HLATag: Selection of Tag SNPs for Capturing Human Leukocyte Antigens

Yao-Ting Huang[*1] and Chung-Chan Tsai[*]
[*]Department of Computer Science and Information Engineering
National Chung Cheng University, Taiwan.

Abstract

Human Leukocyte Antigen (HLA) genes are crucial determinant of transplant rejection and susceptibility to a variety of autoimmune-related diseases. However, large-scale and accurate typing of HLA alleles is laborious and expensive. In this paper, we aim to develop an algorithm for selecting a small subset of Single Nucleotide Polymorphisms (SNPs), called tag SNPs, which is able to capture HLA alleles. The HLA alleles can be indirectly predicted by alleles of these tag SNPs without performing direct HLA typing. The developed program is tested on the SNP and HLA map in four human populations. The experimental results indicate that the prediction accuracy of our tag SNPs is high in all populations. The HLATag program is freely available at http://www.cs.ccu.edu.tw/~ythuang/Tool/HLATag/.

Background

Major Histocompatibility Complex (MHC) is a genomic region located on chromosome 6 containing Human Leukocyte Antigen (HLA) class I and class II genes. These HLA genes encode antigen-presenting proteins which are essential in adaptive immune response. The class I HLA antigens (e.g., HLA-A and HLA-B) produce peptides on the cell-surface, whereas class II HLA antigens (e.g., HLA-DQA and HLA-DQB) present phagocytosed antigens from outside of the cell to T-lymphocytes. Each individual expresses different HLA class I and class II proteins for activating the process of the immune response and pathogen clearance. The HLA class I and class II genes are highly polymorphic among human populations (Hertz and Yanover, 2007). For example, more than one hundred alleles are found on the HLA-A gene. Thus, the chance of two unrelated individuals having identical HLA alleles is relatively low.

The determination of HLA alleles of an individual (called HLA typing) is important in clinical immunology research. In order to reduce the risk of transplant rejection, accurate HLA typing is curial in modern transplantation medicine (Gourraud *et al*., 2005). Furthermore, the variations of HLA alleles have been shown to correlate with a variety of autoimmune-related diseases (e.g., Hepatis C Virus) (Gaudieri *et al*., 2007; Thorsby and Lie, 2007; Timm *et al*., 2007). In practice, HLA typing is often done by sequence-specific oligonucleotide probes (SSOP) or PCR amplification, making large-scale HLA typing laborious and expensive. In addition, if the differences of HLA alleles fall outside the typed region, these methods may output inaccurate results (Listgarten *et al*., 2008). In addition, if an individual is heterozygous at one locus (i.e., possess two distinct HLA alleles), the allele information tends to be mixed up at that locus. As a consequence, efficient and accurate methods for typing HLA alleles are still highly demanded.

Recently, a few studies reported that HLA alleles are often highly correlated with alleles of flanking Single Nucleotide Polymorphisms (SNPs) (de Bakker *et al*. 2006; Malkki *et al*., 2005). SNP is a sort of nucleotide substitution at one locus which is kept through the heredity. A set of linked SNP alleles on the same chromosome is called a *haplotype*. In recent years, the advent of high-throughput genotyping technologies greatly reduces the cost of genotyping thousands of SNPs (Altshuler *et al*., 2005;Zhang *et al*., 2003, 2004). Through analysis of Linkage Disequilibrium (LD) across the entire MHC region, de Bakker *et al*. (2006) showed that HLA alleles tend to exhibit strong LD with flanking SNP alleles. Therefore, a small subset of SNPs (called tag

[1] Corresponding author: ythuang@cs.ccu.edu.tw

SNPs) is sufficient to capture partial HLA alleles. Consequently, some HLA alleles can be indirectly determined by alleles of tag SNPs using cost-effective SNP genotyping, instead of using laborious HLA typing. Figure 1(A) illustrates a high LD example of HLA alleles and surrounding SNPs. The alleles A and C at SNP_2 are perfectly correlated with HLA-C 1801 and 0702 alleles, respectively. Therefore, SNP_2 can be the tag SNP for HLA-C. However, the majority of HLA genes have more than two alleles, which is impossible to be tagged using only one SNP (see Figure 1(B)). Moreover, because of chromosome recombination, the same HLA allele may lie on multiple haplotype backgrounds (see Figure 1(C)). As a consequence, the selection of appropriate tag SNPs for capturing all HLA alleles is challenging.

In this paper, we design and implement algorithms for the selection of tag SNPs and for the prediction of HLA alleles. We use a two-stage approach for selecting tag SNPs, which are able to capture distinct haplotype backgrounds of each HLA allele. The developed program is tested on a variety of real data sets. The experimental results indicate that our program has high prediction accuracy in many data sets. We observed the majority of tag SNPs is specific to each population, indicating many HLA alleles lie on different haplotype backgrounds in distinct populations.

