ELSEVIER

Inferred HLA Haplotype Information for Donors From Hematopoietic Stem Cells Donor Registries

Pierre-Antoine Gourraud, Phillipe Lamiraux, Nabil El-Kadhi, Colette Raffoux, and Anne Cambon-Thomsen

ABSTRACT: Human leukocyte antigen (HLA) matching remains a key issue in the outcome of transplantation. In hematopoietic stem cell transplantation with unrelated donors, the matching for compatible donors is based on the HLA phenotype information. In familial transplantation, the matching is achieved at the haplotype level because donor and recipient share the block-transmitted major histocompatibility complex region. We present a statistical method based on the HLA haplotype inference to refine the HLA information available in an unrelated situation. We implement a systematic statistical inference of the haplotype combinations at the individual level. It computes the most likely haplotype pair given the phenotype and its probability. The method is validated on 301 phase-known phenotypes from CEPH families (Centre d'Etude du Polymorphisme Humain). The method is further applied to 85,933 HLA-A B DR typed unrelated donors from the French Registry of hematopoietic stem cells donors (France Greffe de Moëlle). The average value of prediction probability is 0.761 (SD 0.199) ranging from 0.26 to 1. Correlations between phenotype characteristics and predictions are also given. Homozygosity (OR = 2.08; [2.02–2.14] $p < 10^{-3}$) and linkage disequilibrium ($p < 10^{-3}$) are the major factors influencing the quality of prediction. Limits and relevance of the method are related to limits of haplotype estimation. Relevance of the method is discussed in the context of HLA matching refinement. *Human Immunology 66, 563–570 (2005).* © American Society for Histocompatibility and Immunogenetics, 2005. Published by Elsevier Inc.

KEYWORDS: Donor registry; HLA haplotypes; population immunogenetics; statistical application; transplantation

ABBREVIATIONS

BMD	bone marrow donor
CEPH	Centre d'Etude du Polymorphisme Human
HLA	human leukocyte antigen
HSC	hematopoietic stem cell
OR	odds ratio

INTRODUCTION

Allogeneic hematopoietic stem cell (HSC) transplantation is now a well-established curative therapy for an increasing number of hematologic diseases [1–3]. The role of human leukocyte antigen (HLA) matching between donor and recipient has been studied by many groups over the past years, but its optimal level remains unclear [4, 5]. The development of molecular typing techniques allows a refined matching and thus contributes to reduce risk of graft immunologic failure from host-versus-graft and graft-versus-host allorecognition.

The best donor remains an HLA-matched relative, but such a donor is not always available. In 70% of the cases, a search for an unrelated HLA-matched donor is performed among the 9.1 million bone marrow donors (BMDs) gathered in 54 stem cell donor registries from 40 countries and

From the Faculty of Medicine, INSERM, Toulouse, France (P.A.G., A.C.-T.), Laboratoire Epitech de Recherche en Informatique Appliqueé, Le Kremlin Bicêtre, France (P.L., N.E.-K.), and FGM France Greffe de Moëlle, French Registry of Hematopoïetic cells Donors, Paris, France (C.R.).

Address reprint requests to: Pierre-Antoine Gourraud, INSERM Unit 558, Faculty of Medicine, 37 allées Jules Guesde, F-31000, Toulouse, France; E-mail: gourraud@cict.fr.

Supported by EFG (Etablissement Français des Greffes) Grant 2003, EU contract MADO No: QLG7-CT-2001-00065.

Received December 10, 2004; revised January 3, 2005; accepted January 11, 2005.

0198-8859/05/$–see front matter
doi:10.1016/j.humimm.2005.01.011

37 cord blood registries from 21 countries from BMDs worldwide (http://www.bmdw.org) and the World Marrow Donor Association (http://www.worldmarrow.org/).

Nevertheless, the amount of HLA information taken into account is different. Indeed, through typing the patients' relatives, the actual level of HLA information used in familial HSC transplantation is the HLA haplotypes: matching is thus for two haplotypes (geno-identical situation) or only one (semi–haplo-identical situation) segregating in the family. In contrast, in unrelated situations, the haplotype information is known in the patient but not in the donor: that is, there is an asymmetry in the information available. This is usually solved by taking into account the most minimally shared information; namely, the phenotypic one.

