0.13, 0.09, 0.39, and 0.39, respectively, still quite close to the exact values. These results show that, as expected, SIMPLE is a better alternative than Lin and Speed's or Sobel et al.'s MCMC procedures for this application, due to the simplistic nature of the EA dataset.

When interference, under the chi-square model with  $m=4$ , is accounted for, the results change dramatically. With 100,000 iterations, there are 8 configurations,  $A', B', C', D'$  ( $A'-D'$ ),  $A'', B'', C'', D''$  ( $A''-D''$ ), having probabilities greater than 1%. The relative probabilities within each of the two sets of configurations,  $A'-D'$ , and  $A''-D''$ , are still quite close to their exact values. The corresponding configurations in the sets A-D,  $A'-D'$  and  $A''-D''$  (e.g., A,  $A'$  and  $A''$ ) differ from one another in the haplotype of individual 1001 only (Table I). Configurations  $A'-D'$  and  $A''-D''$  all have 6 recombinations, but involving only single recombinants. In contrast, each of A-D has 5 recombinations, but all involving a double recombinant in individual 100.

These results are not surprising. Under the no-interference model, recombination events are independent of one another, thus a haplotype with only 5 recombinations can be much more probable than a corresponding one with 6 recombinations. In fact, configurations in the set A-D are approximately 33, and 100, times more probable than their corresponding ones in  $A'-D'$ , and  $A''-D''$ , respectively. However, under a realistic interference model, such as the chi-square model with  $m=4$ , double recombinants in a short segment ( $\leq 4$ cM in the current example) are discouraged. Hence, the probabilities of the haplotypes with no double recombinants are much more probable than the corresponding ones with double recombinants, even though there is one more recombination in total. Again, through exact calculation, the configurations in  $A'-D'$  are

TABLE I. Haplotypes of individuals who define the various configurations

Individual	Haplotype configurations <sup>a</sup>					
	A-A''	B-B''	C-C''	D-D''	A-D	A'-D' A''-D''
1006	1 3	1 3	1 3	1 3		
	4 6	4 6	4 6	4 6		
	10 10	10 10	10 10	10 10		
	2 3	2 3	2 3	2 3		
	1 2	1 2	1 2	1 2		
	7 7	7 7	7 7	7 7		
	3 2	3 2	3 2	3 2		
	6 3	6 3	6 3	6 3		
	7 4	7 4	7 4	7 4		
	1007	3 3	3 3	3 3	3 3	
4 6		4 6	4 6	4 6		
1 8		1 8	1 8	1 8		
2 3		2 3	2 3	3 2		
6 6		6 6	6 6	6 6		
4 7		7 4	7 4	4 7		
2 4		4 2	4 2	2 4		
3 6		6 3	6 3	3 6		
4 3		3 4	3 4	4 3		
113		3 3	3 3	3 3	3 3	
	4 6	4 6	4 6	4 6		
	1 10	1 10	1 10	1 10		
	2 3	2 3	3 2	3 2		
	6 1	6 1	6 1	6 1		
	4 7	4 7	4 7	4 7		
	2 3	2 3	2 3	2 3		
	3 6	3 6	3 6	3 6		
	4 7	4 7	4 7	4 7		
	114	3 3	3 3	3 3	3 3	
6 6		6 6	6 6	6 6		
8 10		8 10	8 10	8 10		
3 3		3 3	3 3	3 3		
6 2		6 2	6 2	6 2		
4 7		4 7	4 7	4 7		
2 2		2 2	2 2	2 2		
3 3		3 3	3 3	3 3		
4 4		4 4	4 4	4 4		
1001						1 3
					3 3	3 3
					9 9	9 9
					5 4	5 4
					6 6	6 6
					3 4	4 3
					3 4	4 3
					1 5	5 1

<sup>a</sup>These abbreviations are used for the column headings: A-A''=A, A', A''; B-B''=B, B' B''; C-C''=C, C', C''; D-D''=D, D', D''; A-D=A, B, C, D; A'-D'=A', B', C', D'; A''-D''=A'', B'', C'', D''. Each entry in the table contains two columns representing the two haplotypes for the individual on the left.

found to be about 65 times more probable than their counter parts in A-D, while configurations in A''-D'' are 22 times more probable. Since the

TABLE II. Estimated conditional probabilities of haplotype configurations<sup>a</sup>

Configuration	m=0		m=4	
	100,000 iter.	100 iter.	100,000 iter.	100 iter.
A	0.07 (0.10)	0.07 (0.13)		
B	0.07 (0.10)	0.05 (0.09)		
C	0.28 (0.41)	0.22 (0.39)		
D	0.27 (0.39)	0.22 (0.39)		
A'			0.05 (0.10)	0.04 (0.06)
B'			0.05 (0.10)	0.09 (0.14)
C'			0.21 (0.41)	0.30 (0.46)
D'			0.20 (0.39)	0.22 (0.34)
A''			0.02 (0.12)	
B''			0.02 (0.12)	
C''			0.06 (0.35)	
D''			0.07 (0.41)	