Implementation

Samples and Data Collection

The samples in this experiment are downloaded from the SNP and HLA map created by deBakker *et al.* (2006). This data set contains 90 African (YRI), 90 European descendants (CEU), 45 Chinese (HCB) and 44 Japanese (JPT), corresponding to the samples used in the international HapMap project (Altshuler *et al.*, 2005). There are 7543 SNPs and dele-tion-insertion polymorphisms were genotyped and 5754 passed the quality control. The al-leles of class I HLA genes (HLA-A, HLA-B, and HLA-C) and class II HLA genes (HLA-DQA, HLA-DQB, and HLA-DRB) are typed by PCR-SSOP.

In this section, we describe our algorithms for the selection of tag SNPs and for the prediction of HLA alleles. Our selection algorithm for tag SNPs consists of two stages. The first stage aims to select a minimum set of tag SNPs for distinguishing all distinct HLA alleles. The second stage aims to select a set of tag SNPs for capturing distinct haplotypes sat by one HLA allele. The SNP and HLA data are obtained from deBakker *et al.* (2006). Details of their data set are described in the next section. We retrieve the SNPs locating within the extended 100Kb regions of each HLA gene. In the first stage of our algorithm, we remove all SNPs having more than one allele mapping to the same HLA allele. For the example in Figure 1(B), SNP_1 will be removed, since it contains two alleles C and T mapping to HLA 1801 allele. These removed SNPs are either noise or indicative of HLA alleles lying on multiple haplotype backgrounds, which will be processed in the second stage of our algorithm. The prediction algorithm is used for predicting the HLA allele of one haplotype using the tag SNP alleles on that haplotype.

Stage I: Selection of Tag SNPs for Distinguishing Distinct HLA Alleles

We reformulate the problem of selecting tag SNPs into variant of the Set Covering (SC) Problem. Given a set of elements *E* and a collection of subsets *C* over *E*, the SC problem asks for a minimum subcollection *C'* of *C*, which includes (covers) all elements in *E*. We first reformulate each pair of HLA alleles as elements and map each SNP as a subset. For the example in Figure 1(B), there are three elements (1801,0701), (1801,0702) and (0701,0702), and three subsets corresponding to the three SNPs. For each SNP, the corresponding subset will contain elements that can be distinguished by this SNP. For example, SNP_2 in Figure 1(B) can distinguish two allele pairs: 1801/0702 and 0701/0702. Therefore, the subset corresponding to SNP_2 will contain two elements (1801,0702) and (0701,0702). Note that the subsets corresponding to SNP_2 and SNP_3 covers all the three elements. In other words, the combination of SNP_2 and SNP_3 is sufficient to distinguish all pairs of HLA alleles and thus can be the tag SNPs.

However, the SC problem is known to be NP-hard, implying no polynomial time algorithms are found so far. Hence, we use a greedy approximation algorithm which selects a SNP that distinguishes most pairs of HLA alleles at a time and repeats this selection process until no other pairs of HLA alleles can be distinguished (Huang *et al.*, 2005). If all allele pairs are distinguished, the set of selected SNPs is outputted as the solution. Otherwise, those undistinguished allele pairs are processed in the second stage.

Stage II: Selection of tag SNPs for HLA Alleles on Multiple Haplotype Backgrounds

After the first stage, there could be many pairs of HLA alleles still not distinguished. This is because these HLA alleles lie on multiple haplotypes and SNPs associated with these haplotypes are excluded in the first stage. In the second stage, we retrieve these excluded SNPs and consider those undistinguished HLA allele pairs. Note that these removed SNPs have multiple alleles mapping to some of these undistinguished HLA alleles. We reformulate this relationship into another instance of the SC problem. The element set E of the SC problem are pairs of haplotypes having different HLA alleles, and the subsets for covering elements are the unused SNPs. For the example in Figure 2, there are six haplotypes carrying two distinct HLA alleles. The reformulated SC problem contains nine elements (e.g., (h_1, h_4) and (h_2, h_4)) and two subsets SNP_1 and SNP_2. For each SNP, the corresponding subset will contain elements that can be distinguished by this SNP. For example, SNP_2 can distinguish six pairs of haplotypes (e.g., (h_2, h_4) and (h_3, h_6)). Similarly, if all pairs of haplotypes (having different HLA alleles) are distinguished by a set of SNPs (e.g., combination of SNP_1 and SNP_2), the haplotypes defined by these SNPs will represent the distinct haplotype background sat by each HLA allele.

Consequently, we use a similar greedy algorithm which selects a SNP distinguishing most pairs of haplotypes at one time and repeat this process until all pairs of haplotypes are distinguished. Following the same example, SNP_2 and SNP_3 are the tag SNPs defining distinct haplotypes for each HLA allele. Combining the tag SNPs selected in the first and second stages, these SNPs are the tag SNPs picked by our algorithm.

