JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 3, 2011

Pp. 483–493
DOI: 10.1089/cmb.2010.0258

HLA Type Inference via Haplotypes Identical by Descent

MANU N. SETTY,[1] ALEXANDER GUSEV,[2] and ITSIK PE'ER[2]

ABSTRACT

The human leukocyte antigen (HLA) genes play a major role in adaptive immune response and are used to differentiate self antigens from non-self ones. HLA genes are hypervariable with nearly every locus harboring over a dozen alleles. This variation plays an important role in susceptibility to multiple autoimmune diseases and needs to be matched on for organ transplantation. Unfortunately, HLA typing by serological methods is time consuming and expensive compared to high-throughput single nucleotide polymorphism (SNP) data. We present a new computational method to infer per-locus HLA types using shared segments identical by descent (IBD), inferred from SNP genotype data. IBD information is modeled as graph where shared haplotypes are explored among clusters of individuals with known and unknown HLA types to identify the latter. We analyze performance of the method in a previously typed subset of the HapMap population, achieving accuracy of 96% in HLA-A, 94% in HLA-B, 95% in HLA-C, 77% in HLA-DR1, 93% in HLA-DQA1, and 90% in HLA-DQB1 genes. We compare our method to a tag SNP-based approach, and demonstrate higher sensitivity and specificity. Our method demonstrates the power of using shared haplotype segments for large-scale imputation at the HLA locus.

Key words: algorithms, cancer genomics, DNA arrays, gene expression.

1. INTRODUCTION

The human leukocyte antigen (HLA) region, located on chromosome 6p21, encodes genes for the major histocompatibility complex (MHC) in humans. MHC are cell surface proteins that play an important role in adaptive immune response. These proteins form a complex with the antigenic peptides that are presented on the cell surface. This complex is recognized by the T-cell receptors to trigger the adaptive immune response by inducing the death of the cell and/or production of antibodies.

The HLA genes are classified into two main classes. Class I genes present peptides from within the cell and are recognized by the $CD8^+$/cytotoxic T cells that kill the cells displaying the antigens. The Class I MHC genes are HLA-A, HLA-B, and HLA-C. Class II genes present peptides from the intracellular vacuoles and are recognized by the $CD4^+$/helper T cells that trigger antibody production. The Class II genes are HLA-DP, HLA-DM, HLA-DOA, HLA-DOB, HLA-DQ, and HLA-DR. HLA genes are also

[1]Computational Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York.
[2]Department of Computer Science, Columbia University, New York, New York.

highly polymorphic. For example, HLA-A has 893 alleles, and HLA-DRB has 814 alleles (Robinson et al., 2003). The large number of alleles enables the immune system to respond to a wide range of pathogens.

The HLA nomenclature (Holdsworth et al., 2009) is illustrated in Figure 1. The convention is to use a four-digit code to distinguish alleles that differ in their protein products. The first two digits represent the allele family, determined by serological typing and represent the antigen recognized by the allele family. The antigen A24 is recognized by allele in Figure 1. The third and fourth digits represent amino acid differences due to non-synonymous mutations. The remaining digits represent other non-coding differences. The required resolution depends on the immunological application under consideration.

HLA genes have been implicated in a number of autoimmune diseases such as Crohn's disease (Breese et al., 1993; Brimnes et al., 2005) and multiple sclerosis (Lincoln et al., 2005). Physically, different gene variants have different abilities to present antigens and therefore incur different sensitivities to their presence. Matching HLA types are therefore required for organ transplantation to succeed. However, experimental methods of HLA typing are time consuming and expensive (Leslie et al., 2008). Indirect typing using tag SNPs (de Bakker et al., 2006) is confounded by the unusual patterns of recombination and selection that require locus-specific methods (Leslie et al., 2008). Moreover, the region also contains long stretches of high linkage disequilibrium (LD), often spanning several megabases and HLA loci (Miretti et al., 2005), curbing the performance of standard models of genetic variation.

Leslie et al. (2008) have developed a method for HLA type inference based on allele combinations or haplotypes. A hidden Markov model is used to calculate the probability of observing a specific HLA allele by modeling the chromosome as an imperfect mosaic of the ancestral haplotypes carrying the same allele. The described model assumes that the parental origin (phase) of each allele haplotype is known for SNP data and uses a training set with known, phased HLA types.

With the availability of large and comprehensively genotyped cohorts, tools have been developed to harness many samples for detecting identity by descent (IBD) between pairs of individuals (Gusev et al., 2009; Purcell et al., 2007) that carry copies of the same local haplotype from a recent common ancestor. Such analysis naturally applies to the haplotypic structure of the HLA, with the special interest in this region motivating increased attention. This attention is required because of the special haplotype structure at the HLA (Horton et al., 2004).

We have recently developed a method that accurately detects all long IBD shared regions from genotype marker data (Gusev et al., 2009). The method uses a dictionary-based sliding window approach to identify long, nearly-identical regions between pairs of individuals in linear time. Here, we present a graph-based method that uses segments shared between HLA-typed and un-typed individuals to infer their putative HLA types. We provide theoretic description of the model and offer a software implementation, a unique contribution to the geneticist user.

The article is organized as follows: We define the framework and the problem in Section 3. Section 4 describes our algorithms for HLA imputation. The data used for analysis is explained in Section 5. The results and comparison to tag SNP method are presented in Section 6, followed by a summary discussion in Section 7.

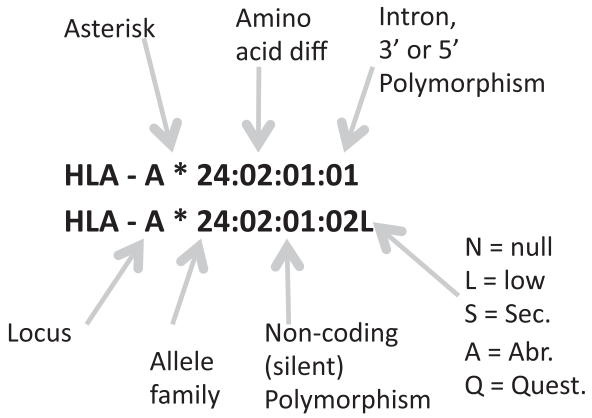

FIG. 1. HLA nomenclature.

2. PRELIMINARIES

We define a model for inferring HLA types at individual loci for unphased data. We study one locus at a time and throughout the methods sections consider only the current locus. The results repeat such analysis for each locus separately along the HLA region. An individual v is associated with a pair of alleles (α, β) at each HLA locus, representing the HLA types. We denote this by $v(\alpha, \beta)$. An individual with $\alpha = \beta$ is homozygous. The input consists of a set of individuals with known HLA types and another set with unknown HLA types. The individuals in these two sets are referred to as *resolved* and *unresolved* individuals, respectively. Unphased IBD segments that are shared pair-wise across resolved and unresolved individuals are inferred using GERMLINE (Gusev et al., 2009) and serve as a starting point for our analysis.

Formally, IBD is represented as an undirected graph called the *IBD-Graph*, G_{IBD}. The nodes V of G_{IBD} map to the individuals with genotypic data (resolved and unresolved) and the edges E represent the IBD shared segments. Ideally we would have G_{IBD} as input for HLA imputation, but in practice we may only assume the input to be a noisy version G^0_{IBD} of the true G_{IBD}. G^0_{IBD} has the same nodes, as G_{IBD} along with many of the same edges (true positives), but it also contains false positives (edges between nodes not related by IBD) and false negatives (missing edges between nodes related by IBD).

An edge in G^0_{IBD} between two nodes $v(\alpha, \beta)$ and $w(\gamma, \delta)$ is suggestive of the nodes sharing one or both the HLA types, i.e., at least one of $(\alpha = \gamma)$, $(\beta = \gamma)$, $(\alpha = \delta)$, or $(\beta = \delta)$ is true. The edges which satisfy these criteria are termed *consistent*. Note that the converse does not hold: if two nodes share a common HLA type, it does not imply they are IBD because the same HLA allele can have multiple SNP-haplotype backgrounds. The HLA imputation problem is intuitively defined as follows:

Input: $G^0_{\text{IBD}}(V, E^0)$ and a set of assigned type pairs $(\alpha_{(r)}, \beta_{(r)})$ for all nodes r in a resolved subset $R \subset V$
Output: Assignment of type pairs $(\alpha_{(u)}, \beta_{(u)})$ for all unresolved nodes $u \in V \backslash R$
Objective: Maximize the correctly assigned nodes.

As the objective is not defined in terms of the available data, we consider a surrogate optimization criterion. We seek an assignment that maximizes the number of consistent edges.

We propose an iterative approach for HLA imputation. While $G^0_{\text{IBD}}(V, E^0)$ is used as the input for the first iteration, the IBD-Graph is adjusted in each iteration to maintain the consistency of edges. Formally, denote the IBD-Graph in the i^{th} iteration as G^i_{IBD}. We detect false positives and false negatives which are removed from and added to the edge set respectively to form E^i, the edge set in the i^{th} iteration. After adjusting the graph, possible HLA types and HLA type-pairs are inferred for unresolved nodes. Possible HLA types represent alleles of one of the chromosomes satisfying the constraints defined by G^i_{IBD} and HLA type-pairs represent alleles of both the chromosomes of the unresolved node.

G^i_{IBD} is examined in triplets of nodes, $T(r_1, r_2, u)$, where r_1, r_2 are resolved and u is unresolved and at least two of the edges (r_1, r_2), (r_1, u) and (r_2, u) are in E_i. The possible HLA types and type-pairs from all triplets containing u are combined based on a likelihood function to assign the most likely HLA types to the unresolved node. We expect a number of unresolved nodes to be resolved within each iteration. This information is then used in subsequent iterations to infer HLA types for the remaining ambiguous or unresolved nodes (Fig. 2)

2.1. Sources of information

The sources of information for defining possible HLA types are triplets generated, matches with homozygote nodes and previously detected false negatives. Triplets and homozygote matches are deduced from G^i_{IBD}.

We define three possible configurations for a triplet based on the sub-graph of G^i_{IBD} induced by (r_1, r_2, u). If this sub-graph is a clique, we call it a *triangle triplet* (Fig. 3a). Alternatively, it is a path along the three nodes and we denote this as an *end triplet* (Fig. 3b) or a *middle triplet* (Fig. 3c) depending on the position of u along the path. Possible HLA types are deduced from each triplet as described below.

Triangle triplets (Fig. 3a) are fully connected by definition. Since only consistent edges are considered for any triangle triplet $T(r_1, r_2, u)$, r_1 and r_2 share one or both HLA types. We consider these cases in turn. (1) One shared HLA-type: We denote the HLA types of the resolved nodes by $r_1(\rho, \alpha)$ and $r_2(\rho, \beta)$ with $\alpha \neq \beta$. In this case the following assignments to u maintain the consistency of the edges: the shared HLA type: ρ or the HLA type-pair formed by (α, β) (Fig. 4A(I)). (2) Both HLA types shared: Here, we denote the

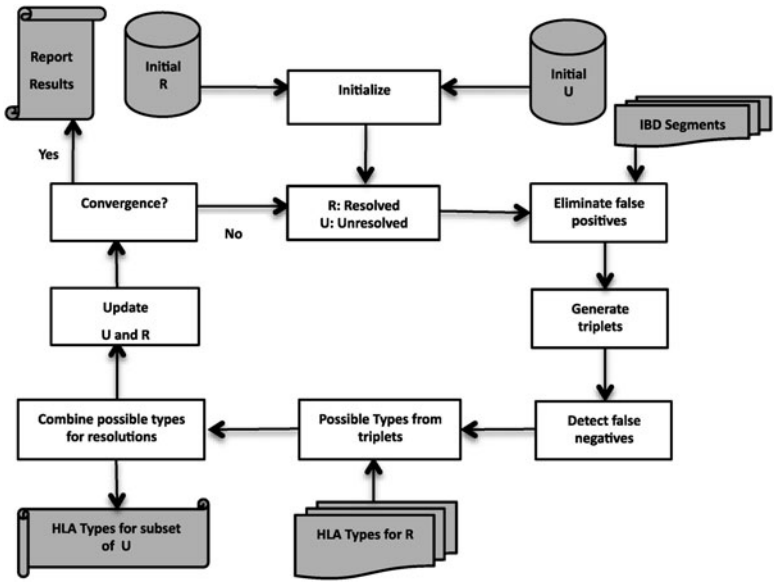

FIG. 2. An iterative-triangulation approach for HLA type inference from unphased data. The method initializes the resolved and unresolved nodes from the training and test sets, respectively. The edges among these individuals are used to generate triplets. These triplets are used to draw up a set of possible HLA type resolutions for each node. The HLA types with highest likelihood are chosen as resolution for the nodes where applicable and the process is repeated for the remaining unresolved nodes.

HLA types by $r_1(\rho, \tau)$ and $r_2(\rho, \tau)$. Thus, the shared types, ρ and τ, are possible HLA types for u (Fig. 4A(II)).

End triplets (Fig. 3b) are processed as follows: For a triplet $T(r_1, r_2, u)$ assume without loss of generality that $(r_1, u) \in E^i$ and $(r_2, u) \notin E^i$. We denote the HLA types of the resolved nodes by $r_1(\rho, \alpha)$ and $r_2(\rho, \beta)$. By definition, if $\alpha \neq \beta$, then assigning the HLA type of r_1 not shared with r_2, i.e., α, to u maintains the consistency of the edges (Fig. 4B(I)). Otherwise, if $\alpha = \beta$, the edge (r_1, r_2) is detected as a false negative and is added to E^i. The triplet is treated as a triangle triplet in the subsequent iterations. For example, the triplet in Figure 4B(II) defines ρ and α as possible HLA types.

Middle triplets (Fig. 3c) do not have an edge between the resolved nodes. For any middle triplet $T(r_1, r_2, u)$, r_1 and r_2 are not known to share any HLA types since $(r_1, r_2) \notin E_i$. Denoting types by $r_1(\alpha, \beta)$ and $r_2(\gamma, \delta)$, all HLA type pairs (ρ, τ) where $\rho \in \{\alpha, \beta\}$ and $\tau \in \{\gamma, \delta\}$ are assigned as possible HLA types of u. Each type-pair maintains the consistency of the edges. The triplet in Figure 4C(I) defines (α, γ), (α, δ), (β, γ) and (β, δ) as possible HLA type-pairs.

