Published in final edited form as:

Genet Epidemiol. 2011 February ; 35(2): 85–92. doi:10.1002/gepi.20549.

Predicting Multiallelic Genes Using Unphased and Flanking Single Nucleotide Polymorphisms

Sue Li¹, Hongwei Wang¹, Anajane Smith², Bo Zhang¹, Xinyi (Cindy) Zhang¹, Gary Schoch², Daniel Geraghty², John A. Hansen^{2,3}, and Lue Ping Zhao^{1,*}

¹ Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

² Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

³ School of Medicine, University of Washington, Seattle, WA 98195, USA

Abstract

Recent advances in genotyping technologies have enabled genomewide association studies (GWAS) of many complex traits including autoimmune disease, infectious disease, cancer and heart disease. To facilitate interpretations and establish biological basis, it could be advantageous to identify alleles of functional genes, beyond just single nucleotide polymorphisms (SNPs) within or nearby genes. Leslie et al (2008) have proposed an Identity-by-Decent method (IBD-based) for predicting human leukocyte antigen (HLA) alleles (multiallelic and highly polymorphic) with SNP data, and predictions have achieved a satisfactory accuracy on the order of 97%. Building upon their success, we introduce a complementary method for predicting highly polymorphic alleles using *unphased SNP data* as the training data set. Due to its generality and flexibility, the new method is readily applicable to large population studies. Applying it to HLA genes in a cohort of 630 healthy individuals as a training set, we constructed predictive models for HLA-A, B, C, DRB1 and DQB1. Then, we performed a validation study with another cohort of 630 healthy individuals, and the predictive models achieved predictive accuracies for HLA alleles defined at intermediate or high resolution ranging as high as (100%, 97%) for HLA-A, (98%, 96%) for B, (98%, 98%) for C, (97%, 96%) for DRB1 and (98%, 95%) for DQB1, respectively. These preliminary results suggest the feasibility of predicting other polymorphic genetic alleles, since HLA loci are almost certainly among most polymorphic genes.

Keywords

GWAS; HCT; HLA; MHC; microsatellite; gene prediction; SNPs

Introduction

Recent advances in array-based Single Nucleotide Polymorphisms (SNPs) genotyping technologies are transforming the landscape of genetic research. It becomes increasingly cost-efficient to genotype hundreds of thousands of SNPs for the whole genome in large population studies. To an extent, multiple SNPs cover or flank many functionally significant genes, such as human leukocyte antigens (HLAs) within the human Major

*For correspondence, contact: Dr. Lue Ping Zhao, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109-1024, Phone: (206) 667-6927 Fax: (206) 667-2437, lzhaol@fhcrc.org.

Histocompatibility Complex (MHC). Unlike biallelic SNPs, a typical HLA locus has multiple alleles and is extremely polymorphic with many alleles. To minimize the terminology confusion, we refer to such a gene locus as a multi-allelic gene (MAG), to differentiate it from bi-allelic SNPs. Through decades of research, many of these HLA alleles are known to have specific immunological functions. Hence, linking SNPs with HLA alleles would enable us to gain further insight into positive discovery with SNPs with supporting evidence for functional validation.

Towards this goal, Leslie et al (2008) recently developed a statistical method for predicting HLA alleles from SNPs, based upon an IBD-based genetic model using phase-resolved genotype data, i.e., *fully phased* haplotype data with multiple SNPs [Leslie, et al. 2008]. To illustrate their method, they used approximately 300 chromosomes as a training set, and built predictive models for HLA alleles, HLA-A, B, C, DRB1 and DQB1 using SNP data from the Affymetrix 500K GeneChip and the Illumina human NS-12 nonsynonymous SNP genotyping beadchip. Validation results on an independent set of samples ranged from 83% to 97%, a remarkable success (see Discussion). Their results have an important implication to ongoing genetic studies. For example, several recent GWAS, focusing on immune mediated diseases, have found very strong associations with the MHC region [2007; Asano, et al. 2009; Hirschfield, et al. 2009; Larsen and Alper 2004; Reveille, et al.; Sabeti, et al. 2007; Stefansson, et al. 2009; Tse, et al. 2009]. To assist with the interpretation of SNP associations and also to pinpoint specific antigens, it would be of great importance to infer the corresponding HLA alleles and to establish their disease associations.

Realizing the potential value of predicting MAG alleles from SNP data, we have developed a complementary method for building predictive models, using unphased SNP data as the training data set. Thus, one can readily gather a much larger number of samples, covering many less common MAG alleles. To illustrate the method, we build predictive models for HLA-A, B, C, DRB1 and DQB1 alleles, with a training data set of 630 healthy unrelated individuals, and then validate the predictive model on a separate set of 630 healthy unrelated individuals. Further, we compare prediction accuracies of our predictive models with those of IBD-based method. We conclude by discussing the strengths and limitations of the proposed method, as well as future improvements.

