Hybrid Approaches for the Selection of Tag SNPs based on Block Partition and Linkage Disequilibrium

Chia-Jung Chang[1,*], Yao-Ting Huang[2], and Kun-Mao Chao[1,3,4]

[1] Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei 106, Taiwan, ROC
d98922004@csie.ntu.edu.tw

[2] Department of Computer Science and Information Engineering,
National Chung Cheng University,
Chia-Yi 621, Taiwan, ROC
ythuang@cs.ccu.edu.tw

[3] Graduate Institute of Biomedical Electronics and Bioinformatics,
National Taiwan University,
Taipei 106, Taiwan, ROC

[4] Graduate Institute of Networking and Multimedia,
National Taiwan University,
Taipei 106, Taiwan, ROC
kmchao@csie.ntu.edu.tw

Received 15 October 2010; Revised 15 November 2010; Accepted 10 December 2010

Abstract. Studies over recent years have revealed that a small subset of SNPs (called tag SNPs) is sufficient to capture the haplotype diversity in high linkage disequilibrium (LD) regions. Methods for finding tag SNPs are mainly based on either the assumption that the haplotypes in the human genome conform to a block-like structure or based on the extent of LD across the human genome. The block-based methods identify tag SNPs that are compression of haplotypes in each block but is sensitive to the the predefined block partition. On the other hand, the LD-based methods is free from the block partition but often produce numerous singleton tag SNPs having no sufficient LD with others. We design and implement two hybrid methods which reduce the number of tag SNPs by both considering the block-like structure and the extent of LD in distinct blocks. A faster algorithm is proposed to boost the computational efficiency of LD among all pairs of SNPs and is internally used by these two hybrid methods. The experimental results indicate that the hybrid methods not only reduce the numbers of tag SNPs but also run much faster than existing LD- or block-based methods such as ldSelect, Tagger, HapBlock and GPT.

Keywords: SNP, haplotype block, linkage disequilibrium

References

[1] Consortium TIH, "A Haplotype Map of the Human Genome," *Nature International Weekly Journal of Science*, Vol. 437, No. 7063, pp. 1299-1320, 2005.

[2] L. Helmuth, "Genome Research: Map of the Human Genome 3.0," *Science Magazine*, Vol. 293, No. 5530, pp. 583-585, 2001.

[3] D. A. Hinds, L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, D. R. Cox, "Whole-Genome Patterns of Common DNA Variation in Three Human Populations," *Science Magazine*, Vol. 307, pp. 1072-1079, 2005.

* Correspondence author

[4] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, E. S. Lander, "High-resolution Haplotype Structure in the Human Genome," *Nature Genetics*, Vol. 29, No. 2, pp. 229-232, 2001.

[5] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, D. Altshuler, "The Structure of Haplotype Blocks in the Human Genome," *Science Magazine*, Vol. 296, No. 5576, pp. 2225-2229, 2002.

[6] 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. N. 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. A. Fodor, D. R. Cox, "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21," *Science Magazine*, Vol. 294, pp. 1719-1723, 2001.

[7] K. Zhang, Z.S. Qin, J.S. Liu, T. Chen, M. S. Waterman, F. Sun, "Haplotype Block Partition and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies," *Genome Research*, Vol. 14, pp. 908-916, 2004.

[8] V. Bafna, B. V. Halldorsson, R. Schwartz, A. G. Clark, S. Istrail, "Haplotypes and Informative SNP Selection Algorithms: Don't Block Out Information," In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, pp. 19-27, 2003.

[9] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, D. A. Nickerson, "Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium," *The American Journal of Human Genetics*, Vol. 74, No. 1, pp. 106-120, 2004.

[10] S. Lin, A. Chakravarti, D. J. Cutler, "Exhaustive Allelic Transmission Disequilibirum Tests as a New Approach to Genome-wide Association Studies," *Nature Genetics*, Vol. 36, pp. 1181-1188, 2004.

[11] Z. S. Qin, S. Gopalakrishnan, G. R. Abecasis, "An Efficient Comprehensive Search Algorithm for TagSNP Selection Using Linkage Disequilibrium Criteria," *Bioinformatics*, Vol. 22, No. 2, pp. 220-225, 2006.

[12] D. O. Stram, C. A. Haiman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel, B. E. Henderson, M. C. Pike, "Choosing Haplotype-tagging SNPs based on Unphased Genotype Data Using a Preliminary Sample of Unrelated Subjects with an Example from the Multiethnic Cohort Study," *International Journal of Human and Medical Genetics*, Vol. 55, No. 1, pp. 27-36, 2003.

[13] M. E. Weale, C. Depondt, S. J. Macdonald, A. Smith, P. S. Lai, S. D. Shorvon, N. W. Wood, D. B. Goldstein, "Selection and Evaluation of Tagging SNPs in the Neuronal-Sodium-Channel Gene SCN1A: Implications for Linkage-Disequilibrium Gene Mapping," *The American Journal of Human Genetics*, Vol. 73, No. 3, pp. 551-565, 2003.

