# *In Silico* Resolution of Ambiguous HLA Typing Data

Jennifer Listgarten,[1] * Zabrina Brumme,[2] Carl Kadie,[1] Gao Xiaojiang,[3] Bruce Walker,[2,4]
Mary Carrington,[3] Philip Goulder,[2,5] David Heckerman [1]*

[1]Microsoft Research, Redmond, WA 98052, USA, [2]Partners AIDS Research Center, Massachusetts, General Hospital, Harvard Medical School, Boston, MA 02129, USA, [3]Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederick, National Cancer Institute, Frederick, Maryland, USA, [4]Howard Hughes Medical Institute, Maryland, 6789 USA, [5]Department of Paediatrics, University of Oxford, Oxford OX1 3SY, UK

*Address email correspondence to:*
*jennl@microsoft.com*
*heckerma@microsoft.com*

## Overview

High-resolution HLA typing plays a central role in many areas of immunology, such as transplant matching, identifying immunogenetic risk factors for disease, studying how the genomes of pathogens evolve in response to immune selection pressures, and vaccine design. However, high-resolution HLA typing is frequently unavailable due to its high cost or the inability to re-type historical data. We recently introduced and evaluated a method for statistical, *in silico* refinement of ambiguous and/or low-resolution HLA data (Listgarten, et al. 2008). Here we present a summary of this work for the histocompatibility community. A tool based on our approach is available for research purposes at http://microsoft.com/science. The user selects an appropriate population of interest (i.e., African, Amerindian, Asian, European, Hispanic) and uploads the low-resolution/ambiguous data; our server then returns the statistically refined version of the HLA data.

## Introduction

HLA types are not always unambiguously determined; rather, they are only determined up to some "resolution" (i.e., level of ambiguity). Additionally, because the number of HLA alleles is constantly increasing, HLA types obtained using molecular methods that depend on the list of known alleles require constant re-interpretation in light of newly discovered alleles. This re-interpretation can result in more ambiguity than originally thought (Voorter, et al. 2007). Perhaps even more importantly, it is often impossible to re-type historic samples that may have been typed using lower-resolution approaches. As such, any method that can help to increase resolution of HLA data, post-hoc and at low cost, should be of use to the scientific and clinical communities.

High-resolution HLA typing is achievable using modern, molecular (DNA-based) methods. Molecular methods for HLA typing include hybridization with sequence-specific oligonucleotide probes (SSOP), PCR amplification with sequence-specific primers (PCR-SSP), and more recently, DNA sequence-based methods. Generally, DNA sequence-based methods involve locus-specific PCR amplification of exons 2 and 3 (for HLA Class I genes), or exon 2 only (for HLA class II), followed by "bulk" DNA sequencing of the amplified product (i.e., sequencing of products derived from both HLA haplotypes). Sequencing is restricted to exons 2 and/or 3 because these regions are the major determinants of HLA peptide-binding specificity and thus contain enough information to discriminate between most allele combinations. If an individual is heterozygous at any locus, direct sequencing of an amplified PCR product will yield nucleotide mixtures at positions in which the two alleles differ in sequence. Consequently, there are two reasons why modern sequence-based typing methods may yield ambiguous typing results: (1) if the differences between the two alleles are located outside the genotyped region (in most cases, exons 2 and/or 3), and (2) if two or more allele combinations yield the exact same pattern of heterozygous nucleotide mixtures when combined into a "bulk" sequence.

Assuming that HLA resolution beyond four digits is ignored, there are various levels of ambiguity that can arise from the second type of molecular (DNA)-based HLA typing ambiguity just described. For example, rather than knowing unambiguously which two HLA-A alleles a person has, one may instead know only a list of possibilities; for example, A*0301-A*3001 or A*0320-A*3001 or A*0326-A*3001. Such intermediate resolution types may result from sequence-specific PCR (SSP)-based typing where testing with the initial set of PCR primers will yield a list of possible genotypes that a particular person might have (which may require further testing with additional combinations of allele-specific primers and/or cloning and sequencing of clones before an unambiguous type is achieved). Depending on the clinical and/or research purpose of the HLA typing, additional laboratory testing required for achieving high-level (i.e., four-digit) resolution is often not performed for reasons relating to time and cost. In many cases, intermediate-level resolution data are truncated to two-digit resolution (low resolution). In the previous example, this individual would be reported as having HLA alleles A*03 and A*30.