If any of $\alpha = \gamma$, $\alpha = \delta$, $\beta = \gamma$ or $\beta = \delta$ is true, then it is an indication of a false negative. Again, we add the edge (r_1, r_2) to E_i and the triplet is treated as a triangle triplet. For example, the triplet in Figure 4C(II) defines α and β as possible HLA types and triplet in Figure 4C(III) defines possible HLA type α and possible HLA type-pair (β, δ).

Lastly, unresolved nodes maybe connected to resolved nodes that are homozygous in HLA alleles. If $(u, r) \in E_i$ where the HLA types of $r(\alpha, \alpha)$, the triplet containing $(u, r) \in E_i$ defines α as a possible type for u.

3. ALGORITHMS

3.1. Triplet generation

The edges of G^i_{IBD} are represented using the adjacency list representation. More precisely, for efficiency reasons, any given individual stores two adjacency lists, for resolved and unresolved neighbors, respectively.

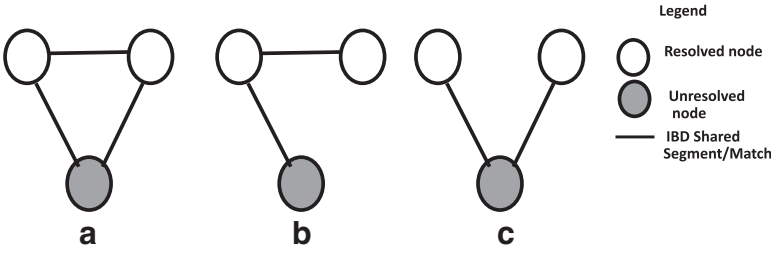

FIG. 3. Types of triplets. (**a**) Triangle triplet: pair-wise matches between all individuals. (**b**) End triplet: unresolved individual has match with only of the resolved individuals. (**c**) Middle triplet: resolved individuals do not have a match.

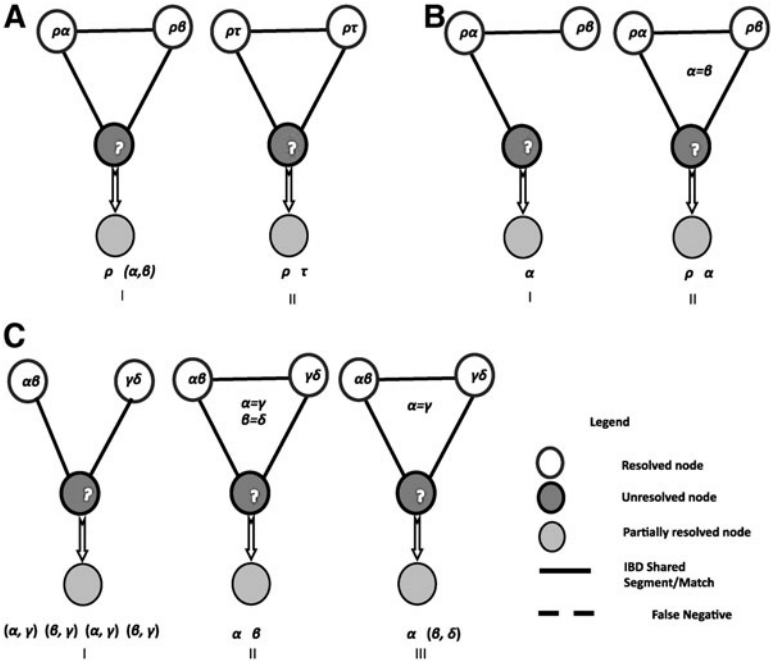

FIG. 4. Possible type from triplets. (**A**) Possible type generation for triangle triplets. (I) One shared type: shared type or the combination of non-shared types. (II) Two shared types. (**B**) Possible type generation for end triplets. (I) One shared type: the non-shared type. (II) False negative: two shared types. (**C**) Possible type generation for middle triplets. (I) Consistent edges: combinations of the types of resolved individuals. (II) False negative with two shared types. (III) False negative with one shared type: shared type and the combination of non shared types.

The algorithm for triplet generation is formally described in Figure 5. Briefly, triplets are generated by traversing the graph for all paths of length 3 containing only one unresolved individual. Each traversal starts from a resolved individual r, and progress in two ways based on the status of the adjacent individual, a:

1. If a is resolved, traverse through all the unresolved adjacent individuals of a. This will generate candidate end triplets.
2. If a is unresolved, traverse through all the resolved adjacent individuals of a to generate candidate middle triplets and traverse through all resolved adjacent individuals of r to generate candidate end triplets.

If a trio of individuals generates both end and middle candidate triplets, a *triangle triplet* of the trio is added. Other candidate end or middle triplets are indeed end or middle triplets respectively. Duplicate generated triplets are identified and removed. The algorithm then proceeds to resolve individuals.

3.2. Type resolution

Define $S = \{t, e, m, f, h\}$ as the categories of information, representing **t**riangle triplets, **e**nd triplets, **m**iddle triplets, **f**alse negatives and **h**omozygous matches respectively.