The large content of the BMD registries enables the estimate of HLA frequencies in a given population [6–10]. The HLA population genetics data have always been a relevant field to apply maximum likelihood estimation of haplotype frequencies [11–13]. Because of the structure of the major histocompatibility complex region, such a method has successfully overcome the lack of phase information at an individual level to produce haplotype frequencies in populations. Besides their interest from the population point of view, we investigate here their possible use for the selection of unrelated donors from BMD registries in individual cases. The aim is to study how much population frequency information can be used to upgrade the donor information taken into account for the individual decision at haplotype level rather than downgrading the patient one at phenotype level. Knowing the genetic background of donors throughout the registries, we implemented a systematic statistical inference of haplotype pairs at the individual level. It computes the most likely haplotype pair given the phenotype and haplotype frequency information in the donors' population as additional information. Incomplete phenotype and use of HLA nomenclature is allowed.

Genetic properties influencing the accuracy of the prediction are discussed and may be of interest in genetic epidemiology as an example of individual haplotype inference procedure.

POPULATION AND METHODS

As reminded in the following sections, a diploid three contiguous locus phenotype can result in a maximum of four distinct phase configurations on the chromosomes.

$$\begin{array}{c} \textit{HLA Phenotype} \\ \textit{Locus A-Locus B-Locus DR} \\ (1,2)\text{-}(8,44)\text{-}(4,3) \end{array} \Rightarrow \begin{cases} 1\text{-}8\text{-}3,\ 2\text{-}44\text{-}4 \\ 1\text{-}44\text{-}3,\ 2\text{-}8\text{-}4 \\ 1\text{-}8\text{-}4,\ 2\text{-}44\text{-}3 \\ 1\text{-}44\text{-}4,\ 2\text{-}8\text{-}3 \end{cases} \begin{array}{l} \textit{Possible} \\ \textit{pairs of} \\ \textit{haplotype} \end{array}$$

For a K-ploid phenotype of R contiguous loci, n, the number of possible pairs of haplotypes is $n = K^{Hr-1}$, where the number of heterozygous loci is H_R. There is only one possible pair if only one locus is heterozygous. The proposed algorithm deals with this issue.

Algorithm

Given haplotype frequencies, the algorithm computes the likelihood for each possible phase. Then, it selects the one with the maximum value:

$$\begin{array}{c} \textit{HLA Phenotype} \\ \textit{Locus 1-locus 2-locus} \end{array} \Rightarrow \begin{cases} A\text{-}B\text{-}C,\ a\text{-}b\text{-}c \mid L_1 = 2 \times f_{ABC} \times f_{abc} \\ A\text{-}B\text{-}c,\ a\text{-}b\text{-}C \mid L_2 = 2 \times f_{ABc} \times f_{abC} \\ A\text{-}b\text{-}C,\ a\text{-}B\text{-}c \mid L_3 = 2 \times f_{AbC} \times f_{aBc} \\ A\text{-}b\text{-}c,\ a\text{-}B\text{-}C \mid L_4 = 2 \times f_{Abc} \times f_{aBC} \end{cases}$$

If the obtained pair of haplotypes is homozygous, the likelihood of such (unambiguous) pair is the squared value of the haplotype frequency estimation.

The probability p of the most likely pair of haplotypes is:

$$P = \frac{\max(L_i, i \in n * i \leq I)}{\sum_{i=1}^{I} L_i} \quad (1)$$

Where p is the prediction probability of the most likely haplotype pair; i is a natural integer used to enumerate the different haplotype pairs, Li is the likelihood of haplotype pair I as defined previously, given haplotype frequencies and Hardy-Weinberg equilibrium; and I is the overall number of possible haplotype pairs indexed by i.

The method ability to find the most likely haplotype pair is given by mean median (measure of central tendency) and percentiles (a value on a scale of 100 that indicates the percentage of the distribution of the phase prediction value that is equal to or below it) of the distribution of P probability defined in Equation 1 over the considered sample.

Several alternative estimations can be provided: