Genetic Epidemiology 27:80–83 (2004)

Letter to the Editor

Is Haplotype Block Identification Useful for Association Mapping Studies?

Weiwei Zhai,* Melissa Jane Todd, and Rasmus Nielsen

Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York

Grant sponsor: HFSP; Grant number: RGY0055/2001-M; Grant sponsor: NSF; Grant number: DEB-0089487; Grant sponsor: NSF/NIH; Grant number: DMS/NIGMS 0201037.
*Correspondence to: Weiwei Zhai, Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, 14853. E-mail: wz36@cornell.edu
Received 5 February 2004; Accepted 22 March 2004
Published online 14 May in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20014

To the Editor:

Much interest has recently focused on the prospect of using Single Nucleotide Polymorphisms (SNPs) for linkage disequilibrium mapping [e.g., Kruglyak, 1999; Reich et al., 2001; Pritchard and Przeworski, 2001]. Despite the rapid improvements in SNP typing technology, it is widely recognized that there are too many human SNPs to make it practical to genotype all known SNPs in a sample, so the problem arises: how do we select the SNPs that will be most informative for association studies? Several recent studies have suggested that LD in humans occurs in a block-like structure with large regions of high LD separated by short regions of low LD [Daly et al., 2001; Patil et al., 2001; Gabriel et al., 2002]. This observation has prompted the suggestions that identification of sets of so-called Haplotype Tag SNPs (htSNP) in blocks of high LD, representing the haplotype variation in the block, may be an efficient strategy for selecting SNPs for LD-mapping studies [Johnson et al., 2001; Zhang et al., 2002a]. Here we provide evidence in favor of the view that the haplotype block methods provide little or no increase in the statistical power of association mapping studies over more naive methods for identifying Tag SNPs.

To address the question whether protocols based on identifying haplotype blocks and then subsequently choosing Tag SNPs within blocks are the best strategy for maximizing mapping power, Zhang et al. [2002b] performed coalescence simulations to generate case-control samples. They compared the power of an association mapping test based on randomly chosen SNPs to the power based on htSNPs as identified by the method of Patil et al. [2001]. They showed that choosing htSNP based on this scheme led to improved power compared to randomly chosen SNPs. However, any real study will probably not be based on randomly chosen SNPs but rather on SNPs chosen on the basis of their frequency and their distribution in the genome [Kruglyak, 1999; Muller-Myhsok and Abel, 1997; Kaplan and Morris, 2001]. A very simple but realistic strategy for choosing SNPs without consideration of haplotype blocks, which researchers might apply in the absence of more sophisticated methods, might be simply to choose a fixed number of SNPs with a population frequency larger than a certain cut-off that are as evenly distributed as possible in the region of the genome being studied. Here we evaluate the efficiency of the haplotype block-based method by comparing it to this naive selection strategy.

Data were simulated under the neutral coalescent model using a modified version of the program by Hudson [2002]. We assumed an infinite sites model of mutation, and a scaled mutation rate of $4\,N_e\mu=200$, approximately corresponding to a 200-kb region in humans. We simulated cases with a constant recombination rate across the whole segment (scaled recombination rate $4N_er=200$) and cases with recombination hotspots [Jeffreys et al., 2001]. In the simulations with recombination hotspots, it was assumed that there were four hotspots (three of length 2 kb and one of length 4 kb [Jeffreys et al., 2001; Cullen et al., 2002]), in which the rate of recombination was elevated 10-fold. The total expected numbers of recombination events were identical in the two sets

of simulations, but in the simulations with hotspots about 1/3 of the recombination events occurred in the hotspots instead of 5% in the even case.

Two different models of population structure were assumed. In the first (no population subdivision), 26,000 haplotypes were generated assuming one panmictic population, and from these haplotypes, 13,000 diploid genotypes were formed assuming random mating. Fifty of these genotypes were randomly selected to form an ascertainment sample for SNP discovery [Carlson et al., 2003], haplotype block inference, and tag SNP selection. Case-control samples were then constructed from the remaining 12,950 genotypes (see below). In the second model (population subdivision), data from two populations were simulated according to a two-population symmetric island model [e.g., Hudson 2002] with scaled migration rate $2N_em=5.0$. 13,000 genotypes were generated, 50 from one population and 12,950 from the other population. The rate of recombination and mutation in each population was $4N_e\mu=4N_er=100$ [Tajima, 1989]. The first 50 genotypes were used for SNP discovery, haplotype block inference, and tag SNP selection, and the remaining genotypes from the other population were used for constructing case-control samples. The chosen value of $2N_em=5$ corresponds to an F_{ST} value on the order of 0.05 as typically observed among the major human ethnic groups [Makova et al., 2001].

In both simulation models, a multiplicative model for the disease risk was used. A random locus of minor allele frequency between 0.13 and 0.17 was selected as the disease locus. Labeling the disease alleles as d, and the other allele as D, individuals of genotype dd, dD, and DD have probability $c\gamma^2$, $c\gamma$, and c of having the disease. We let $c=0.042$ and $\gamma=1.6$ so that the prevalence of the disease in the population is approximately 0.05. These parameter values were chosen to ensure that a sufficient number of diseased individuals would occur in the simulations and to ensure that the power of the test was somewhat intermediate between 1 and the nominal level of the test. Five hundred cases and 500 controls were the selected among the population of 12,950 individuals.

For Tag SNP identification, we used the algorithm of Patil et al. [2001] implemented in the HapBlock computer program [Zhang et al., 2002a]. In short, a segment of consecutive SNPs is defined as being within a block if common haplotypes (whose frequency exceeds β) account for at least α percent of the observed haplotypes. As in Zhang et al. [2002b], we let $\alpha=80\%$ and $\beta=5\%$. Tag SNPs are selected as the minimum set of SNPs that can distinguish at least η (we let $\eta=80\%$) percent of the haplotypes. All SNPs with minor allele frequency less than 0.10 are eliminated before Tag SNP inference (for further details please refer to Zhang et al. [2002b]).

For the SNP selection criterion based on selecting the set of SNPs with the most even distribution along the sequence (even selection scheme), we chose the exact same number of SNPs as identified by the haplotype Tag SNP procedure. A heuristic algorithm was implemented to choose the set of SNPs with the least variance in the distance between the SNPs and with minor allele frequency larger than u, and u was allowed to vary from 0.1 to 0.25. In order to compare with the previous study of Zhang et al. [2002b], we also implemented an SNP selection procedure in which SNPs were chosen uniformly at random among all SNPs with minor allele frequency greater than u (random selection scheme). Case control association tests were performed for each marker using a Chi-Square test (or by Fisher's-Exact when the expected value of a cell entry was less than 5.0) and a Bonferroni correction was employed to control the type I error.

The results of 500 replicate simulations for each simulation scheme are presented in Table I. As in Zhang et al. [2002b], we find that, in the case of no hotspots and no population subdivision, the power of the HapBlock approach is higher than the power of the randomly chosen SNPs and not much reduced compared to the case where all SNPs are available. However, we also find that the SNP selection protocol based on finding a set of evenly distributed SNPs performs almost as well, or sometimes even better, than the Haplotype Block method. This general pattern also holds true in the presence of recombination hotspots. In the presence of population subdivision as modeled here, the HapBlock approach has a slight advantage over the evenly distributed selection scheme. In general, evenly distributed SNPs have the highest power if the cut-off for the minor allele frequency is relatively high. However, this particular aspect of the simulation study may be sensitive to the specific attributes of the assumed genetic disease model. In particular, it has been suggested that power is maximized when the frequency of the marker SNPs is close to the frequency of the disease allele [Muller-Myhsok and Abel, 1997; Kaplan and Morris, 2001].

Both the genetic model and the method for performing association testing used here are

TABLE I. Power of different marker selection schemes

	HapBlock	Even (0.1)	Even (0.15)	Even (0.2)	Even (0.25)	Random (0.1)	Random (0.15)	Random (0.20)	Random (0.25)	All (0.1)	All (0.15)	All (0.2)	All (0.25)
Uniform rec. rate No population structure 5% significance level	0.772	0.732	0.778	0.768	0.762	0.692	0.694	0.698	0.684	0.814	0.834	0.826	0.784
Uniform rec. rate No population structure 1% significance level	0.592	0.564	0.632	0.638	0.596	0.544	0.560	0.562	0.542	0.680	0.686	0.680	0.630
HotSpots No population structure 5% significance level	0.780	0.758	0.770	0.782	0.750	0.728	0.724	0.728	0.692	0.820	0.826	0.824	0.774
HotSpots No population structure 1% significance level	0.612	0.584	0.624	0.638	0.616	0.556	0.578	0.578	0.546	0.678	0.686	0.684	0.640
Uniform rec. rate Population structure 5% significance level	0.728	0.704	0.710	0.714	0.710	0.682	0.676	0.696	0.686	0.808	0.796	0.770	0.744
Uniform rec. rate Population structure 1% significance level	0.584	0.516	0.530	0.558	0.550	0.506	0.534	0.542	0.524	0.630	0.598	0.596	0.568

relatively simple. Nonetheless, the results clearly show that for this simple testing scheme, there is often no advantage in selecting Tag SNPs based on haplotype blocks because simpler methods based only on SNP frequency and spacing have equal power. The conclusions of this study seriously challenge the utility of concentrating further efforts on devising SNP selection strategies based on inferred haplotype blocks. Future algorithmic improvements of the haplotype block-based selection schemes may improve their utility. However, the results of this study suggest that other considerations, beyond haplotype block structure, such as SNP frequency and disease model, should be taken into consideration when choosing Tag SNPs.

ACKNOWLEDGMENTS

We thank Andy Clark for discussion and two anonymous referees for their valuable comments. This work was supported by HFSP grant RGY0055/2001-M, NSF grant DEB-0089487, and NSF/NIH grant DMS/NIGMS 0201037 to R.N.

REFERENCES

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33:518–521.

Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. 2002. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet 71:759–776.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High resolution haplotype structure in the human genome. Nat Genet 29:229–232.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222.

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Genova GD, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Phillipa C, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237.

Kaplan N, Morris R. 2001. Issues concerning association studies for fine mapping as a susceptibility gene for a complex disease. Genet Epidemiol 20:432–457.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144.

Makova KD, Ramsay M, Jenkins T, Li W. 2001. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics 158:1253–1268.

Muller-Myhsok B, Abel L. 1997. Genetic analysis of complex diseases. Science 275:1328-1329.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Lee DH, McDonough DP, Nguyen BN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas

KR, Frazer KA, Fodor SPA, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Pitcher DJ, Lavery T, Kouyoumjian R, Rfarhadian SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Tajima F. 1989. DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 123:229–240.

Zhang K, Deng M, Chen T, Waterman MS, Sun F. 2002a. A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339.

Zhang K, Calabrese P, Nordborg M, Sun F. 2002b. Haplotype block structure and its application to association studies: power and study designs. Am J Hum Genet 71:1386–1394.