Type resolution assigns the most likely HLA types to an unresolved individual. Define $C_s(\alpha)$ as the number of instances of category s defining α as a type for the unresolved node, u under examination.

The quintuple $(C_t(\alpha), C_t(\alpha), C_m(\alpha), C_f(\alpha), C_h(\alpha))$ is the sufficient statistic for calculating the likelihood of α being assigned as follows: Define $L_t(s)$ to be the likelihood of triplet of category s being correct and $L_f(s)$ as likelihood of triplet of category s being incorrect. Define C_s^u to be the total number of triplets of category s for unresolved individual u. The likelihood of α being the resolution for individual u is calculated as

$$Likelihood(\alpha|u, \text{Counts}) = \prod_{s \in S} L_t(s)^{C_s(\alpha)} \prod_{s \in S} L_f(s)^{C_s^u - C_s(\alpha)} \tag{1}$$

For HLA type-pairs (α, β), define $E_s(\alpha, \beta)$ to be the effective count triplets of category s defining (α, β) as a possible HLA type-pair. The likelihood calculation of HLA type-pair uses the same formula but after calculating the effective counts given by

$$E_s(\alpha, \beta) = C_s(\alpha) + C_s(\beta) - C_s(\alpha, \beta) \tag{2}$$

where $C_s(\alpha, \beta)$ represents the triplets of category s, defining both α and β as possible types.

```
GENERATE-TRIPLET (G^i_IBD, R_{i-1}):
G^i_IBD: IBD-Graph
R_{i-1}: Set of individuals resolved in iteration (i-1)
define sets T_t, T_e, T_m: Set of triangle, end
and middle triplets respectively

for r in R_{i-1} do
    for v s.t (r, v) ∈ E^i

        if v.RESOLVED = TRUE
        then
            T'_e = T'_e ∪ {(r,v,u)|(v,u) ∈ G^i_IBD, u.RESOLVED = FALSE}
        else
            T'_m = T'_m ∪ {(r,v,w)|(v, w) ∈ G^i_IBD, w.RESOLVED = TRUE}
            T'_e = T'_e ∪ {(r,v,w)|(v,w) ∈ G^i_IBD, w.RESOLVED = FALSE}

T_t = T'_e ∩ T'_m
T_e = T'_e \ T_t
T_m = T'_m \ T_t
```

FIG. 5. Algorithm for triplet generation.

The score for HLA type-pair (α, β) and individual u is calculated as

$$Likelihood((\alpha,\beta)|u,\text{Counts}) = \prod_{s\in S} L_t(s)^{E_s(\alpha,\beta)} \prod_{s\in S} L_f(s)^{C_s^u - E_s(\alpha,\beta)} \tag{3}$$

Define p^+, q^+, p^- and q^- to be the likelihoods of true positive, false positive, true negative, and false negative edges, respectively. The likelihoods for the triplets being correct or incorrect are calculated as in Table 1.

We estimate the following rates from GERMLINE $p^+ : 0.9$, $p^- : 0.85$, $q^+ : 0.1$, $q^- : 0.15$ (Gusev et al., 2009).

The algorithm for type resolution is formally described in Figure 6. The algorithm greedily resolves individuals by the likelihood calculation. Our implementation maintains a hash-map of all possible HLA types and type-pairs for each individual. The value of the hash-map is the quintuple, with the type or the type-pair being the key.

The most likely HLA type, η is first chosen. If the individual is determined to be homozygous genotypically, η is assigned as the resolution for the individual. If η satisfies all the edges with the resolved individuals, the individual is considered *potentially homozygous* in HLA types. In such cases, the individual is left unresolved and retained for processing in the further iterations.

HLA type-pairs are formed by combining each possible HLA type with η. The two most likely type-pairs are determined and the HLA type-pair with highest likelihood is assigned as resolution, if the difference between their likelihoods is greater than zero.

At the end of each iteration, the adjacency lists of individuals containing newly resolved individuals are updated to move the newly resolved individuals to the head of the list. Thus, the entire adjacency matrix needs to be constructed only once at the start of the algorithm. All the steps are repeated until convergence where no more resolutions are possible.

TABLE 1. LIKELIHOOD TERMS FOR CORRECT AND INCORRECT TRIPLETS

Category	*Likelihood of correct triplet*	*Likelihood of incorrect triplet*
Triangle triplet	$(p^+)^3$	$(q^+)^3$
End triplet, middle triplet	$(p^+)^2 \times p^-$	$(q^+)^2 \times q^-$
False negative	$(p^+)^2 \times q^-$	$(q^+)^2 \times p^-$
Triangle triplet	p^+	q^+

```
RESOLUTION (U)
U: Unresolved individuals

for u in U do
    determine η: highest scoring possible type
    if u is homozygous
    then
        u.TYPES := (η, η)
        u.RESOLVED := TRUE
    else
        if η satisfies all the matches with resolved r: (u , r) ∈ G^i_IBD
        then
            u cannot be completely resolved in the current iteration
            u.RESOLVED := FALSE
            this is an indication of possible homozygous HLA types
        else
            for each possible type α do
                if τ ≠ η
                then
                    define type-pair (α, η)
                    False negative detection based on η (Fig.7)
                    Calculate LIKELIHOOD((α, η))

            find (α, τ) and (β, τ), the two most likely type-pairs
            if LIKELIHOOD((α, η)) ≠ LIKELIHOOD((β, η))
            then
                u.TYPES := (α, τ)
                u.RESOLVED := TRUE
            else
                u.RESOLVED := FALSE
```

FIG. 6. Algorithm for type resolution.

