Pacific Symposium on Biocomputing 11:454-466(2006)

RELAXING HAPLOTYPE BLOCK MODELS FOR ASSOCIATION TESTING

NATALIE CASTELLANA, KEDAR DHAMDHERE, SRINATH SRIDHAR,
AND RUSSELL SCHWARTZ*

Computer Science Department and Biological Sciences Department
Carnegie Mellon University
4400 Fifth Avenue
Pittsburgh, PA 15213 USA
Email: russells@andrew.cmu.edu

The arrival of publicly available genome-wide variation data is creating new opportunities for reconciling model-based methods for associating genotypes and phenotypes with the complexities of real genome data. Such data is particularly valuable for testing the utility of models of conserved haplotype structure to association studies. While there is much interest in "haplotype block" models that assume population-wide regions of low diversity, there is also evidence that such models eliminate correlations potentially useful to association studies. We investigate the value of relaxing the rigidity of block models by developing an association testing method using the previously developed "haplotype motif" model, which retains the notion of representing haploid sequences as concatenations of conserved haplotypes but abandons the assumption of population-wide block boundaries. We compare the effectiveness of motif, block, and single-variant models at finding association with simulated phenotypes using real and simulated data. We conclude that the benefits of haplotype models in any form are modest, but that haplotype models in general and block-free models in particular are useful in picking up correlations near the boundaries of the detectable level.

1. Introduction

Searches for correlations between human genetic variations and disease phenotypes have often been fruitful for strongly hereditary diseases, but have had limited success at finding genetic risk factors for complex diseases. This failure is likely due at least in part to the challenges of distinguishing many relatively weak correlations from the noise produced by chance associations with the millions of known sites of common variation. One approach to address this problem involves identifying segments of correlated variations

*to whom correspondence should be addressed

Pacific Symposium on Biocomputing 11:454-466(2006)

known as haplotypes. By finding these co-associating sets of variations, one can in principle reduce the amount of data to be collected and analyzed in an association study and avoid some of the confounding effects of testing many variant sites. The prospects of such methods may be greatly facilitated by the recent construction of the HapMap[6], a publicly available collection of genome-wide single nucleotide polymorphism (SNP) variations separated by donor to allow for haplotype inference.

Studies of simulation models[22] and limited amounts of real data[1] have suggested potentially large advantages to haplotype-based association methods. Many such methods are based on the haplotype block model[3], which proposed that the genome consists of discrete regions of strongly correlated variations separated by recombination hotspots, across which correlations have been eliminated by frequent historical recombination. Numerous block construction criteria have since been proposed, generally based either on haplotype diversity or similar metrics[13] or on linkage disequilibrium statistics[5]. A haplotype-based association test may be conducted by directly testing for differences in frequencies of common haplotypes in individuals affected (cases) or unaffected (controls) by a disease[2,1]. Or they may use haplotypes to identify "haplotype tagging SNPs" (htSNPs), a subset of SNPs that contain most of the information contained in the full SNP set[9], which can reduce the cost of genotyping and the difficulty of finding meaningful associations in the resulting data.

While haplotype block models appear useful in facilitating association studies by reducing data complexity, there is evidence that they do not robustly capture true underlying haplotype conservation patterns[15]. In prior work, we developed the "haplotype motif" model[16], which explains individual genomes as concatenations of conserved haplotypes (or isolated variant sites). This model relaxes some of the rigidity of the block models, while still maintaining enough structure to allow for robust fitting[16] and efficient application to various computational analyses[17]. Figure 1 illustrates the difference between block and motif models of haplotype structure on a small hypothetical set of sequences. Other groups have since also developed "haplotype motif" models using various optimization metrics[18,10].

Here, we focus on the ultimate test of such relaxed conserved haplotype models: Does the extra information they preserve relative to block models provide an advantage in association testing? We approach that question with an empirical study of single-SNP, haplotype block, and haplotype motif methods for finding associations between genotype and phenotype, using simulated and HapMap data to understand how idealized models

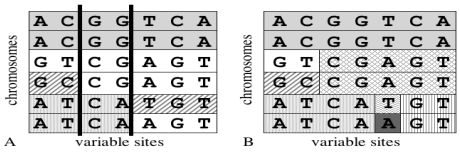

Figure 1. Illustration of possible block and motif partitions of a hypothetical set of sequences, each corresponding to variable sites in a given region of one chromosome from one individual. A: A block model, in which each chromosome is explained as a choice of one haplotype in each of three blocks. B: A motif model, in which each chromosome is explained as a concatenation of a set of "haplotype motifs" of varying length.

might mislead us with regard to real data. While we focus on the haplotype motif model, our interest is not specifically in that model *per se*, but rather in whether exploiting correlation information across haplotype block boundaries can lead to improvements in association study effectiveness.

2. Methods

2.1. *Haplotype Structure and Tagging SNP Inference*

Haplotype motif structure was inferred as described in Schwartz[16] with significance level 0.001 and maximum motif length 10. The method first identifies candidate motifs by finding each subsequence whose population frequency is significantly higher than would be predicted from the frequencies of substrings from which it might be assembled. It then uses an iterative algorithm to repeatedly explain the training sequences as maximum-likelihood concatenations of individual motifs then use these explanations to improve estimates of the motif frequencies. The reader is referred to Schwartz[16] for algorithm details. Tag SNPs were determined from the motif structure by a dynamic programming algorithm[17] to give an estimated maximum prediction error of 5% for each hidden SNP site on the training data. The maximum motif length was dictated by the prohibitive computational cost of htSNP selection using long motifs. The significance level is a program default selected to provide high confidence that almost all motifs represent truly conserved haplotypes.

We used two haplotype block methods, both implemented with the dynamic programming algorithm of Zhang et al.[21] One method we call bounded blocks finds block partitions by minimizing the total number of observed haplotypes over all blocks, subject to a maximum block length. The results reported here used a maximum length of 5, although we found

nearly identical results with maximum length 10 (data not shown). We also used a simplified version of LD-based testing called four-gamete blocks, which minimizes the number of blocks given that sequences in each block must be consistent with a perfect phylogeny, or, equivalently, cannot have all four possible gametes for any pair of SNP sites. While many block metrics have been proposed, we chose these two as representatives because they tend to stress two different criteria that should make for a "good" model for association testing: few blocks and thus few distinct association tests (four gamete) or few haplotypes per block and thus less confounding from unassociated haplotypes (bounded block). Minimal tag SNPs were selected within each block by exhaustive enumeration.

2.2. *Association Testing Methods*

For each SNP in a data set, we counted occurrences of all motifs overlapping that SNP for which the total population frequency of the motif exceeded 10%. All other motifs were grouped into a common "other" class. We then performed a chi-square test of association on the contingency table of motif classes and case/control status. As the motif method does not separate the genome into putatively uncorrelated regions, we established p-values by a permutation test. We randomly reassigned case and control labels while preserving the size of each class and computed chi-square values as above, recording the maximum statistic value for each degree of freedom. P-values were estimated from one thousand permutations per data set.

With the two block methods, one class was developed for each of the three most common haplotypes in each block. All other haplotypes were assigned to a fourth class. A chi-square test of significance was applied to the two-by-four table of the haplotype classes and case/control statuses. As our concern in this study is whether correlations across block boundaries are useful to association studies, we exclude such cross-boundary correlation information by assuming no such correlations and using Bonferroni correction for multiple hypothesis control.

Association tests were also performed on individual SNPs using a chi-square test of the two-by-two contingency table of SNP allele and case-control status. Control for multiple hypotheses was again performed by Bonferroni correction assuming no correlations between SNPs. The same protocol was used to test association with tag SNPs.

To assess the influence of cross-block correlations, we repeated all Bonferroni-corrected tests with permutation tests using 1,000 permutations.