Materials and Methods

Study population

We have recently assembled a cohort of ~1,500 patient-donor pairs for a GWAS of hematopoietic cell transplant (HCT) outcome. The normal healthy donor set includes 1,260 Caucasians. The entire GWAS cohort was genotyped using Affymetrix 5.0 human GeneChip.

MHC SNPs and HLA genotyping

The Affymetrix 5.0 array includes 1,273 SNPs located within the extended human MHC region spanning approximately 6MB from position 28 to 34MB on chromosome 6p [de Bakker, et al. 2006]. The intra-MHC region is bounded by the HLA-A locus at the centromeric end and the HLA-DP locus at the telomeric end, and includes the class I HLA-A, B, C genes, and the class II DRB1 and DQB1 genes [de Bakker, et al. 2006]. Historically, HLA antigens were typed by serology. The introduction of DNA-based HLA typing, however, revealed that serology was capable of only an intermediate level of resolution. The majority of the classical HLA antigens, e.g., HLA-A2, represent distinct families of alleles that share one or more serologically-defined epitopes. The classification and identification

of HLA alleles in this study follows the recently published guidelines of the WHO HLA Nomenclature Committee (<http://hla.alleles.org/>).

Genotyping for the class I HLA-A, B and C genes was performed by sequencing exon 2 and 3, and genotyping for the class II DRB1 and DQB1 genes was performed by a combination of SSOP typing and sequencing of exon 2.

A Likelihood Model

Consider a random sample of N subjects. On each subject, we genotype an HLA gene, denoted by $h_i = \dot{h}_i \ddot{h}_i$, where the subscript $i (= 1, 2, \dots, N)$ denotes the i th subject, each $\dot{h}_i$ or $\ddot{h}_i$ is an HLA allele, taking one of the categorical values, such as HLA-A*01, A*02, A*03, A*11, etc. (see Supplementary Table S1 for more alleles at the intermediate resolution) or HLA-A*0101, A*0201, A*0202, etc. (see Supplementary Table S1 for more alleles at the high resolution). Suppose that on each i th subject, we have m flanking SNPs $g_i = (\dot{g}_{i1} \ddot{g}_{i1}, \dot{g}_{i2} \ddot{g}_{i2} \dots \dot{g}_{im} \ddot{g}_{im})$, where each $\dot{g}_{ij}$ or $\ddot{g}_{ij}$ takes one of two values as a single SNP allele at the j th locus. Given the expected local LD among the HLA gene and all SNPs, one naturally depicts their joint distribution via the following probability function:

$$f(h_i, g_i) = f(\dot{h}_i \ddot{h}_i, \dot{g}_{i1} \ddot{g}_{i1}, \dot{g}_{i2} \ddot{g}_{i2}, \dots, \dot{g}_{im} \ddot{g}_{im}) = \sum_{\Omega(h_i, g_i)} f(\dot{h}_i \dot{G}_i) f(\ddot{h}_i \ddot{G}_i), \quad [1]$$

where the set $\Omega(h_i, g_i)$ represents all haplotype pairs whose genotypes are consistent with those observed genotypes (h_i, g_i) , the summation is over all possible haplotypes given the observed genotypes (h_i, g_i) , $\dot{h}_i \dot{G}_i$ and $\ddot{h}_i \ddot{G}_i$ are two extended haplotypes of HLA gene and SNPs under the Hardy-Weinberg equilibrium (HWE), $(\dot{G}_i, \ddot{G}_i)$ represents two haplotypes of observed SNP alleles, and, finally, $f(\dot{h}_i \dot{G}_i)$ represents the probability of observing the haplotype $\dot{h}_i \dot{G}_i$. Without imposing any additional assumption, the haplotype probability distribution can be represented by a multinomial distribution, which may be written as:

$$f(\dot{h}_i \dot{G}_i) = \prod_{hG} \Pr(hG)^{I(\dot{h}_i \dot{G}_i = hG)}, \quad [2]$$

where $\Pr(hG)$ is the haplotype frequency for observing hG , and the indicator function $I(\dot{h}_i \dot{G}_i = hG)$ equals one if the inside equality is true.

Estimating the haplotype frequency $\Pr(hG)$ is of primary interest. To do so, we employ the likelihood method, via maximizing the following log likelihood function:

$$l = \sum_i \log \left(\sum_{\Omega(h_i, g_i)} \prod_{hG} \Pr(hG)^{I(\dot{h}_i \dot{G}_i = hG) + I(\ddot{h}_i \ddot{G}_i = hG)} \right), \quad [3]$$

where the first summation is over all N independent subjects. Given the double summations, one commonly used approach to maximize this log likelihood function [3] is via expectation-maximization (EM) algorithm EM [Excoffier and Slatkin 1995; Hawley and Kidd 1995]. However, when the number of possible haplotypes becomes fairly large, such as this case, the computational burden could be prohibitive.