3.3. Complexity and implementation

Triplet generation explores all paths of length three containing only one unresolved node. Thus, the time complexity for triplet generation can be estimated as $O(|R|^2|V\backslash R|)$, where R is the set of resolved nodes and V is the set of all nodes in the graph.

The runtime of type resolution is linearly dependent on the number of triplets generated since each triplet is examined only once to identify the possible HLA types and type-pairs. Thus, $O(|R|^2|V\backslash R|)$ is the bound on complexity.

The program was implemented in Java 1.5, and testing was done on a Linux node of 2×2.4 GHz Xeon CPUs with 2 GB of memory. The average runtimes per individual for cross validation of HLA-A, HLA-B, HLA-C, HLA-DRB1, HLADQA1, HLADQB1 genes were 15, 4, 3, 0.7, 0.7, 1.3 seconds, respectively, analyzing 5475, 3328, 3387, 2899, 2545, 2426 IBD shared segments. The software has been made available for download at http:/www1.cs.columbia.edu/∼itsik/hla_ibd/index.html.

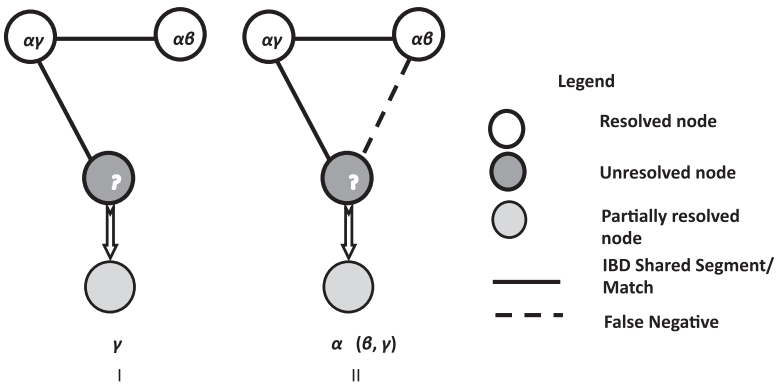

FIG. 7. Illustration of end triplets detecting false negatives. (I) The generated end triplet. (II) The same end triplet becomes a false negative if β is chosen as η.

4. DATA

The data used for analysis has been described in de Bakker et al. (2006). Briefly, the data includes 90 individuals (30 parent-offspring trios) of the Yoruba people from Ibadan, Nigeria (YRI); 182 Utah residents (29 extended families of European ancestry, from the Centre dEtude du Polymorphisme 6 is available. HLA typing was carried out for class I (HLA-A, HLA-B, HLA-C) and class II (HLA- DRB1, HLA-DQA1, HLA-DQB1) genes using the PCR-SSOP protocols. The CEU and YRI populations were used for analysis of the model and all the data was assumed to be unphased for *Centre d'Etude du Polymorphisme Humain (CEPH)* collection (CEU); 45 unrelated Han Chinese from Beijing, China (CHB); and 44 unrelated Japanese from Tokyo, Japan(JPT), 6338 variants located in a 7.5Mb region on chromosome 6 is available. HLA typing was carried out for class I (HLA-A, HLA-B, HLA-C) and class II (HLA-DRB1, HLA-DQA1, HLA-DQB1) genes using the PCR-SSOP protocols. The CEU and YRI populations were used for analysis of the model, and all the data were assumed to be unphased for analysis.

5. RESULTS

We established the accuracy of our method on the CEU HapMap data. Intuitively, the effectiveness of the model is dependent on the presence of IBD among the individuals under consideration. ''Leave one out'' cross-validation offers a good approach to test the model since it utilizes all the available IBD shared segment information for the individual being tested. An individual being tested is instantiated as unresolved, and all the other individuals form the resolved individuals or training set for the model. The individual can either be inferred as resolved, if types on both chromosomes are inferred; as ambiguous, if two types cannot be inferred or more than two types are equally likely; as potentially homozygous, if only a single type is present and inferred; or as unresolved, if no inference can be made. The model is not dependent on learning any parameters from the training data and therefore leave one out cross-validation does not bias the results.