Pacific Symposium on Biocomputing 11:454-466(2006)

2.3. *Data Processing*

We evaluated the methods using two real and three simulated data sets. We downloaded phased data from a high-density 500 kb region of 7q21.13 from the ENCODE resequencing project[4] and the full chromosome 22 HapMap data set[6]. We believe the 7q21 data is a good approximation of the data to be expected in a candidate gene study while the larger but sparser chromosome 22 data provides a better approximation to the challenges involved in whole-genome studies. For each real data set, we removed all SNPs that were not variant in all four HapMap population groups: CEPH (Utah Residents with Northern and Western European ancestry); Han Chinese in Beijing, China; Japanese in Tokyo, Japan; and Yoruba in Ibadan, Nigeria. We were left with 548 such universal SNPs out of 1,523 total for the 7q21 data, an average marker distance of 912 bases, and 11,900 universal SNPs out of 19,250 for chromosome 22, an average marker distance of 4.7 kb.

Simulated data was generated by coalescent simulation under a Wright-Fisher neutral model using the ms program[7]. We followed a protocol developed for a prior empirical study of the utility of block and motif models for information compression[19]. We used a mutation rate of 2.5×10^{-8} per nucleotide per generation, a recombination rate of 10^{-8} per pair of sites per generation, and an effective population size of 10,000 based on estimated values of the human mutation[11] and recombination[8] rates and effective population size[14]. Each simulated data set consisted of 2,000 chromosomes in a region of 100,000 segregating sites, representative of a 100 kb genomic region. The resulting sequences were screened to remove any SNPs with population frequency below 10%. Pairs of sequences were combined at random assuming Hardy-Weinberg equilibrium to assign chromosomes to individuals. A total of 220 simulated population samples were created.

Simulated disease phenotypes were artificially imposed on all data sets. A disease SNP was assigned for each sample from among SNPs having population frequency between 40% and 60%. The disease SNP was also required to be within a region between 45% and 55% of the distance along the chromosome for real data or between 40% and 60% for simulated data. For simplicity, an additive model with a single disease penetrance parameter, p, was used to determine disease risk. Individuals homozygous for the disease allele had probability p of having the disease, those homozygous for the non-disease allele had probability $1 - p$ of the disease, and others were assumed to have equal probability of having the disease or not. Individuals (pairs of chromosomes) were assigned to case and control sets accordingly. For the

real data, cases and controls were assigned independently for each of the four population groups and any excess of cases over controls or vice-versa was discarded for each before pooling the four ethnicities for association testing. This protocol ensures an equal number of members of each group in the cases and controls in order to better simulate the demographically matched cases and controls to be expected in real association study data sets. For each simulated sample, a single case and a single control set were assigned and any excess of cases over controls or vice-versa was discarded.

For the 7q21 data, five case/control sets were constructed for each penetrance value from 55% to 100% in increments of 5%. For the chromosome 22 data, five case/control sets were constructed for each penetrance value from 60% to 100% in increments of 10%. One simulated data set was constructed by creating ten case/control sets for each penetrance from 55% to 100% in increments of 5% and a second by creating ten case/control sets for each penetrance from 51% to 60% in increments of 1%. A final set was developed for specificity testing using twenty sets of individuals assigned randomly to cases and controls independent of genotype.

Each association method was applied to all data sets. Success was evaluated by testing the fraction of associations detected at LOD cutoff values of 3 (p-value 0.001), 2.5 (p-value 0.0032), 2 (p-value 0.01), and 1.5 (p-value 0.032). Sensitivity was tested on the randomly assigned cases and controls at LOD cutoffs 3, 2.5, 2, 1.5, and 1 (p-value 0.1).