To overcome the computational challenge, we adopt an estimating equation approach, which has been used by us and detailed elsewhere [Li, et al. 2003]. Briefly, we used the above log

likelihood function to derive its score estimating equation by the first derivative with respect to haplotype probabilities. Within the score estimating equation, we identified a covariance matrix component, which requires a substantial computation, following the likelihood calculation. However, this component could be modified, to simplify the computational burden, and such a modification does not invalidate the consistency of estimated haplotype probabilities nor negatively impacts the efficiency of the estimates. Thus modifying the score estimating equation leads to the formation of the estimating equation which gives rise to valid and efficient estimates of haplotype probabilities [Li, et al. 2007].

A Predictive Model

Our goal is to predict HLA alleles, given unphased SNP genotypes $(\dot{g}_1\dot{g}_1, \dot{g}_2\dot{g}_2, \dots, \dot{g}_m\dot{g}_m)$. Using the Bayes' theorem, we constructed the following probability calculation:

$$\Pr(\dot{h}\dot{h}|\dot{g}_1\dot{g}_1, \dot{g}_2\dot{g}_2, \dots, \dot{g}_m\dot{g}_m) = \frac{f(\dot{h}\dot{h}, g)}{\sum_{\dot{H}\dot{H}} f(\dot{H}\dot{H}, g)} = \frac{\sum_{\Omega(h, g)} \prod_{hG} \Pr(hG)^{I(\dot{h}\dot{G}=hG)+I(\dot{h}\dot{G}=\bar{h}\bar{G})}}{\sum_{\dot{H}\dot{H}} \sum_{\Omega(H, g)} \prod_{hG} \Pr(hG)^{I(\dot{H}\dot{G}=hG)+I(\dot{H}\dot{G}=\bar{h}\bar{G})}}, \quad [4]$$

where the first summation in the denominator is over all possible genotypes at the HLA locus $(\dot{G}\dot{G})$, the second summation is over all possible haplotype pairs that are consistent to observed genotype data, and $\Pr(hG)$ is the haplotype frequency that is estimated from the above likelihood function on the training data set.

The above predictive probability takes value from 0 to 1, for all possible HLA allelic pairs under prediction. There are two predictive strategies in using the above predictive probability [4]. The first strategy is to compute predictive probabilities for all possible HLA genotypes. The predicted HLA genotype is chosen to be the one that has the highest predictive probability, regardless of its value. This strategy would always lead to a HLA genotype prediction, i.e., the call rate is 100%. One potential downside to this strategy is that best predictive probability may still take a very small value, given a large number of alleles at HLA loci. An alternative predictive strategy is to require the best probability to exceed call threshold, say CT=0.5 or 0.9. For those predictions with unacceptable predictive probabilities, HLA genotype is left as an ambiguous prediction, i.e., the call rate may be less than 100%.

SNP Selections

To build a successful predictive model for a target gene, it is essential to select informative SNPs, while minimizing redundant SNPs in close proximity and also in high linkage-disequilibrium with each other. Secondly, one may want to include certain distant flanking SNPs that share ancestry haplotypes with certain HLA alleles. Thirdly, it is also important to realize that different SNP haplotype backgrounds may carry the same HLA alleles, as HLA genes may undergo faster evolutionary selections [Walsh, et al. 2003]. Lastly, following the parsimonious principle, the predictive model should have the fewest predictive SNPs possible, without sacrificing predictive accuracy. To take into account all of these considerations, we constructed an objective function, based upon the Akaike criterion (AIC) [Koehler and Murphree 1988], i.e., penalizing on the number of additional haplotype parameters to be estimated, when adding each SNP to the predictive model. The resulting objected function may be written as:

$$Q = - \sum_i \log \left(\frac{\sum_{\Omega(h_i, g_i)} \prod_{\underline{hG}} \Pr(\underline{hG})^{I(\underline{h_i}, \underline{G_i} = \underline{hG}) + I(\underline{\tilde{h_i}}, \underline{\tilde{G_i}} = \underline{hG})}}{\sum_{\underline{H}\underline{\tilde{H}}} \sum_{\Omega(\underline{H}, \underline{g_i})} \prod_{\underline{hG}} \Pr(\underline{hG})^{I(\underline{H}\underline{\tilde{G_i}} = \underline{hG}) + I(\underline{\tilde{H}}\underline{\tilde{G_i}} = \underline{hG})}} \right) + (m - k). \quad [5]$$

where the first logarithmic term in the above equation is the negative log likelihood of predictive probabilities given all SNP genotype data in the training set, and the second term equals the difference of the number of haplotypes of SNP–HLA and the number of HLA alleles (k). Note that the first term of the objective function is monotonically decreasing with ever increasing number of haplotype parameters to be estimated due to increasing numbers of SNPs. This function is penalized by the increasing number of haplotype parameters to be estimated.

