Block Partition and Tag Selection in Human SNP Haplotypes

Yaw-Ling Lin*, Guan-Jie Hua, and Wen-Pei Chen

Department of Computer Science and Information Engineering,
Providence University,
Taichung 433, Taiwan, ROC
yllin@pu.edu.tw, gt758215@gmail.com, g9471023@pu.edu.tw

Received 18 June 2010; Revised 19 August 2010; Accepted 20 September 2010

Abstract. Recent studies show that the patterns of linkage disequilibrium (LD) observed in human chromosome reveal a block-like structure; the high LD regions are called haplotype blocks. The existence of haplotype block structures has serious implications for association-based methods in mapping of disease genes. A Single Nucleotide Polymorphism or SNP is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of species. In this paper, we propose several efficient algorithms for identifying haplotype blocks in the genome. Especially, we develop a dynamic programming algorithm for haplotype block partitioning to minimize the number of tagSNPs required to account for most of the common haplotypes in each block. We implement these algorithms and analyze the chromosome 21 haplotype data given by Patil et al. [1]. As a result, we identify a total of 2,432 blocks (3,333 tagSNPs) which is 41.2% (27%) smaller than those identified by Patil et al. or Zhang et al. [2].

Keywords: Diversity, dynamic programming, SNP, haplotype block, tagSNP, haplotype block partition

References

[1] 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*, Vol. 294, No. 5547, pp. 1719-1723, 2001.

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

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

[4] J.D. Rioux, M.J. Daly, M.S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E.J. Kulbokas, S. O'Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S.B. Bull, R.S. McLeod, A.M. Griffiths, A. Bitton, G.R. Greenberg, E.S. Lander, K.A. Siminovitch, T.J. Hudson, "Genetic Variation in the 5q31 Cytokine Gene Cluster Confers Susceptibility to Crohn Disease," *Nature Genetics*, Vol. 29, No. 2, pp. 223-228, 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*, Vol. 296, No. 5576, pp. 2225-2229, 2002.

[6] G.C.L. Johnson, L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G.D. Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C.J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto,

* Correspondence author

S.C.L. Gough, D.G. Clayton, J.A. Todd, “Haplotype Tagging for the Identification of Common Disease Genes,” *Nat Genet*, Vol. 29, No. 2, pp. 233-237, 2001.

[7] J.D. Wall and J.K Pritchard, “Haplotype Blocks and Linkage Disequilibrium in the Human Genome,” *Nature Reviews Genetics*, Vol. 4, No. 8, pp. 587-597, 2003.

[8] R. R. Hudson and N. L. Kaplan, “Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences,” *Genetics*, Vol. 111, No. 1, pp. 147-164, 1985.

[9] N.Wang, J.M. Akey, K. Zhang, R. Chakraborty, L. Jin., “Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: the Interplay of Population History, Recombination, and Mutation,” *Am. J. Human Genetics*, Vol. 71, No. 5, pp. 1227-1234, 2002.

[10] E.C. Anderson and J. Novembre, “Finding Haplotype Block Boundaries by Using the Minimum-description-length principle,” *Am. J. of Human Genetics*, Vol. 73, No. 2, pp. :336-354, 2003.

[11] G. Greenspan and D. Geiger, “Model-based Inference of Haplotype Block Variation,” *Journal of computational biology*, Vol. 11, No. 2, pp. 493-504, 2004.

[12] M. Koivisto, M. Perola, R. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, E. Ukkonen, H. Mannila, “An MDL Method for Finding Haplotype Blocks and for Estimating the Strength of Haplotype Block Boundaries,” *8th Pacific Symposium on Biocomputing*, pp. 502-513, 2003.

[13] D. Clayton, “Choosing a Set of Haplotype Tagging SNPs from a Larger Set of Diallelic Loci,” *Nature Genetics*, Vol. 29, No. 2, 2001.