In this paper, we describe and evaluate a statistical approach to resolving these low and intermediate resolution types. The approach uses linkage disequilibrium among HLA allele loci—the fact that certain alleles tend to be inherited together—in order to probabilistically resolve these ambiguities.

## Methods

We derive a method for ambiguity resolution of HLA types from a haplotype model—a statistical model that captures linkage disequilibrium across the HLA loci, and also from the over-

*Table 1: Summary of data used in experiments*

| Ethnicity | # of individuals | # of unique A alleles | # of unique B alleles | # of unique C alleles |
|---|---|---|---|---|
| North American European | 7526 | 81 | 129 | 48 |
| North American African | 3545 | 60 | 106 | 42 |
| Asian | 1318 | 43 | 76 | 30 |
| North American Hispanic | 881 | 47 | 106 | 35 |

all frequencies of alleles at each locus. The haplotype modeling part of our approach is related to the EM-based maximum-likelihood methods (Excoffier 1995; Hawley and Kidd 1995) found in the HLA literature (Cao, et al. 2001; Kollman, et al. 2004; Leffell, et al. 2007; Maiers, Gragert and Klitz 2006; Müller, Ehninger and Goldmann 2003; Thriskos, Zintzaras and Germenis 2007), although it differs in several crucial respects that make it more accurate when limited data is available, and allows it to more easily be applied to many loci with many alleles. We refer the reader to Listgarten, et al. 2008 for technical details of the approach.

Our approach involves two steps: (1) training a statistical model on high-resolution typing data (that is, fitting the parameters of the model to the training data), and (2) applying this trained model to our limited-resolution data of interest. For example, if a patient in our data set of interest was typed ambiguously at the A locus as having either (a) A*0243, A*0101, or (b) A*0243, A*0122, then our fitted statistical model assigns a probability to each of these two possibilities and we could take the one with the largest probability as an estimate of the high-resolution allele present.[1] More generally, our model assigns a probability to any number of possibilities (not just two), and over many loci. To date, we have used our method without computational difficulty to refine up to four loci with 20-130 alleles at each locus, and on data sets with up to half a million possible haplotypes.

Our web server makes use of a pre-compiled training data set, comprising data derived from a large collection of disease cohorts and controls (see Acknowledgments) that were all typed in the laboratory of Mary Carrington (see Listgarten, et al. 2008 for the typing protocols), as well as data from the NMDP as reported in Maiers, Gragert and Klitz 2006. Together, these two data sets represent 6,057 individuals of African-descent, 256 individuals of Amerindian descent, 3,088 individuals of Asian descent, 8,067 individuals of European descent, and 2,860 individuals of Hispanic descent, at the A, B and C loci. For the experiments described below, the NMDP portion of the data was not used as it was unavailable to us at the time we conducted them. Because alleles Cw17, Cw18 and A74 were almost never fully resolved to four digits in this data set, we left these as two digit designations.

To re-iterate our approach, our core methodology relies on accurately capturing haplotype and frequency information from the large set of high-resolution training data, treating each ethnicity separately. Once we have captured this information from the training data in the form of a fitted model, we then use this model to compute the probability of every four-digit haplotype consistent with the observed genotypes in our limited resolution data set, and in doing so, statistically resolve the ambiguities in it. By "capture this information," we mean that once we have fitted our model to the training data, this model then contains the haplotype and frequency information of the HLA alleles in the training data (and thus also in the limited-resolution data set of interest, if it is similar enough). Specifically, our fitted statistical model is able to compute the probability of every possible four-digit haplotype.