As noted above, given the current chip design, we expect to include all available SNPs within genes as well as those SNPs in flanking regions. Since LDs of SNPs and their haplotypes with HLA alleles decrease with increasing distance, we propose to use a combined forward selection and backward elimination scheme, starting from the HLA locus and gradually adding one SNP at a time. Our selection scheme is summarized as follows:

1. Include all SNPs (S) within HLA locus (if there are no SNPs genotyped within the HLA locus, include three flanking SNPs on each side). Calculate the objective function in [5] and denote as Q^0 .
2. Perform a forward selection by adding the next adjacent SNP to S from the left and calculating the objective function, which is denoted as Q . If $Q < Q^0$, add this SNP to the set of SNPs. Set $Q^0 = Q$.
3. Perform a backward elimination: if a new SNP added into S in step 2, calculate the objective function $Q(s)$ by removing each SNP in S . If $\min(Q(s)) < Q^0$, remove the SNP s with the minimum $Q(s)$ from S and reset $Q^0 = \min(Q(s))$.
4. Repeat steps 2 and 3 for the next adjacent SNP to S from the right side of HLA locus.
5. Repeat steps 2–4 until the pre-set boundary on both sides is reached.
6. Determine the boundary via evaluating the objective functions with varying sizes of the flanking region, from 5KB to 400KB, and choose the boundary size such that the objective function approaches a “minimum” value in the objective function.

Missing SNP Genotype Data

Despite the ever-improving genotyping technologies, SNP genotype data may still be untyped for a proportion of individuals. When the percentage of missing data is large, say, over 5% across all samples per SNP or 5% across all SNPs per individual, the corresponding SNP or sample is best to be excluded from the training/validation sets, due to the quality concern. For the remaining SNPs or samples, we propose to generalize the likelihood method [3] and [5] by allowing missing genotype data. The missing mechanism is assumed to be missing at random [Efron 1994]. Under such an assumption, we can add another level summation over missing genotype, whenever a SNP genotype is missing.

Predictive Model Validation

Upon the completion of the model training process, we can assemble a list of selected SNPs for each HLA gene, together with their haplotypes ($\underline{hG}$) and frequencies $\Pr \underline{h}(\underline{G})$ as the

estimated haplotype distributions for the target population. Now to predict HLA allele pairs ($\hat{h}\hat{h}$) given an array of any genotypes at selected SNP loci ($\hat{g}_1\hat{g}_1, \hat{g}_2\hat{g}_2, \dots, \hat{g}_m\hat{g}_m$), one can compute the predictive probability via equation [4]. Once an HLA genotype is predicted, one can thus compare the predicted HLA genotype ($\hat{h}^p\hat{h}^p$) with the observed HLA genotype ($\hat{h}^o\hat{h}^o$). We propose an “accuracy” index, to measure the number of alleles that are correctly predicted, i.e., for an individual, the accuracy takes a value of 0, 0.5 and 1.0, for completely wrong prediction, correctly predicting one allele, and correctly predicting both alleles, respectively.

Results

We used 630 individuals in the training set to build HLA predictive models, and an independent set of 630 individuals for validating predictive models for the HLA-A, B, C, DRB1 and DQB1 genes. In the training set, high resolution allele level data for HLA-A, B, C and DRB1 is available for all 630 individuals and 622 individuals have high resolution data available for DQB1. In the validation set, high resolution allele level data for HLA-A, B, C and DRB1 is available for all 630 individuals and 578 individuals have high resolution allele level data available for DQB1. When building predictive models for intermediate level resolution, we collapsed the high resolution allele nomenclature to the intermediate resolution, i.e., antigen equivalency. Allelic frequencies for HLA-A, B, C, DRB1 and DQB1, in training and validation sets, are listed in the supplementary tables (intermediate resolution in Table S1 and high resolution in Table S2).

Using the training set, we built separate predictive models for five HLA genes (HLA-A, B, C, DRB and DQB1) using MHC SNPs obtained from the Affymetrix 5.0 array. The selected SNPs were located in the flanking regions of the HLA loci and none were intragenic. By the computational algorithm described above, we evaluated the objective functions for given flanking region sizes. Figure 1A shows that the objective function (y-axis) decreases as the size of flanking region increases, while building HLA-A, B, C, DRB1 and DQB1 predictive models. For predicting intermediate resolution of HLA-A, it appears that a flanking region of $\pm 250\text{KB}$ is sufficient. In total, 13 SNPs (covering 29,885,178-30,134,125 on chromosome 6) were identified. On the other hand, it seems that the comparable flanking region would be required for selecting informative SNPs to achieve high resolution predictions for HLA-A (Figure 1B). The actual SNP selection identifies 19 SNPs covering the region (29,885,178-30,134,125) (all SNPs are identified in supplementary Table S3). Similarly, flanking boundaries for HLA-B, C, DRB1 and DQB1 can be determined based upon the objective functions (Figure 1A and 1B). Following the same SNP selection procedure, we have selected informative SNPs for intermediate and high resolution predictions for the remaining HLA genes (see Table S3).