Each locus is analyzed separately and the accuracy and coverage are defined with respect to the number of chromosomes analyzed. Formally, let $u(\rho, \tau)$ be the individual being tested. If u is inferred as resolved with HLA types (α, β), both the chromosomes are accounted as called and both are correct if $(\alpha, \beta) = (\rho, \tau)$. Only one of the chromosomes is considered correct if $\alpha \in (\rho, \tau)$ and $\beta \notin (\rho, \tau)$ or vice versa. If u is inferred as ambiguous with HLA type α, one chromosomes is called and is correct if $\alpha = \rho$ or $\alpha = \tau$. If u is inferred as potentially homozygous with type α, both chromosomes are called; both are correct is $\alpha = \rho = \tau$, one is correct if $\rho \neq \tau$ and, $\alpha = \rho$ or $\alpha = \tau$.

The coverage and accuracy are measures as follows

$$Coverage = \frac{TotalCalled}{TotalAnalyzed} \tag{4}$$

$$Accuracy = \frac{TotalCorrect}{TotalCalled} \tag{5}$$

Table 2 lists the results of leave one out cross-validation tests on the CEU HapMap population. The analysis examined shared segments which span 100kb upstream and downstream of the gene under consideration. Results are shown for both four-digit and two-digit resolutions. HLA types occurring only once in the population and types which are not resolved to the required extent are excluded from analysis. The model predicts results with high accuracy in the HLA-A, HLA-B, HLA-C, HLA-DQA1, and HLA-DQB1 alleles at four-digit resolution. The accuracy for class I genes remains the approximately the same at both the resolutions, but using two-digit resolution leads to higher accuracy in the class II genes. This can be attributed to a reduction in the false positive matches at lower resolution.

The main sources of error are false positives and non-availability/non-detection of IBD between individuals. The distribution of the false positives at four-digit resolution for the different genes is illustrated in Figure 8. For each individual, the false positive percentage in a region is the percentage of matches of the resolved adjacent nodes which are false positive. The HLA-DRB1 region has a higher number of individuals with large false positive percentages which is reflected in the low accuracy prediction. Fitting the parameters of IBD detection, especially the specific span used will improve results, but we chose to present benchmarks with vanilla parameters across all genes.

Table 2. Results of Leave One Out Cross-Validation for CEU Population by Considering IBD Shared Segments Which Span 100kb Upstream and Downstream of the Gene

	Four-digit		*Two-digit*	
Gene	*Analyzed*	*Accuracy*	*Analyzed*	*Accuracy*
HLAA	314	96.5	322	96.4
HLAB	281	94.3	311	93.4
HLAC	328	94.2	316	94.9
HLADRB1	308	77.6	294	91.3
HLADQA1	350	92.6	330	94.3
HLADQB1	350	90.3	323	92.3

We compare our results to the results from phasing the data. Phased version of the CEU data based on trios is available from de Bakker et al. (2006). We used GERMLINE (Gusev et al., 2009) to obtain pairwise IBD segments between the haplotypes. Leave one out cross-validation is again used for testing. If M is the set of matches determined by GERMLINE, the likelihood of allele α being the resolution for chromosome c is calculated as below

$$Likelihood(\alpha|c, M) = \frac{H(\alpha)}{\sum_{\beta \in A} H(\beta)} \tag{6}$$

where $H(\alpha) = \sum_{(y,c)\in M} \delta(\beta_y, \alpha)$ where β_y is the HLA type of chromosome y and δ function given by

$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{if } \alpha \neq \beta \end{cases} \tag{7}$$

The HLA type with highest likelihood as assigned as resolution. If two or more HLA types are tied for the highest likelihood, the chromosomes is left unresolved and considered ambigious.

We also used fastPHASE (Scheet and Stephens, 2006) to perform phasing without using the trio information and again use GERMLINE to determine set of matches using haplotypic extensions of matches rather than genotypic extensions. The triplet-based algorithm is used to determine the resolutions in a leave one out cross-validation setting.

The performance is assessed by means of the sensitivity and specificity differences. If A is the set of alleles under examination and $P_\alpha^+, P_\alpha^-, N_\alpha^+$ and N_α^- are the positive, false positive, true negative, and false negatives for allele α, respectively, then sensitivity and specificity are calculated as below:

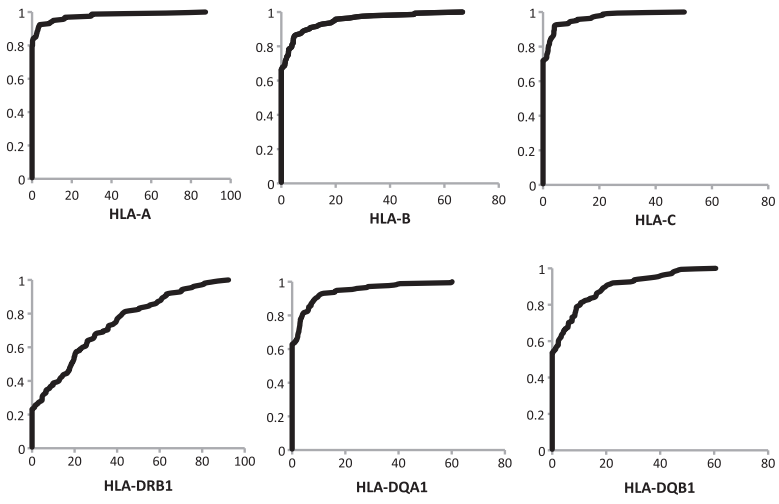

FIG. 8. False positive rate distribution. The *x*-axis represents the false positive percentage, and the *y*-axis represents the fraction of individuals. Each point on the graph represents the fraction of individuals with false positive percentage less than the corresponding reference value.