## Data and Experiments

In order to evaluate our statistical approach to resolving HLA ambiguities, we used data sets consisting of four-digit resolution HLA data from individuals at the A, B and C loci, as described above and shown in Table 1. Then we synthetically masked the known four-digit allele designation for some loci and some individuals, at random, and then tried to recover them using our statistical approach. In this way, ground truth is available for quantitative assessment. We assessed accuracy within each ethnicity. The way in which we did so for data of a particular ethnicity (e.g., Caucasian) was as follows: First, we selected 80 percent of the individuals[2] in this data set randomly as the training data set that we use to fit our statistical model (the first of the two steps mentioned at the start of the Methods section). Then, we used the remaining 20 percent of individuals[3] as a test set, as a proxy for a real data set of limited resolution (the second of the two steps mentioned in the Methods section).

Because this test set is not actually of limited resolution, we needed to synthetically make it so, and achieved this by randomly masking X percent of the allele calls in this data set to two-digits (using X=30 percent[4] and X=100 percent in two separate experiments). Once we had done this synthetic masking, we then resolved these masked alleles back to four digits with our model by taking the most probable four-digit genotype as computed by our model. Last, we counted the percentage of these masked alleles that were correctly estimated—what we refer to below as resolution accuracy. Additionally, for a baseline comparison, we computed the accuracy in resolved alleles when using a baseline model, one consisting of just the allele frequencies at each locus (thus throwing away any potential haplotype information).

A summary of the data used for these experiments is shown in Table 1. All but 0.1 percent of HLA alleles represented common and well-defined alleles (as classified in Cano, et al. 2007). For details on deviance from Hardy-Weinberg Proportions (HWP) in these data, see Listgarten, et al. 2008.

## Results

First we tried masking X=30 percent of the alleles in the test set to two-digit resolution (leaving the remainder at four-digit resolution). Use of our model here, which incorporates haplotype information and allele frequencies, resulted in ambiguity resolution accuracies of 95 percent, 90 percent, 90 percent and 86 percent, respectively, in the European, African, Asian, and Hispanic descent data sets. In contrast, use of the baseline model, which ignores haplotype information, resulted in ambiguity resolution accuracies of only 86 percent, 82 percent, 81 percent and 73 percent, respectively, in the European, African, Asian, and Hispanic descent data sets. Using the statistical test reported in Listgarten, et al. 2008, these differences are statistically significant ($p=10^{-4}$). Thus we see that overall, our statistical approach to resolving HLA ambiguities is useful, and also that using the haplotype information confers a significant benefit over not using this information. Note that the accuracy on the European data set may be better than that on the African data set for two reasons: the European data set is twice as large and thus has twice as much training data with which to fit the statistical model, and the European population is known to have undergone a population bottleneck that makes linkage disequilibrium in this population stronger than in African populations.

Next we tried masking 100 percent of the alleles in the test set to two-digit resolution (so that the entire test set was low-resolution). Surprisingly, the resulting accuracies were comparable to those in the 30 percent masking experiment described above, with 90 percent, 89 percent and 86 percent accuracies in the African, Asian, and Hispanic descent data sets, respectively, (we did not use the European data set here) as compared to the 90 percent, 90 percent, and 86 percent accuracies in the 30 percent masking experiment described previously. This suggests that most of the haplotype information at the four digit level can be captured at just the two digit level.

Further experimentation showing (1) how additional training data are likely to improve the resolution accuracies, (2) that our approach fares well on test data that is drawn from independent data sets such as from dbMHC, and (3) the locus-specific, accuracies can be found in Listgarten, et al. 2008.

## Conclusion

We have introduced a method for statistical refinement of low or intermediate resolution HLA data when a full resolution training data set from a similar population is available. Experimentally, we have shown that it is feasible to use statistical approaches for HLA refinement. Our HLA refinement method helps to mitigate the limiting factor of cost in HLA typing today, and allows for lower/intermediate resolution or historical data to be statistically refined when it cannot be refined by assay. A web tool that incorporates models based on a large collection of full resolution data is available for research purposes at http://microsoft.com/science.