Table 1 displays computed accuracies in the training set for intermediate and high genotype resolution. At intermediate resolution, accuracy ranges from 95% (DRB1) to 100% (HLA-C), with CT=0. By increasing CT to 0.5 and to 0.9, accuracy levels approach 100%. The call rate, especially for DRB1, falls to 78%. Accuracy levels for predicting high resolution genotypes for HLA-A, B and C are on the order of 97% to 98% with CT=0. However, accuracies drop to 84% and 88% for DQB1 and DRB1, respectively. When CT is raised to 0.5 or 0.9, accuracies for predicting HLA-A, B and C improve towards 99%, at the expense of reducing call rates. For DRB1 and DQB1, however, accuracies are improved, but call rates are substantially reduced to 37% and 55% for DQB1 and DRB1, respectively. To gain insights into those poor predictions, we listed all of the incorrectly predicted alleles, and examined reasons for incorrect predictions (see Discussion).

To validate these predictive models, we next evaluated accuracies of predicted HLA alleles with an independent validation data set. For predicting genotypes at intermediate resolution, accuracies range from 93% to 98%, with CT=0 (Table 2). By raising CT to 0.5 or 0.9, accuracies are uniformly increased, at the expense of reduced call rates, resulting in accuracies ranging from 97% (HLA-C) to 100% (HLA-A). In contrast, with high resolution calls, patterns of accuracies largely remain, with minor drops in accuracies. Notably, accuracies for HLA-DRB1 and HLA-DQB1 are particularly problematic, with accuracies of 79% and 83%. Accuracies improve to 96% and 95%, once the CT is raised to 0.9.

Discussion

In this manuscript, we have described a general methodology for building predictive models for MAG alleles with multiple flanking SNP genotypes. The key idea underlying this methodology is to build extended haplotype structures of SNPs and target genes in the training set, and then select a minimal but informative set of SNPs for the polymorphic alleles of the gene. Based upon these results, one can compute predictive probabilities for all possible MAG alleles, given SNP genotype data. We have illustrated the model building exercise for predicting HLA alleles, and resulted predictive models have very respectable accuracies in the validation study.

As noted above, our intent of developing this methodology goes beyond its valuable application to predicting HLA genes. An immediate application is to predict microsatellite (MS) marker genotypes, from SNP genotypes. MS markers are traditionally used in genetic research, from linkage to association analysis. In comparison with SNPs, MS markers have multiple alleles, which are typically fewer than HLA genes, and are thus more informative for many genetic analyses. However, genotyping cost of MS markers is much higher than genotyping SNPs per locus. One possibility is to build a SNP-MS conversion table, so that SNP genotype data may be interpretable via MS markers. Another application of this methodology is to predict full sequence variants from SNPs. Despite the advances in sequencing technologies, the cost of sequencing genes or genetic regions on hundreds or thousands of subjects remains high for the next few years. One strategy moving forward is to utilize well-categorized sequence data, such as those from the 1000 Genome Project (<http://www.1000genomes.org>), as the training set, and to code pertinent sequence variants (nucleotide variants, insertions/deletions, structural variants) as different alleles. Based upon SNP data from HapMap project (<http://hapmap.ncbi.nlm.nih.gov>), one can train predictive models of interest.

Despite achieving prediction accuracies ranging between 95% and 98%, these predictions are not perfect. Here we discuss three factors that may contribute to these imperfect predictions. The first factor is associated with the uninformative prediction with more than one HLA allele on a SNP single haplotype, as noted above. For example, for HLA-A allele “25” (intermediate resolution), its 22 predictions are incorrect, since the predicted allele is the allele “26” (Table S1). Further investigation suggested that the haplotype of selected SNPs carries both allele “25” and “26”. Since the allele “26” has higher allelic frequency ($n=42$) than allele “25” ($n=22$) in the training set (Table S1), the predictive probability always favors the allele “26” (64%) as oppose to the allele “25” (36%). In HLA literature [Madrigal, et al. 1993], HLA-A25 and A26 share a common serological epitope designated HLA-A10. Prior phylogenetic analyses of the HLA-A locus based on allele sequence shows that A10 is one of the major branches of the HLA-A family tree. A10 is “old”, while A25 and A26 are “young”. Therefore, unless such “young SNPs” are included in our training set, it is unrealistic to expect correct predictions of less frequent allele “HLA-A*25”. In the absence of such “young SNPs” for practice, one may combine these two alleles into a composite allele (25/26). If such composite allele were introduced, the predictive accuracy

would be noticeably improved by a couple of percentage points. Similar observations and remedies can be made for several other alleles in HLA genes. To improve the resolution for those ambiguous alleles, it is prudent to introduce additional SNPs, especially those within genes.

The second factor that may contribute to imperfect predictions is the stringency in making the prediction. Consider the high resolution HLA-B predictions (Table 2). In the training set, the accuracy equals 97%, 97% and 99%, with CT=0, 0.5 and 0.9, respectively. By raising CT, the prediction accuracy improves. It becomes more obvious for the validation data set, where the accuracy improves from 93%, to 94% and to 96% with the increasing CT.

The third factor contributing to imperfect prediction is that some HLA alleles are rare, being observed less than three times in the training set. For example, for HLA-A locus, the allele "69" is completely absent in the training set, and is present twice in the validation data set. Given the rarity, it is probably unrealistic to expect accurate predictions in such cases, as noted as the second limit within the current prediction model (probably true for most prediction models). Again, if incorrect predictions on such alleles were excluded, the resulting accuracies of our predictions would be improved further.

In general, as the number of copies of each allele present in the training set increases, the accuracy of predicted alleles in the validation data set also increases, with a few exceptions. Specifically, accuracies for HLA-A, B and C at either intermediate or high resolution increase as the copy number of observed alleles increases (see Figure 2A and 2B). There are a few outliers, for example, the prediction accuracy equals zero for the outlier with 22 copies of the allele HLA-A*25, because all of them have been predicted to be HLA-A*26. Another example is HLA-B*39, which is observed 18 times, but a total of 8 samples are predicted to have HLA-B*38 (Table S4). Again, these two alleles (38 and 39) are rather close to each other, and one or more additional SNPs would be able to distinguish them, which would improve the accuracy. Similar phenomenon holds for many cases in high resolution (Figure 2B).

However, for DRB1 and DQB1, relationships between accuracies and allele numbers seem to be more complicated. For intermediate resolution, the prediction is relatively strong with accuracies hovering around 0.9 or better, with only one exception (in the middle) associated with the HLA-DRB1*11 (Table S4). Even with 174 copies observed in the training set, these alleles are often predicted to HAL-DRB1*01, *04, *07 and *03 in the validation data set. When dealing with high resolution, predictive accuracies on several relatively common alleles are variable. One possible explanation is that selected SNPs and their haplotypes do not tag these alleles well, or sequence variations are not tagged by available SNPs. Potentially, different genetic markers, such as copy number variations or copy number polymorphisms may be more informative for such predictions [McCarroll, et al. 2008]. Of course, given the historical nature of some DRB1 and DQB1 genotype data, transcriptional errors from clinical records to research database could be another contributing factor.

Before HLA predictive models can be generally useful, it is important to realize that the number of HLA alleles covered by our training and validation sets are much smaller than HLA alleles documented in IMGT (<http://www.ebi.ac.uk/imgt/hla/>) (see Table 3). Of course, without coverage of those alleles, predictions of corresponding HLA alleles are not realistic. To improve the general applicability of predictive models, it is essential to apply the current method to large population studies so that relatively uncommon alleles will be observed. Additionally, one could apply this method to population studies of different ethnic groups, which will cover many more HLA alleles unique to those populations. Preliminary results from investigating this issue indicate that inclusion of multiple ethnic groups in the training

set leads to predictive models of HLA alleles with much more favorable accuracies, in comparison with those ethnic-specific predictive models (not shown). Full results from exploring this issue and several other practical issues will be reported under a separate cover.

Another area for further improvement is to build predictive models with SNPs from SNP array technologies other than the Affymetrix 5.0 chip, which was used in constructing and validating of the predictive models for alleles encoded by five HLA genes. Affymetrix has produced several new chips, including SNP array 6.0 and Axiom (<http://www.affymetrix.com>). Illumina, on the other hand, has 660W, Omni1, 1M and customized MHC chips (<http://www.illumina.com/>). All of these chips have many more SNPs covering MHC, e.g., Axiom array includes 7,391 SNPs to cover MHC region.

Despite its generality, this method can be further extended to incorporate the partially phased genotype data informed by family data, such as trios, if such data are available. When working with highly polymorphic genes, one often has a collection of family data, so that rare variants can be unambiguously determined within families. With family data in place, one typically can obtain partial phase information between SNPs and target genes. In such cases, one could incorporate the partial phase information by modifying the log-likelihood function [3] and the objective function [5], to re-write the log-likelihood function to incorporate the family structures, in the same manner as in our earlier works [Zhao, et al. 2007].