3. Results

We derived haplotype motif structures for all data sets and performed a visual inspection of the motif patterns for the real data. Figure 2 depicts the motif patterns assigned to the 7q21 region and a representative subregion of chromosome 22 selected for illustrative purposes. Both datasets are overwhelmingly assigned to motifs of the maximum allowed length (10 SNPs), suggesting considerable conserved structure at both marker densities. Common motifs are nearly identical between the two Asian samples, often shared between the Asian and European-ancestry samples, occasionally shared between the Asian and Yoruba samples, and very rarely shared by Yoruba and European but not Asian samples. These results provide an informal check on the method as they are consistent with recent reconstruction of the likely human evolutionary tree[20].

We began our quantitative analysis using the 7q21 data set. Figure 3 shows the results. All methods consistently fail for penetrance below 70%

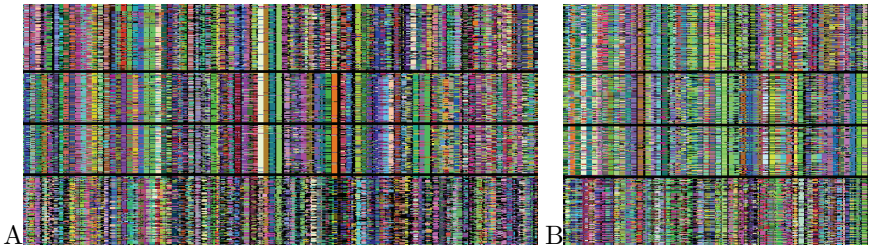

Figure 2. Visualization of haplotype motif assignments. Rows correspond to different chromosomes and columns to different SNP sites. Contiguous bars of a single color represent a single motif, with bars of the same color above and below them representing copies of that motif in other individuals. Each image shows motifs for the four HapMap populations (European-ancestry, Han Chinese, Japanese, and Yoruba) in order from top to bottom, separated by solid black rows. A: Motif assignments for a representative region of chromosome 22. B: Motif assignments for the 7q21 region.

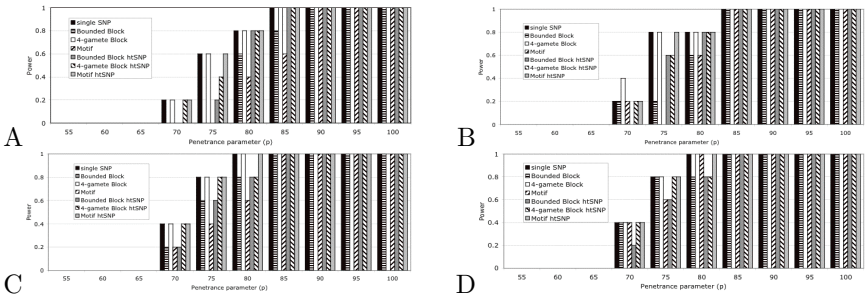

Figure 3. Power in identifying associations in 7q21 data. A: LOD cutoff 3; B: LOD cutoff 2.5; C: LOD cutoff 2; D: LOD cutoff 1.5

and succeed for penetrance values of at least 90% for LOD 3 or 85% for LOD 2.5 and below. The methods are therefore distinguishable only on a relatively narrow set of penetrances. Within that range, the straightforward motif method generally showed the least power. The single-SNP, 4-gamete block, and motif-based htSNP methods were most successful, with the 4-gamete block method outperforming the other two in one case.

We next examined the chromosome 22 data set. Preliminary visual inspection of the results confirms that the methods are finding significant associations only in the region of the disease SNP. While space does not permit us to present all of the detailed SNP-by-SNP scans, Fig. 4 shows a representative set of images from a sample with 90% penetrance. At the full-chromosome resolution, all methods show a single significant spike at