[14] K. Zhang, Z.S. Qin, J.S. Liu, T. Chen T, M.S. Waterman, F. Sun, “Haplotype Block Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies,” *Genome Research*, Vol. 14, No. 5, pp. 908-916, 2004.

[15] W.H. Li and D. Graur, *Fundamentals of Molecular Evolution*, Sinauer Associates, Inc, 1991.

[16] Y.L. Lin and W.S. Su, “Identifying Long Haplotype Blocks with Low Diversity,” *Proceedings of the 23rd Workshop on Combinatorial Mathematics and Computation Theory*, pp. 151-159, 2006.

[17] W.P. Chen, T.C. Lee, Y.L. Lin, “Haplotype Block Partitioning and TagSNP Selection on Human Chromosome 21,” *Proceedings of the International Computer Symposium 2006*, pp. 1278-1283, 2006.

[18] Providence University SNP and Haplotype Research Center. http://bioinfo.cs.pu.edu.tw/hap/.

[19] A.E. Darling, L.Carey, W.C. Feng, “The Design, Implementation, and Evaluation of mpiBLAST,” *Proceedings of ClusterWorld*, 2003.

[20] B. Halligan, J. Geiger, A. Vallejos, A. Greene, S. Twigger, “Low Cost, Scalable Proteomics Data Analysis Using Amazon's Cloud Computing Services and Open Source Search Algorithms,” *Journal of Proteome Research*, Vol. 8, No. 6, pp. 3148-3153, 2009.

[21] A. Matsunaga, M. Tsugawa, J. Fortes, Cloudblast: Combining Mapreduce and Virtualization on Distributed Resources for Bioinformatics Applications,” *Fourth IEEE International Conference on eScience*, pp. 222-229, 2008.

[22] M.C. Schatz, “Cloudburst: Highly Sensitive Read Mapping with Mapreduce,” *Bioinformatics (Oxford, England)*, Vol. 25, No. 11, pp.1363-1369, 2009.

[23] R. Buyya, C.S. Yeo, S. Venugopal, “Market-oriented Cloud computing: Vision, hype, and reality for delivering it services as computing utilities,” *Department of Computer Science and Software Engineering (CSSE), The University of Melbourne, Australia. He,* pp. 10-1016, 2008.

[24] B.F. Cooper, A. Silberstein, E. Tam, R., Sears, R. Benchmarking, “Cloud Serving Systems with YCSB,” *Proceedings of the 1st ACM symposium on Cloud computing*, pp. 143-154, 2010.

[25] P. Barham, B. Dragovic, K. Fraser, H. Steven, H. Tim, A. Ho, R. Neugebauer, I. Pratt, A. Warfield, “Xen and the Art of Virtualization,” *Symposium on Operating Systems Principles*, pp. 164-177, 2003.

[26] S. Hazelhurst, “Scientific Computing Using Virtual High-performance Computing: a Case Study Using the Amazon Elastic Computing Cloud,” *Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology***,** pp. 94-103, 2008.

[27] K. Keahey, I. Foster, T. Freeman, X. Zhang, “Virtual Workspaces: Achieving Quality of Service and Quality,” *Life in the Grid. Scientific Programming Journal*, Vol. 13, No. 4, pp. 265-276, 2005.

[28] Apache, Hadoop project, http://hadoop.apache.org/core/.

[29] T. White, Hadoop, *The Definitive Guide*, O'Reilly Media, 1 edition, 2009.

[30] J. Dean and S. Ghemawat, “Mapreduce: a Flexible Data Processing Tool,” *Communications of the ACM*, Vol. 53, No. 1, pp. 72-77, 2010.

[31] J. Dean, S. Ghemawat, Google Inc, “Mapreduce: Simplified Data Processing on Large Clusters,” *Proceedings of the 6th conference on Symposium on Opearting Systems Design and Implementation*, Vol. 51, No. 1, pp. 107-113, 2004.