We have implemented this new method in MATLAB, and a compiled version is available upon request. Also, to facilitate predictions of HLA genes from SNPs (for those using Affymetrix 5.0), we have combined the training and validation data sets, to select the “most informative SNPs” for five HLA genes from our cohort. Figure S1 shows the objective function with the window sizes, leading to the choice of flanking regions. Interestingly, the patterns of the objective functions closely resemble those in the training set (Figure 1), which indicates the relative robustness of choosing the flanking window sizes. Following the same selection procedure, we obtain a revised list of SNPs for HLA predictions (Table S6).

Our HLA predictive models are closely connected with those built by Leslie et al (2008), despite that the required training data, phased versus unphased, is different. From the perspective of actual practice in predicting HLA alleles, we compared accuracies of our final predictive models with estimated accuracies extracted from Leslie et al (2008), using the same validation cohort (Table 4). It appears that both call rates and accuracies from our predictive models outperform those from the Leslie et al (2008), with an exception for DRB1; the call rates at high resolution are lower for CT=0.5 and 0.9. Even for this exception, the accuracy of our model at CT=0 equals 85%, which is higher than 72% of the Leslie et al's model. Besides some technical differences, the main contributing factor for this superior performance is that we were able to train the models with a much larger sample size.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Authors would like to thank the many clinical investigators, research nurses, biostatisticians, and database specialists who have conducted this cohort study and have maintained this valuable data resource. Equally, we would like to acknowledge all of the donors whose genotype data are used in our analysis. We would also like to thank NIH for providing partial funding support to this project through the following grants: NIH/NHLBI

R01HL087690 (PI: John Hansen), NIH/NIMH R01MH084621 (PI: Lue Ping Zhao), NIH/NCI R01CA106320 (PI: Lue Ping Zhao), NIH/NCI R01CA119225 (PI: Lue Ping Zhao).

References

- Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447(7145):661–78. [PubMed: 17554300]
- Asano K, Matsushita T, Umeno J, Hosono N, Takahashi A, Kawaguchi T, Matsumoto T, Matsui T, Kakuta Y, Kinouchi Y, et al. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat Genet* 2009;41(12):1325–9. [PubMed: 19915573]
- de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 2006;38(10):1166–72. [PubMed: 16998491]
- Efron B. Missing data, imputation, and the bootstrap. *J Am Stat Assco* 1994;89(426):463–475.
- Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;12(5):921–7. [PubMed: 7476138]
- Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 1995;86(5):409–11. [PubMed: 7560877]
- Hirschfield GM, Liu X, Xu C, Lu Y, Xie G, Gu X, Walker EJ, Jing K, Juran BD, Mason AL, et al. Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. *N Engl J Med* 2009;360(24):2544–55. [PubMed: 19458352]
- Koehler AB, Murphree ES. A comparison of the akaike and schwarz criteria for selectin model order. *Appl Stat* 1988;37(2):187–195.
- Larsen CE, Alper CA. The genetics of HLA-associated disease. *Curr Opin Immunol* 2004;16(5):660–7. [PubMed: 15342014]
- Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. *Am J Hum Genet* 2008;82(1):48–56. [PubMed: 18179884]
- Li S, Khalid N, Carlson C, Zhao LP. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics* 2003;4(4):513–22. [PubMed: 14557108]
- Li SS, Cheng JJ, Zhao LP. Empirical vs Bayesian approach for estimating haplotypes from genotypes of unrelated individuals. *BMC Genet* 2007;8:2. [PubMed: 17261196]
- Madrigal JA, Hildebrand WH, Belich MP, Benjamin RJ, Little AM, Zemmour J, Ennis PD, Ward FE, Petzl-Erler ML, du Toit ED, et al. Structural diversity in the HLA-A10 family of alleles: correlations with serology. *Tissue Antigens* 1993;41(2):72–80. [PubMed: 8475492]
- McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 2008;40(10):1166–74. [PubMed: 18776908]
- Reveille JD, Sims AM, Danoy P, Evans DM, Leo P, Pointon JJ, Jin R, Zhou X, Bradbury LA, Appleton LH, et al. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat Genet* 2006;38(2):123–7. [PubMed: 16620622]
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007;449(7164):913–8. [PubMed: 17943131]
- Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, et al. Common variants conferring risk of schizophrenia. *Nature* 2009;460(7256):744–7. [PubMed: 19571808]
- Tse KP, Su WH, Chang KP, Tsang NM, Yu CJ, Tang P, See LC, Hsueh C, Yang ML, Hao SP, et al. Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. *Am J Hum Genet* 2009;85(2):194–203. [PubMed: 19664746]
- Walsh EC, Mather KA, Schaffner SF, Farwell L, Daly MJ, Patterson N, Cullen M, Carrington M, Bugawan TL, Erlich H, et al. An integrated haplotype map of the human major histocompatibility complex. *Am J Hum Genet* 2003;73(3):580–90. [PubMed: 12920676]

Zhao LP, Li SS, Shen F. A haplotype-linkage analysis method for estimating recombination rates using dense SNP trio data. *Genet Epidemiol* 2007;31(2):154–72. [PubMed: 17219374]

Figure 1.
The objective functions over the flanking window sizes, for HLA-A, B, C, DRB1 and DQB1, where the window size is sum of the flanking windows from both sides.