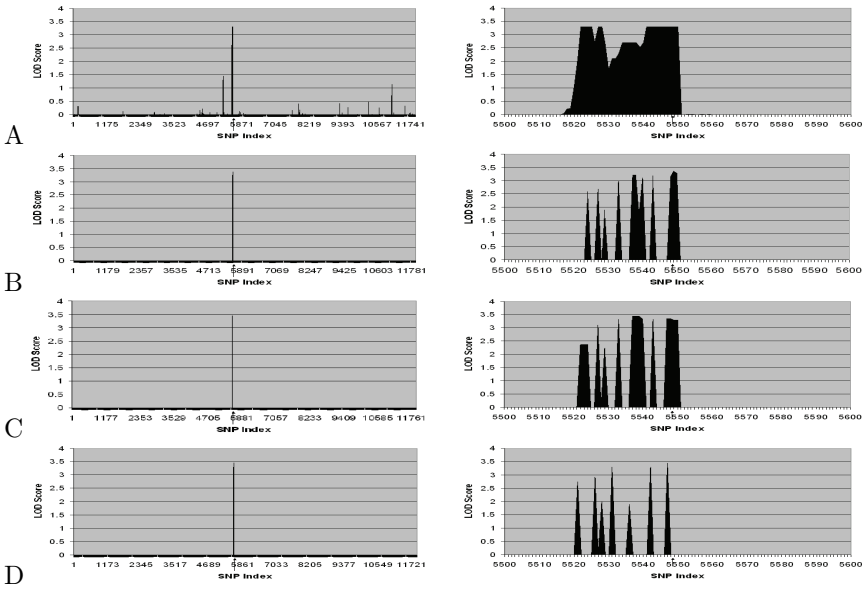

Figure 4. Representative selection of association scans for chromosome 22 with a simulated 90% penetrance disease SNP. Each line shows SNP-by-SNP LOD scores for the full chromosome (left) and a one hundred SNP region (right) around the disease site (marked by arrows). A: Motifs; B: Individual SNPs; C: Four-gamete blocks; D: Motif-selected htSNPs. Values below zero after Bonferroni correction are truncated to zero. Motif LOD scores above 3 (p-value 0.001) cannot be accurately estimated due to the use of a 1,000 trial permutation test and are therefore arbitrarily set to 3.3 (p-value 0.0005).

the location of the disease SNP (SNP position 5549). The motif method shows many smaller distant spikes, while the other methods do not. This is attributable to the fact that we assumed independence between tests for the other methods when correcting for multiple hypotheses and the resulting correction appears overly conservative. In the close-up view, all methods show a region of significant association centered slightly to the left of the disease SNP. The motif method shows significant associations across this entire region, while the others all show isolated spikes of association separated by regions of no association. This suggests that there are conserved haplotypes associated with the disease SNP spanning the entire region that are found by the motif method but often lost to the block methods.

Figure 5 shows the sensitivity of the methods on the chromosome 22 data. As with the 7q21 data, the methods behave identically for most

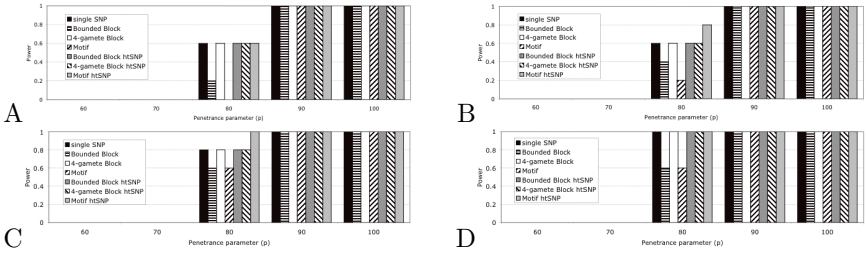

Figure 5. Power in identifying associations in chromosome 22 data. A: LOD cutoff 3; B: LOD cutoff 2.5; C: LOD cutoff 2; D: LOD cutoff 1.5

parameter values, successfully identifying association for all 90-100% penetrance data sets and failing for all 60-70% penetrance data sets. At 80% penetrance, the motif-based htSNP method is the most powerful, while the standard motif method is the least powerful. Bounded blocks also perform poorly, while all other methods perform equally well.