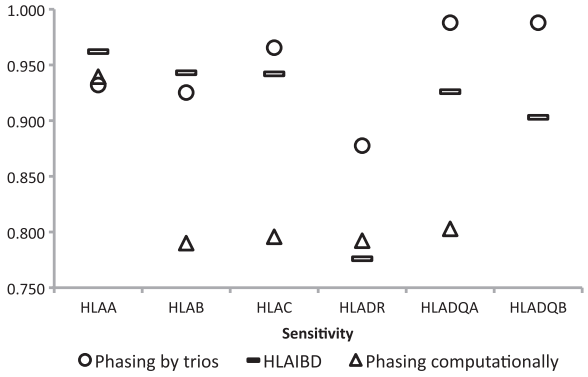

FIG. 9. Sensitivity of the different approaches.

$$Sensitivity = \frac{\sum_{\alpha \in A} P_{\alpha}^{+}}{\sum_{\alpha \in A} (P_{\alpha}^{+} + N_{\alpha}^{-})} \tag{8}$$

$$Specificity = \frac{\sum_{\alpha \in A} N_{\alpha}^{+}}{\sum_{\alpha \in A} (N_{\alpha}^{+} + P_{\alpha}^{-})} \tag{9}$$

The comparison plot is shown in Figure 9. All the methods show very high specificity (>0.95 in all the genes). The performance of our method compares well with trio-based phased data results in the class I genes, whereas having phased data has significant benefits for the class II genes. This could be possibly because of a reduction in the false positive rates in IBD matches when using trio-based phasing. This demonstrates that our method can be applied to unphased data with accuracy comparable to phased data when the false positive rates in the IBD segment determination are low. Phasing computationally without using the trio information performs worse in both class I and class II genes, demonstrating the effectiveness of using genotypic extension when trio-based phased data is not available.

6. CONCLUSION

We have developed a method for inferring HLA types using genotypic data by examining IBD shared segments with individuals of known HLA types. HLA types are predicted with high accuracy by using the CEU HapMap data. HLA genes play a critical role in adaptive immune response and autoimmune diseases. Our model can be used as a starting point in the analysis of similar diseases. The further applicability of SNP-based methods of HLA type inference has been described elsewhere (Leslie et al., 2008).

Although advances have been made in phasing, inferring haplotype structure in small cohorts or unrelated individuals remains a challenge (Marchini et al., 2006). Our method analyzes genotypic data without consulting the haplotype phases. This broadens the applicability of the method to data where the phase is unknown.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Breese, E., Braegger, C.P., Corrigan, C.J., et al. 1993. Interleukin-2- and interferon-gamma-secreting T cells in normal and diseased human intestinal mucosa. *Immunology* 78, 127–131.

Brimnes, J., Allez, M., Dotan, I., et al. 2005. Defects in $CD8^{+}$ regulatory T cells in the lamina propria of patients with inflammatory bowel disease. *J. Immunol.* 174, 5814–5822.

de Bakker, P.I., McVean, G., Sabeti, P.C., et al. 2006. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* 38, 1166–1172.

Gusev, A., Lowe, J.K., Stoffel, M., et al. 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* 19, 318–326.

Holdsworth, R., Hurley, C.K., Marsh, S.G., et al. 2009. The HLA dictionary 2008: a summary of HLA-A, -B, -C, DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens* 73, 95–170.

Horton, R., Wilming, L., Rand, V., et al. 2004. Gene map of the extended human MHC. *Nat. Rev. Genet.* 5, 889–899.

Leslie, S., Donnelly, P., and McVean, G. 2008. A statistical method for predicting classical HLA alleles from SNP data. *Am. J. Hum. Genet.* 82, 48–56.

Lincoln, M.R., Montpetit, A., Cader, M.Z., et al. 2005. A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis. *Nat. Genet.* 37, 1108–1112.

Marchini, J., Cutler, D., Patterson, N., et al. 2006. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 78, 437–450.

Miretti, M.M., Walsh, E.C., Ke, X., et al. 2005. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 76, 634–646.

Purcell, S., Neale, B., Todd-Brown, K., et al. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575.

Robinson, J., Waller, M.J., Parham, P., et al. 2003. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.* 31, 311–314.

Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644.

Address correspondence to:
Dr. Itsik Pe'er
Department of Computer Science Columbia University
506 Computer Science Buildling
500 West 120th Street
New York, NY 10027-7003

E-mail: itsik@cs.columbia.edu