## Acknowledgments

We wish to acknowledge the following cohorts and investigators from whom HLA data used in this study were derived: the International HIV Controllers Study, the Multicenter AIDS Cohort Study, the Multicenter Hemophilia Cohort Study, the Washington and New York Men's Cohort Study, the San Francisco City Clinic Cohort, the AIDS Linked to Intravenous Experience, the Swiss HIV Cohort, the Urban Health Study, the NIH Focal Segmental Glomerulosclerosis Genetic Study, Hepatitis C Antiviral Long-term Treatment against Cirrhosis, National Cancer Institute Surveillance Epidemiology and End Results Non-Hodgkin Lymphoma Case-Control Study, Woman Interagency Health Study, Classic Kaposi Sarcoma Case-Control Study I and II, Genetic Modifiers Study, Nairobi CTL Cohort, Grace John-Stewart, Stephen O'Brien, and Thomas O'Brien. We thank John Hansen for useful discussion, David Ross for code which contributed to the softmax regression, Giuseppe Cardillo for his publically available implementation of Guo & Thompson, 1992 and Fuh-Mei Duh for compiling the HLA data sets. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.  This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. We thank the National Marrow Donor Program for providing us with data used on our web server.

## Notes

[1] Even better, we could make use of the probabilities as explained on our web server and in Listgarten, et al. 2008.

[2] For example, the European data set contains 7,526 individuals according to Table 1, thus our training data set for the European populations consisted of 0.8 x 7,526=6,020 individuals.

[3] For example, the European data set contains 7,526, of which 6,020 were used for training, leaving 20% or 1,506 individuals for the test set.

[4] For example, in the European data set which had a test set of 1,506 individuals, X=30% means that 0.3 x 1,506 x 6 =2,710 alleles were masked, where 0.3 represents X=30% masking, 1,506 is the test set size as described earlier, and 6 is the number of allele calls available for each individual (two A alleles, B alleles and C alleles). The resolution accuracy that we use to evaluate our approach is defined as the percentage of these 2,710 alleles that had their four-digit resolution correctly estimated by the statistical model.

## *Bibliography*

1. Cano, P, et al. "Common and well-documented HLA alleles: report of the ad-hoc committee of the American Society for Histocompatiblity and Immunogenetics." Human Immunology, no. 168 (2007): 392–417.
2. Cao, K, J Hollenbach, X Shi, W Shi, M Chopek, and MA Fernández-Viña. "Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations." Hum Immunol 62, no. 9 ( 2001): 1009-30.
3. Excoffier, L, Slatkin, M. "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." Mol Biol Evol, no. 12 (1995): 921–7.
4. Guo, S., & Thompson, E. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics (48), 361–372.
5. Hawley, M, and K Kidd. "HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes." J Hered, no. 86 (1995): 409–411.
6. Kollman, C, et al. "Assessment of optimal size and composition of the u.s. national registry of hematopoietic stem cell donors." Transplantation, no. 78 (2004): 89–95.
7. Leffell, MS, WS Cherikh, G Land, and AA Zacharay. "Improved definition of human leukocyte antigen frequencies among minorities and applicability to estimates of transplant compatibility." Transplantation, no. 83 (2007): 964–972.
8. Listgarten, Jennifer, Zabrina Brumme, Carl Kadie, Gao Xiaojiang, Bruce Walker, and David Heckerman. "Statistical Resolution of Ambiguous HLA Typing Data." PLoS Computational Biology 4, no. 2 (2008): e1000016.
9. Maiers, M, L Gragert, and W Klitz. "High resolution HLA alleles and haplotypes in the US population." Human Immunology 67 (2006): S16.
10. Müller, CR, G Ehninger, and SF Goldmann. "Gene and haplotype frequencies for the loci HLA-A, HLA-B, and HLA-DR based on over 13,000 german blood donors." Hum Immunol, no. 64 (2003): 137–51.
11. Thriskos, P, E Zintzaras, and A Germenis. "DHLAS: A web-based information system for statistical genetic analysis of HLA population data." Comput Methods Prog Biomed, no. 85 (2007): 267-272.
12. Voorter, C, E Mulkers, P Liebelt, Sleyster E, and E van den Berg-Loonen. "Reanalysis of sequence-based HLA-A, -B and -Cw typings: how ambiguous is today's SBT typing tomorrow." Tissue Antigens, no. 70 (2007): 383-9.