Although we examined here only one chromosome, we can extrapolate our results to a full-genome scan. Chromosome 22 is approximately 56 Mb, or about 1.8% of the human genome, which allows us to estimate that LOD scores detected by our methods would be approximately 1.8 lower if corrected for analysis to the full data set. Thus, the LOD 3 cutoff results would correspond to approximately a 94% confidence in a full genome scan.

We then analyzed the simulated data sets. We began by considering a broad range of penetrance parameters, 55% to 100% in increments of 5%. All methods are consistently successful for penetrance values of at least 75%. Occasional successes are observed even as low as 55%, suggesting that the larger population size used in the simulated test allows even relatively weak effects to be detected. The motif method appears most successful at detecting the weakest effects (penetrance 55%) but is the least successful on high-penetrance data sets. Overall, the motif-based htSNP method appears marginally the best at detecting associations in these data.

We then focused on the most difficult cases, examining simulated data with penetrances for each integer value from 51% to 60%. Figure 7 shows the results of these trials. No method was consistently dominant. The motif method appeared most successful on the hardest examples (penetrance 51%-55%). This success may be attributable to its better ability to find some conserved haplotype correlating with the disease SNP if any exists or it may be because its permutation test allows it to exploit correlations across block boundaries, a capability not permitted for the other methods.

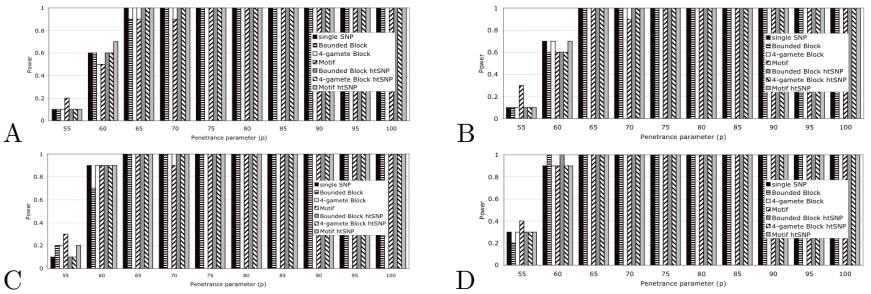

Figure 6. Power for penetrances 55%-100% using coalescent simulated data. A: LOD cutoff 3; ; B: LOD cutoff 2.5; C: LOD cutoff 2; D: LOD cutoff 1.5

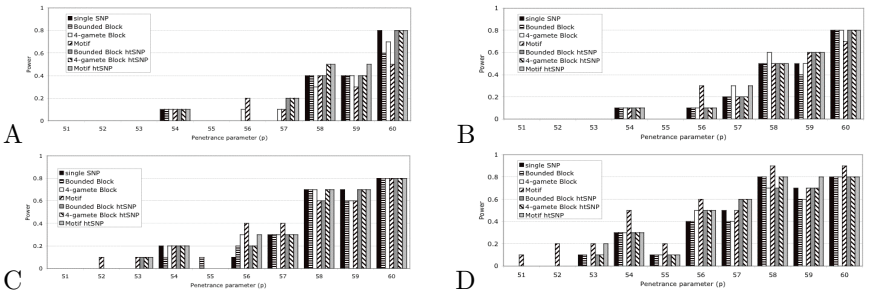

Figure 7. Power for penetrances 51%-60% using coalescent simulated data. A: LOD cutoff 3; B: LOD cutoff 2.5; C: LOD cutoff 2; D: LOD cutoff 1.5

Motif htSNPs become the most successful toward higher penetrance values.