[14] K. Zhang, M. Deng, T. Chen, M. S. Waterman, F. Sun, "A Dynamic Programming Algorithm for Haplotype Block Partitioning," In *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99, No. 11, pp. 7335-7339, 2002.

[15] D. C. Crawford and D. A. Nickerson, "Definition and Clinical Importance of Haplotypes," *Annual Review of Medicine*, Vol. 56, pp. 303-320, 2005.

[16] Y.T. Huang, K. Zhang, T. Chen, K.M. Chao, "Selecting Additional Tag SNPs for Tolerating Missing Data in Genotyping," *BMC Bioinformatics*, Vol. 6, pp. 263-278, 2005.

[17] C.K. Ting, W.T. Lin, Y.T. Huang, "Multi-objective Tag SNPs Selection Using Evolutionary Algorithms," *Bioinformatics*, Vol. 26, No. 11, pp. 1446-1452, 2010.

[18] J. C. Barrett, B. Fry, J. Maller, M. J. Daly, "Haploview: Analysis and Visualization of LD and Haplotype Maps," *Bioinformatics*, Vol. 21, No. 2, pp. 263-265, 2005.

[19] K. Zhang and L. Jin, "HaploBlockFinder: Haplotype Block Analyses," *Bioinformatics*, Vol. 19, No. 10, pp. 1300-1301, 2003.

[20] C.J. Chang, Y.T. Hunag, K.M. Chao, "A Greedier Approach for Finding Tag SNPs," *Bioinformatics*, Vol. 22, No. 6, pp. 685-691, 2006.

[21] K. Zhang, Z. Qin, T. Chen, J.S. Liu, M. S. Waterman, F. Sun, "HapBlock: Haplotype Block Partition and Tag SNP Selection Software Using as Set of Dynamic Programming Algorithms," *Bioinformatics*, Vol. 21, No. 1, pp. 131-134, 2005.

[22] P. I. D. Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, D Altshuler, "Efficiency and Power in Genetic Association Studies," *Nature Genetics*, Vol. 37,No. 11, pp. 1217-1223, 2005.

[23] E. Halperin and E. Eskin, "Haplotype Reconstruction from Genotype Data Using Imperfect Phylogeny," *Bioinformatics*, Vol. 20, No. 12, pp. 1842-1849, 2004.

[24] Y.T. Huang, K.M. Chao, T. Chen, "An Approximation Algorithm for Haplotype Inference by Maximum Parsimony," *Journal of Computational Biology*, Vol. 12, pp. 1261-1274, 2005.

[25] Z.S. Qin, T. Niu, J.S. Liu, "Partitioning-ligation-expectation-maximization Algorithm for Haplotype Inference with Single-nucleotide Ploymorphisms," *The American Journal of Human Genetics*, Vol. 71, No. 5, pp. 1242-1247, 2002.

[26] M. Stephens and P. Donnelly, "A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data," *The American Journal of Human Genetics*, Vol. 73, No. 5, pp. 1162-1169, 2003.

[27] B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, "A Map of Recent Positive Selection in the Human Genome," *Public Library of Science Biology*, Vol. 4, No. 3, pp. 446-458, 2006.

[28] P. I. W de Bakker, G. McVean, P. C. Sabeti, M. M. Miretti, T. Green, J. Marchini, X. Ke, A. J Monsuur, P. Whittaker, M. Delgado, J. Morrison, A. Richardson, E. C Walsh, X. Gao, L. Galver, J. Hart, D. A Hafler, M. Pericak-Vance, J. A Todd, M. J Daly, J. Trowsdale, C. Wijmenga, T. J Vyse, S. Beck, S. S. Murray, M. Carrington, S. Gregory, P. Deloukas, J. D Rioux, "A High-resolution HLA and SNP Haplotype Map for Disease Association Studies in the Extended Human MHC," *Nature Genetics*, Vol. 36, pp. 1166-1172, 2006.

[29] K.M. Chao and L. Zhang, *Sequence Comparison: Theory and Methods*, Springer, 2009.

[30] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, WH Freeman and Company, 1979.

[31] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, The MIT Press, 2001.

[32] D. S. Houchbaum, "Approximation Algorithms for the Set Covering and Vertex Cover Problems," *SIAM Journal on Computing*, Vol. 11, No. 3, pp. 555-556, 1982.

[33] S. I. Ao, K. Yip, M. Ng, D. Cheung, P.Y. Fong, I. Melhado, P. C. Sham, "CLUSTAG: Hierarchical Clustering and Graph Methods for Selecting Tag SNPs," *Bioinformatics*, Vol. 21, No. 8, pp. 1735-1736, 2005.

[34] C. H. Papadimitriou and M. Yannakakis, "Optimization, Approximation, and Complexity Classes," *Journal of Computer and System Sciences*, Vol. 43, No. 3, pp. 425-440, 1991.