EDITORIAL

# From the Editor-in-Chief

Sharon Adams, *MT, CHS(ABHI)*

As Editor-in-Chief, I am happy to present to you an exciting issue that has some information for all types of laboratories. Some practical topics are discussed as well as some novel ideas for HLA typing and possible resolution of ambiguous allele combinations. As always, the efforts of the *ASHI Quarterly* will be to provide information that is relevant to ASHI community members. It is with this intent that, as Editor-in-Chief, I continue to implore members to provide suggestions for future articles and also to encourage the submission of articles to the magazine.

This issue of the *ASHI Quarterly* has some very interesting practical topics for reflection. The first article, "Got Milk? A Simple Method for Reducing the Undesirable Effects of Pronase Treatment of Lymphocytes," submitted by Paul Warner from the Puget Sound Blood Center, will be of particular interest to laboratories performing flow cytometry crossmatching. The article presents three types of lymphocyte treatments: pronase only, pronase and milk, and milk only. Review of the data provides the conclusion that lymphocyte treatment with both pronase and milk provides the optimal effects on the crossmatch. The article presents data that is very practical in application. Laboratories performing this methodology may be encouraged to experiment with this technique after reading this article.

The second article, "Introduction to BioArray Solutions' BeadChip Technology," submitted by W. Tait Stevens from Loma Linda University Medical Center, presents a novel technique for HLA typing. The technique is very early in its development but combines approaches that most HLA laboratories are familiar with due to other techniques which have been utilized in the past. The Beadchip technology requires attaching DNA oligonucleotides to polystyrene beads of various colors. DNA fragments are then exposed to the surface of the chip. Bound DNA is elongated with fluorescence-labeled dNTPs. Following elongation, a photomicrograph, a decoder image, is taken of the fluorescence and then analyzed. This is an exciting new technology for the future. The article presents the opportunity to the community to understand the technique in its very early development.

The third article, "*In Silico* Resolution of Ambiguous HLA Typing Data," submitted by Jennifer Listgarten and David Heckerman from Microsoft Research, presents an alternative approach to the resolution of ambiguous allele typing combinations. As more laboratories perform high resolution HLA typing, the need to establish a method to handle the ambiguous allele combinations is becoming extremely apparent. The article describes a statistical approach that can resolve low resolution/ambiguous HLA typing data. A large volume of data was utilized to determine the efficacy of this approach. The article provides some promising insight into the future of how the HLA community could potentially resolve this ever-growing problem.

As you read this issue, consider how these ideas might be incorporated in the future of your own laboratory. And I encourage you to provide feedback to these articles, so that as Editor-in-Chief, I can determine if I am providing topics of interest to community members. I hope you will enjoy this issue of the *ASHI Quarterly*. Hopefully, it will spark some ideas for you to explore.

## ASHI Quarterly Scientific Communications Guidelines for Authors

The *ASHI Quarterly* invites submissions of scientific articles of interest to the histocompatibility and immunogenetics community. In general, the articles should relate to clinically-relevant topics in transplantation immunobiology, including reviews of novel techniques or assays, testing/assay applications, specific disease states in transplantation medicine, characteristics of specific lymphocyte subpopulations or lymphocyte markers, and anthropology/population genetics. Articles examining the history and impact of immunogenetics on transplantation medicine are also appropriate. The articles should be 2-4 pages in length (double spaced), and include relevant references.

In order to further the goal of providing educational content with the articles and provide Continuing Education Credit to the ASHI membership, authors are requested to submit five quiz questions pertinent to the article, in multiple choice format. The questions should relate to the main points of the article, and provide a means to ascertain the reader understands the relevance and principles presented in the article. Please provide the reasoning behind why the right answer is considered correct, and more importantly, why the alternative answers are not correct, unless it's obvious. The answers to the questions will be published in the *ASHI Quarterly* immediately following the edition in which the quiz is presented, in order to provide educational feedback to the reader membership.