Given indications that there is information useful to association tests lost by the assumption of independent blocks, we further asked whether dropping this assumption could recover some of that information. We therefore repeated all block and single-variant tests, replacing Bonferroni correction with permutation tests. Figure 8 shows power for chromosome 22 and coalescent simulated data at LOD 3. Compared to the prior Bonferroni-corrected graphs, the permutation tests yield a noticeable improvement in the block methods and a slight improvement for the block htSNPs. The motif htSNP method does not improve, suggesting that the assumption of independence is more nearly true of motif-selected htSNPs than block-selected htSNPs. Additional tests at other LOD values (data not shown) confirm that permutation tests lead to improved sensitivity of block, block htSNP, and single SNP tests, but not to motif htSNP tests.

We finally assessed the specificity of the methods using twenty samples

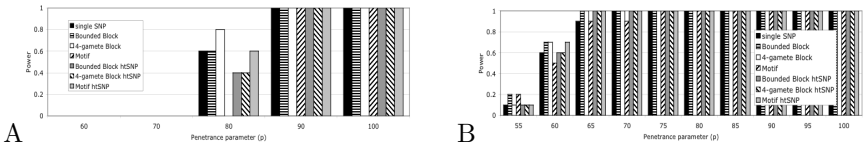

Figure 8. Power for permutation-test variants of all methods. A: chromosome 22 with LOD cutoff 3; B: simulated data with LOD cutoff 3.

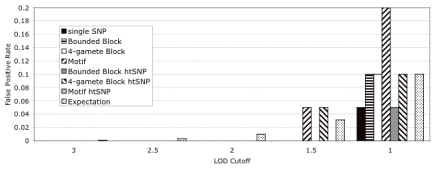

Figure 9. False positive identifications on twenty trials of randomly assorted simulated cases and controls as a function of LOD cutoff.

of randomly assorted simulated cases and controls. Figure 9 shows false positive rates as functions of the LOD cutoff. All values are within what can reasonably be expected by chance for each LOD score. All but motif htSNPs produce at least one false positive on 20 trials at LOD 1, while no methods produce any false positive values at LOD 2 or higher. Half of the methods using Bonferroni correction achieved exactly the expected number of errors at LOD 1 (2 errors); this appears to undermine our prior indication from chromosome 22 data that the Bonferroni correction is overconservative, suggesting that it is not excessively so.

4. Discussion

This study was intended to determine whether relaxing rigid block boundaries in models of haplotype structure would improve their utility for association tests. Our results provide an ambiguous answer to that question. Only a narrow range of parameter values discriminates between the methods for any data set, suggesting that the benefits of any one method over any other are relatively modest. The motif-based htSNP test appears to be marginally the best overall for both real and simulated data, consistent with prior work showing that motifs provide a clear advantage over comparable block methods at robustly selecting small htSNP sets[17]. The pure motif method does generally poorly, which we conjecture occurs because it tends

to produce too many motifs covering each site, confounding the chi-square statistic. This could be an inherent problem of the motif approach or might be resolved by a different motif inference method or test statistic. Motifs do, however, appear to be the best for detecting the weakest correlations.

Both motif-based methods work comparatively better on simulated than on real data, suggesting that block-like patterns are more pronounced in real than in simulated data, even if they do not fully describe either. This conclusion is consistent with an emerging consensus in the field that inferred block patterns do not entirely reflect an inherent "blockiness" due to recombination hotspots, as was first proposed[3], but neither are they are fully explicable from uniform recombination rate models of human population history[12]. The conclusion is further supported by the improvement exhibited in block tests when using a permutation test rather than Bonferroni correction. The assumption of the block model that correlation information is captured within blocks is only partially valid, and methods for recovering that information either at the stage of model construction or application of the association test can lead to improved power. Both motif methods work better on the chromosome 22 data than on the denser 7q21 data, possibly because computational resource constraints limit us to short motifs (maximum of 10 SNPs in these tests), preventing them from taking advantage of long-range correlations in a dense marker set.