Algorithm for Predicting HLA Alleles

Given the alleles of selected tag SNPs on one individual, the prediction algorithm is used for predicting the HLA allele carried by the individual. Our prediction algorithm consists of two phases. The first phase used tag SNPs selected in the first stage and the second phase used the tag SNPs in the second stage. The tag SNPs selected in the first phase are used for identifying the set of possible HLA alleles. If there is only one HLA allele remained, this individual is predicted to carry that allele. Otherwise, we will use the tag SNPs selected in the second stage and compare the haplotype of this individual with those of all HLA alleles. The most similar one is outputted as the predicted HLA allele.

Results and Discussion

We download the database of SNP and HLA map from deBakker *et al*. (2006). This data set contains 90 African (YRI), 90 European descendants (CEU), 45 Chinese (HCB) and 44 Japanese (JPT), corresponding to the samples used in the international HapMap project (Altshuler *et al*., 2005). There are 7543 SNPs genotyped. The alleles of class I HLA genes (HLA-A, HLA-B, and HLA-C) and class II HLA genes (HLA-DQA, HLA-DQB, and HLA-DRB) are typed by PCR-SSOP. For each HLA gene, we retrieve the SNPs within the 100kb window centered on each gene. The tag SNP selection algorithm is run on each of the class I and class II HLA genes, separately for each of the four populations. Table 1 summarizes the number of tag SNPs found on each HLA gene for each population. In general, the number of tag SNPs is proportional to the ratio of distinct haplotypes to distinct HLA alleles at the gene locus within the population. For example, HLA-DRB is the most polymorphic gene with 40 distinct alleles and thus requires more tag SNPs than other genes in class II.

The prediction accuracy of our tag SNPs are evaluated using a leave-one-out cross-validation. That is, each haplotype is removed alternately and the other haplotypes are used for selecting tag SNPs. These tag SNPs are then use for predicting the HLA allele of the removed haplotype. In this experiment, we remove singleton HLA alleles and only consider alleles typed with four digit resolution (e.g., HLA-A 0301). The prediction accuracy of our tag SNPs are shown in Table 2. The experimental results indicate that the prediction accuracy are very high in class I HLA genes in all populations. The average prediction accuracy of class I HLA genes in all populations is about 95%. On the other hand, the average prediction accuracy in class II HLA genes is relatively low (88%). However, some of these class II HLA genes still achieve near 100% accuracy. In particular, the HLA-B and HLA-DRB have the lowest accuracy in class I and class II, respectively. This is because these two genes have the largest number of alleles in each class (e.g., 29 alleles in HLA-B in YRI), but the sample size in our training data set is insufficient. We also observe that when mixing Chinese and Japanese samples as one population, the accuracies (average 86.87% in class I and 80.57 in class II) are lower than separate results of each population. Thus, this phenomenon implies that the same HLA alleles of Chinese and Japanese populations lie on distinct haplotype background with different phylogenetic history.

Conclusion

This paper presented algorithms for the selection of tag SNPs which are able to capture untyped HLA alleles. The HLA alleles can be indirectly predicted by our prediction algorithm using alleles at these tag SNPs. The developed program was tested on a number of real data sets. The experimental results indicated that the prediction accuracy of our tag SNPs is high in all HLA genes and populations. We observed the majority of tag SNPs is specific to each population, indicating many HLA alleles lie on different haplotype backgrounds in distinct populations.

Acknowledgement

Yao-Ting Huang was supported in part by NSC grant 97-2221-E-194-051 from the National Science Council, Taiwan.

Availability and requirements

Project name: HLATag
Project homepage: http://www.cs.ccu.edu.tw/~ythuang/Tool/HLATag/.
Operating system: Linux, Windows
Programming language: C

Authors' contributions

YTH designed the algorithm and wrote the manuscript. CCT implemented and conducted the experiments. Both authors approved this manuscript.