Figure 2.
Relationships of accuracies in validation predictions with frequencies of observed alleles in the training set.

Accuracy in predicting HLA alleles in the training data set (630 WGA Caucasian unrelated donor samples, with n=2*630 haplotypes. Among them 8 samples missing or having intermediate resolution DQB1 genotypes therefore excluded in DQB1 prediction)

Table 1

Locus	Intermediate resolution (%)			High resolution (%)		
	CT=0	CT=0.5	CT=0.9	CT=0	CT=0.5	CT=0.9
HLA-A	98	98(100)	100(89)	97	97(100)	99(88)
HLA-B	98	98(100)	100(91)	97	97(99)	99(79)
HLA-C	100	100(100)	100(97)	98	98(100)	99(95)
HLA-DRB1	95	95(99)	99(78)	84	88(83)	98(37)
HLA-DQB1	99	99(100)	100(98)	88	93(81)	99(55)

+: Call threshold

Accuracy in predicting HLA alleles in the validation data set (630 WGA Caucasian related donor samples, n=2*630 haplotypes. Among them 52 samples missing or having intermediate resolution DQB1 genotypes therefore excluded in DQB1 validation)

Table 2

Locus	Intermediate resolution (%)			High resolution (%)		
	CT=0	CT=0.5	CT=0.9	CT=0	CT=0.5	CT=0.9
HLA-A	97	97(100)	100(86)	95	95(100)	97(86)
HLA-B	96	96(99)	98(88)	93	94(97)	96(75)
HLA-C	98	98(100)	98(97)	97	97(99)	98(93)
HLA-DRB1	93	93(99)	97(78)	79	85(83)	96(39)
HLA-DQB1	97	97(99)	98(95)	83	86(86)	95(57)

+: Call threshold

Number of known HLA alleles, number of HLA alleles represented in the training and validation sets, as well as those reported by Leslie et al (2008).

Table 3

	A-locus	B-locus	C-locus	DRB1-locus	DQB1-locus
Number of known alleles ^a	965	1543	626	762	107
Number alleles (intermediate/high resolution) represented in					
Training set	17/30	28/60	13/33	13/31	5/18
Validation set	16/30	27/56	13/30	13/34	5/16
Training set used by Leslie et al (2008)	13/17	22/29	13/17	12/23	5/14

^aAccording to the IMGT-HLA Database, March 2010 release: <http://www.ebi.ac.uk/imgt/hla/>

Comparison of HLA allele prediction accuracy among the British 1958 Birth Cohort Data (genotyped by Affymetrix 500K), using predictive models produced by Leslie et al and by us

Table 4

Locus	Predictive model	Number of SNPs selected ^c	Number of Haplotypes ^d	Accuracy at High Resolution (%) and Call Rate (%)			Accuracy at Intermediate Resolution (%) and Call Rate (%)		
				CT=0	CT=0.5	CT=0.9	CT=0	CT=0.5	CT=0.9
HLA-A	Leslie et al. ^a	10/10	876/1792	89	91 (93)	97 (58)	93	94 (94)	95 (29)
	Li et al. ^b	16/15	1482/2936	96	96 (99)	98 (88)	97	97 (100)	99 (90)
HLA-B	Leslie et al.	40/40	1630/1708	82	85 (88)	93 (66)	84	87 (89)	94 (65)
	Li et al.	26/25	2582/2708	94	95 (98)	97 (82)	95	95 (100)	97 (89)
HLA_DRB1	Leslie et al.	34/34	834/1798	72	76 (88)	83 (51)	86	90 (88)	95 (55)
	Li et al.	15/14	1910/2950	85	90 (81)	99 (38)	97	98 (99)	98 (77)
HLA_DQB1	Leslie et al.	22/22	1088/1774	77	80 (88)	93 (29)	90	91 (89)	97 (31)
	Li et al.	16/12	1726/2942	94	95 (98)	98 (80)	99	99 (100)	99 (97)

^aIn Leslie et al., the training cohort are the parents (genotyped at over 7500 SNPs, [Leslie et al. 2008]) of the 30 CEU family in the HapMap Project and 15 additional families from the same population as the CEU

^bIn the current manuscript, the training set includes 1260 Caucasian donors (genotyped by Affymetrix 5.0) for patients who received HCT at the Fred Hutchinson Cancer Research Center

^cPredictive models are built separately at the required resolutions (high resolution/intermediate resolution)

^dAccuracy is only assessed at those individuals in which both alleles are typed at the required resolutions (four digit/two digit)