None of the methods considered — SNP, block, or motif — consistently dominates the others in all conditions; each can be expected to find some associations that would be missed by others. It would be self-defeating in practice to apply many methods to every data set, as the correction for multiple hypotheses would likely eliminate the small advantages of different methods for different cases. However, using a small number of very different methods may have advantages over applying only one "best" method. If we were to recommend one method from among those we examined, it would be an htSNP method. But if we were to recommend two, the second would be the motif method, as it is most likely to find associations the first misses. The field of association testing may benefit most by seeking a diversity of approaches, with particular emphasis on finding a few niche methods, like motifs, that are strongest in cases where others methods are weakest.

Acknowledgments

We thank R. Ravi and G. Blelloch for helpful discussions and comments on this manuscript. This work was supported in part by NSF awards #0122581

and #0346981 and by the Merck Program for Computational Biology and Chemistry at Carnegie Mellon University.

References

1. J. Akey, L. Jin, and M. Xiong. *Eur. J. Hum. Genet.* **9**, 291, (2001).
2. N.H. Chapman and E.M. Wijsman. *Am. J. Hum. Genet.* **63**, 1872 (1998).
3. M.J. Daly, J.D. Rioux, S.F. Schaffner, and T.J. Hudson, *Nat. Genet.* **29**, 229 (2001).
4. The ENCODE Project Consortium. *Science.* **306**, 636 (2004).
5. S. Gabriel, S. Schaffner, H. Nguyen, J. Moore, J. Roy, B. Blumenstiel, J. Higgens, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altschuler. *Science.* **296**, 2225 (2002).
6. The International HapMap Consortium. *Nature.* **426**, 789 (2003).
7. R.H. Hudson. *Bioinform.* **18**, 337 (2002).
8. M.I. Jensen-Seaman, T.S. Furey, B.A. Payseur, Y. Lu, K.M. Roskin, C.-F. Chen, M.A. Thomas, D. Haussler, and H.J. Jacob. *Genome Res.* **14**, 528 (2004).
9. G.C. Johnson, L. Esposito, B.J. Barret, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbrigde, R.C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C. Gough, D.G. Clayton, and J.A. Todd. *Nat. Genet.* **29**, 233 (2001).
10. M. Koivisto, P. Rastas, and E. Ukkonen. *Lect. Notes Comp. Sci.* **3113**, 159 (2004).
11. M.W. Nachman and S.L. Crowell. *Genetics.* **156**, 297 (2000).
12. M. Nordborg and S. Tavaré, *Trends Genet.* **18**, 83 (2002).
13. N. Patil, A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, C. Marjoribanks, D.P. McDonough, B.T. Nguyen, M.C. Norris, J.B. Sheehan, N. Shen, D. Stern, R.P. Stokowski, D.J. Thomas, M.O. Trulson, K.R. Vyas, K.A. Frazer, S.P. Fodor, and D.R. Cox. *Science.* **294**, 1719 (2001).
14. B. Rannala and Z. Yang. *Genetics*, **164**, 1645, (2003).
15. R. Schwartz, B. Halldórsson, V. Bafna, A.G. Clark, and S. Istrail. *J. Comp. Biol.* **10**, 13 (2003).
16. R. Schwartz, *Proc. IEEE Comp. Sys. Biotech. Conf.*, 306 (2003).
17. R. Schwartz, *Proc. IEEE Comp. Sys. Biotech. Conf.*, 90 (2004).
18. J. Sheffi. *MIT Comp. Sci. M.Eng. Thesis*, 2004.
19. S. Sridhar, K. Dhamdhere, G. E. Blelloch, R. Ravi and R. Schwartz. *Carnegie Mellon Comp. Sci. Tech Report,* **CMU-CS-040166** (2004).
20. S. Tishkoff and K.K. Kidd. *Nat. Genet.* **36**, S21 (2004).
21. K. Zhang, M. Deng, T. Chen, M. S. Waterman and F. Sun. *Proc. Natl. Acad. Sci. USA*. **99**, 7335 (2002).
22. K. Zhang, P. Calabrese, M. Nordborg, and F. Sun. *Am. J. Hum. Genet.* **71**, 1386 (2002).