<sup>a</sup>The numbers in parentheses in each column are relative probabilities, calculated based on the estimated conditional probabilities in the table within each of the three sets: A-D, A'-D', and A''-D''. The exact relative probabilities for each of these three sets of configurations are all found to be approximately 0.1, 0.1, 0.4, and 0.4. Only configurations with estimated conditional probabilities greater than 1% are shown here.

probabilities of the configurations in A''-D'' are only 1/3 of the probabilities of the corresponding configurations in A'-D', it should be harder to estimate them accurately, which is confirmed by the results shown in Table II. To obtain more accurate estimates for these smaller probabilities, more iterations are needed. Table II also shows that, with 100 iterations, only A'-D' has estimated probabilities greater than 1%, indicating that 100 iterations is insufficient for sampling the target distribution adequately under interference.

The above example with the EA dataset demonstrates several advantages of SIMPLE. First, multiple haplotype configurations exceeding a user-set probability threshold can be obtained with accurately estimated probabilities (with a reasonable number of Monte Carlo iterations). Second, compared to rule-based methods such as the MRH algorithm, SIMPLE is amenable to datasets with a larger amount of missing data. Although the EA dataset can be successfully tackled by both the MRH algorithm and SIMPLE, the former may not produce complete haplotype configurations in datasets with a larger amount of missing data, hence is limited in its usefulness in such situations. In contrast, SIMPLE and MCMC methods can realize their full potentials under these more difficult circumstances, as they are applicable to pedigrees with arbitrary amount and patterns of missing data.

Most importantly, by accounting for interference, the chance of inferring multiple recombinations in a short segment is greatly reduced. This is indeed a strength of likelihood-based approaches like SIMPLE, as such events are often regarded as errors [Schaid et al. 2002]. If interference is known to exist but is being ignored, the haplotypes inferred can be misleading, which may diminish the usefulness of these haplotypes in genetic studies or even lead to incorrect results. In this regard, the MRH criterion may not be the right objective. An MRHC with multiple recombinant haplotypes in a short segment of the genome may not be preferable to a haplotype configuration having all single recombinants, albeit a slight increase in the total number of recombinations. It is also worth noting that the popular software package GENEHUNTER [Kruglyak et al. 1996] does not take interference into account either for haplotype reconstruction. Furthermore, GENEHUNTER can only handle small pedigrees; it would have to drop 10 individuals in order to analyze the EA dataset, resulting in an incomplete haplotype configuration.

Despite these advantages, results obtained from SIMPLE, or any other Monte Carlo methods, should be carefully evaluated. Monte Carlo algorithms can take longer to run, and are subject to slow convergence, especially when there is a large amount of missing data. Furthermore, for pedigrees that are too complex to be peeled even at a single locus, SIMPLE is no longer feasible, in which case, MCMC methods are the only viable alternatives to date. Finally, we would also like to point out that, in the current implementation, SIMPLE assumes Hardy-Weinberg and linkage equilibrium, which, as cautioned by Schaid et al. [2002], can produce misleading haplotypes if the

markers are in fact in linkage disequilibrium. An investigation on how to lift these assumptions is currently underway.

#### ELECTRONIC-DATABASE INFORMATION

The URL for the software package SIMPLE is: <http://www.stat.ohio-state.edu/~statgen/SOFTWARE/SIMPLE>

#### ACKNOWLEDGMENTS

This work was supported in part by NSF grant DMS-9971770 (to S.L.).

#### REFERENCES

- Broman KW, Weber JL. 2000. Characterization of human crossover interference. *Am J Hum Genet* 66:1911–1926.
- Kruglyak L, Daly M, Reeve-Daly M, Lander E. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 58:1347–1363.
- Lin S, Speed TP. 1997. An algorithm for haplotype analysis. *J Comp Biol* 4:535–546.
- Lin S, Cheng R, Wright FA. 2001. Genetic crossover interference in the human genome. *Ann Hum Genet* 65:79–93.
- Litt M, Kramer P, Browne D, Ganchar S, Brunt ERP, Root D, Phromchotikul T, Dubay CJ, Nutt J. 1994. A gene for Episodic Ataxia/Myokymia maps to chromosome 12p13. *Am J Hum Genet* 55:702–709.
- Qian D, Beckmann L. 2002. MRH in Pedigrees. *Am J Hum Genet* 70:1434–1445.
- Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. 2002. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet* 71:992–995.
- Skrivanek Z, Lin S, Irwin M. 2003. Linkage analysis with sequential imputation. *Genet Epidemiol* 25:25–35.
- Sobel E, Lange K, O'Connell J, Weeks D. 1996. Haplotyping algorithms. In: Speed TP, Waterman MS, editors. Genetic mapping and DNA sequencing, IMA volumes in mathematics and its applications. New York: Springer-Verlag. p 89–110.