References

1. Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., and Donnelly, P. A haplotype map of the human genome. *Nature*, 437:1299-1320, 2005.
2. Bafna, V., Halldorsson, B.V., Schwartz, R., Clark, A.G., and Istrail, S. Haplotypes and informative SNP selection algorithms: don't block out information. *In Proc. RECOMB'03*, pages 19-27, 2003.
3. Barrett J.C., Fry B., Maller J., and Daly M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263-265, 2005.
4. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium, *Am. J. Hum. Genet*., 74:106-120, 2004.
5. Chang, C.-J., Huang, Y.-T., and Chao, K.-M. A greedier approach for finding tag SNPs. *Bioinformatics*, 22: 685-691, 2006.
6. de Bakker, P.I., Yelensky, R., Pe’er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. Efficiency and power in genetic association studies. *Nat. Genet*., pages 1217-1223, 2005.
7. de Bakker, P.I., McVean G., Sabeti, P.C., *et al*. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics*, 2006.
8. Gaudieri, S., Rauch, A., Park, L., Freitas, E., Herrmann, S., *et al*. Erratum: Evidence of viral adaptation to HLA class I-restricted immune pressure in chronic hepatitis C virus infection. *J Virol*, 81: 8846-8848, 2007.
9. Gourraud P, Lamiraux P, El-Kadhi M, Raffoux C, Cambon-Thomsen A. Inferred HLA haplotype information for donors from hematopoietic stem cells donor registries. *Human Immunology*, 66: 563-570, 2005.
10. Hertz T, Yanover, C.Identifying HLA supertypes by learning distance functions. *Bioinformatics*, 23: e148-155, 2007.
11. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307:1072-1079, 2005.
12. Huang, Y.-T., Zhang, K., Chen, T., and Chao, K.-M. Selecting additional tag SNPs forh tolerating missing data in genotyping. *BMC Bioinformatics*, 6:263, 2005.

13. Huang Y.-T. and Chao, K.-M. A new framework for the selection of tag SNPs using multimarker haplotypes. *Journal of Biomedical Informatics*, 41:953-961, 2008
14. Leslie, S., Donnelly, P., and McVean G. A Statistical Method for Predicting Classical HLA Alleles from SNP Data. *American Journal of Human Genetics*, 2008.
15. Listgarten, J., Brumme, Z., Kadie, C., Xiaojiang, G., Walker, B. *et al*. Statistical resolution of ambiguous HLA typing data. *PLOS Computational Biology*, 4(2): e1000016, 2008.
16. Malkki, M., Single, R., Carrington, M., Thomson, G., and Petersdorf, E. MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: Implications for unrelated donor hematopoietic transplantation and disease association studies. *Tissue Antigens*, 66, 114-124, 2005.
17. Thorsby, E and Lie, B.A. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. *Transplant Immunology*, 14: 175-182, 2005.
18. Timm, J., Li B., Daniels, M.G., Bhattacharya, T., Reyor, L., *et al*. Human leukocyte antigen-associated sequence polymorphisms in hepatitis c virus reveal reproducible immune responses and constraints on viral evolution. *Hepatology*, 46: 339-349, 2007.
19. Zhang, K., Sun, F., Waterman, M.S., and Chen, T. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *American Journal of Human Genetics*, 73:63-73, 2003.
20. Zhang, K., Qin, Z.S., Liu, J.S., Chen, T., Waterman, M.S., and Sun, F. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Research*, 14(5):908-916, 2004.

Figures

Figure 1 – Examples of LD among tag SNPs and HLA alleles.

(A) SNP_2 can be the tag SNP for capturing alleles at HLA-C.; (B) There are three alleles at the HLA-C locus. The combination of SNP_2 and SNP_3 are predictive of alleles at HLA-C; (C) The allele 1801 at HLA-C lies on multiple haplotype backgrounds.

Figure 2 – Examples of reformulation into a set cover instance

The reformulated SC problem for one pair of undistinguished HLA alleles. There are six haplotypes (h_1, …, h_6) carrying two distinct HLA alleles and create a set of nine elements E to be covered. The combination of SNP_2 and SNP_3 can still be the tag SNPs distinguishing these two HLA allele classes.

Additional Files

Additional File 1 – The HLATag source code and program

The HLA.zip is compressed using WinZip and contains the C code and the binary code compiled on Linux platform.

Tables

Table 1. The number of tag SNPs selected for each HLA gene in each population. The total number of tag SNPs are the summation of numbers in the four populations. The shared number of tag SNPs is the number of tag SNPs included in all populations.

	YRI	CEU	HCB	JPT	Total
HLA-A	160	47	68	33	195
HLA-B	24	85	41	97	126
HLA-C	19	29	27	9	58
HLA-DQA	5	4	28	19	30
HLA-DQB	10	8	11	6	13
HLA-DRB	39	41	37	56	68

Table 2. Prediction accuracy (%) of our tag SNPs in leave-one-out cross-validation. The average accuracy in each row is based on the four populations.

	YRI	CEU	HCB	JPT	Average
HLA-A	94	97	97.8	95.2	96
HLA-B	89.8	95.6	83.5	96.3	91.3
HLA-C	98.3	94.9	100	98.8	98
Average	94.03	95.83	93.77	96.77	95.1
HLA-DQA	97.5	96.7	76.1	93	90.83
HLA-DQB	98.3	99.4	92	74.7	91.1
HLA-DRB	77.6	93.1	76.2	82.1	82.25
Average	91.13	96.4	81.43	83.27	88.06