Human Immunology 74 (2013) 189–195

Contents lists available at SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

Analysis of the adequate size of a cord blood bank and comparison of HLA haplotype distributions between four populations

Katri Haimila *, Antti Penttilä [1], Anne Arvola, Marja-Kaisa Auvinen [2], Matti Korhonen

Finnish Red Cross Blood Service, Kivihaantie 7, FI-00310 Helsinki, Finland

ARTICLE INFO

Article history:
Received 27 February 2012
Accepted 26 October 2012
Available online 5 November 2012

ABSTRACT

The number of units and especially the number of different HLA haplotypes present in a cord blood (CB) bank is a crucial determinant of its usefulness. We generated data relevant to the development of our national CB in Finland. The HLA haplotype distribution was examined between specific populations. We developed graphical ways of data presentation that enable easy visualization of differences. First, we estimated the optimal size of a CB bank for Finland and found that approximately 1700 units are needed to provide a 5/6 HLA-matched donor for 80% of Finnish patients. Secondly, we evaluated HLA haplotype distributions between four locations, Finland, Japan, Sweden and Belgium. Our results showed that the Japanese Tokyo Cord Blood Bank differs in both the frequency and distribution of haplotypes from the European banks. The European banks (Finnish Cord Blood Registry, The Swedish National Cord Blood Bank, and Marrow Donor Program-Belgium) have similar frequencies of common haplotypes, but 26% of the haplotypes in the Finnish CB bank are unique, which justifies the existence of a national bank. The tendency to a homogenous HLA haplotype distribution in banks underlines the need for targeting recruitment at the poorly represented minority populations.

1. Introduction

Allogeneic hematopoietic cell transplantation is an effective treatment for selected patients with high-risk hematological disorders, specific primary immunodeficiencies, refractory autoimmune diseases or solid tumors. Although, the bone marrow donor registries include over 18 million volunteer donors worldwide, a suitable HLA-matched donor is found only for 34–70% of patients in need of a stem cell transplant, with ethnic minorities being in a particularly unfavorable situation [1]. Umbilical cord blood (CB) is a good alternative for a hematopoietic cell source, and already 500,000 CB units have been collected in public CB banks worldwide in order to expand the donor pool. Both the cell count and the HLA-match level of the CB graft have been shown to affect engraftment and survival. Survival rates improve with a higher match level [2], although CB requires less stringent HLA matching than bone marrow or peripheral blood stem cell transplantation because of the reduced risk of severe graft-versus-host disease (GVHD) in mismatched CB transplantation. It has been reported that transplantations with CB units containing mismatches only to graft-versus-host direction have as good an outcome as 6/6 HLA-matched CB units [3]. The often faced problem of too low a cell count of a CB unit can be overcome by transplanting two CB units at the same time, making the cell dose adequate also for adult patients [4,5].

The banking of high-quality units is expensive, and thus a balance between a large enough bank and reasonable costs must be determined. In the present study, our first aim was to delineate an optimal size for a national CB bank. The size, however, is not the only crucial characteristic of a bank, the diversity of HLA haplotypes must also be considered. Consequently, our second aim was to scrutinize the HLA heterogeneity of banked CB units by comparing three European and one Asian CB bank. We used several appropriate, sophisticated statistical methods producing visualized results.

2. Materials and methods

2.1. The HLA data

The Finnish CB bank contained 2910 units at the time of the study (2009) and the Finnish bone marrow donor registry had 18,624 potential donors. We compared the HLA haplotype distributions between these two registries and found no differences (data not shown). These two data sets were combined to form a

Abbreviations: CB, cord blood; GVHD, graft-versus-host disease; HLA, human leukocyte antigen; PCA, principle component analysis; TNC, total nucleated cells.

* Corresponding author: Fax: +358 9 5801495.
E-mail address: katri.haimila@bts.redcross.fi (K. Haimila).
[1] Present address: Department of Physics, P.O. Box 64, FI-00014 University of Helsinki, Finland.
[2] Present address: Clinical Immunology and Transfusion Medicine, Karolinska University Hospital, Huddinge C2:66, SE-14186 Stockholm, Sweden.

0198-8859/$36.00 - see front matter © 2012 American Society for Histocompatibility and Immunogenetics. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.humimm.2012.10.018

sample of 21,534 units of the total Finnish HLA haplotype distribution.

When calculating the probability of finding an HLA-matched donor, different levels of matching were used. In addition to the optimal 6/6 (in loci A, B and DRB1) match, matching levels of 4/6 and 5/6 were also accepted with one restriction: only one mismatch was allowed for HLA-DRB1 locus.

The Swedish National Cord Blood Bank, the Marrow Donor Program-Belgium (including data from four affiliated CB Banks CHU de Liege, UCL-St Luc, ULB-Institut Bordet and Navelstrengbloedbank UZ Gent) and the Tokyo Cord Blood Bank were chosen for the comparison of the HLA haplotype distributions between the populations. The Belgian and the Japanese CB banks were medium-sized national banks containing 6867 units (Belgium) and 4183 units (Japan) in April 2009. The first is located in the middle of a continent and the latter on an isolated island. The Swedish CB bank with 647 units (2009) was chosen for the study because Sweden is a neighboring country to Finland. Sweden and Finland have a common history going back for centuries, with migrations to both directions. The Swedish National Cord Blood Bank is under development and has a goal of collecting 50% of its units from donors of non-European descent. The Swedish bank comprised over 2000 units at the end of 2011.

The haplotypes containing three HLA loci (A, B and DRB1) were retrieved from the database of Bone Marrow Donors Worldwide (BMDW, www.bmdw.org) in April 2009. The haplotypes are composed by the maximum likelihood method [6]. All the units were HLA typed by DNA-based methods for class II but for class I, the typing was done serologically for over half of them (55%); therefore the haplotypes from the different banks were converted to a similar typing resolution to make the comparison possible between the populations. The resolution was kept as informative as possible, but for example the serological splits B60 and B61 had to be converted to B40 as well as B64 and B65 to B14 due to the large number of units typed at a serological broad level or with DNA-based methods. The serological split level was used for B62 and B63 because this information was available for most of the units and it is more informative than B15. Some rare haplotypes may have been missed because some units typed at a broad level (B15) were ignored.

2.2. Bootstrap resampling

The effect of the CB bank size on the probability of finding a suitable donor for a patient was studied with simulations based on the bootstrap resampling method (see, e.g., [7]). The principle behind bootstrap is to randomly sample the original data with replacement, thus forming other possible samples of the hypothetical population. We can draw both the samples of patients needing a transplant (of size k), and samples of possible CB bank (of size n). In practice, the value of k is the size of the combined HLA data, 21,534, and n can vary within reasonable CB bank sizes. Using values of n almost as large as or larger than the whole data (21,534 units) will introduce a positive bias to donor probabilities because the bootstrap sampling cannot produce novel HLA types to the sample, but only resample the present types. This feature of the bootstrap method is not a problem in our study, at least not for the 4/6 and 5/6 matches, since the simulated CB bank sizes for these cases remain below 3% and 50% of the data size, respectively.

With every simulation round i, a patient group G_i (of size k) and a simulated CB bank C_i of size n_i are drawn. For every patient in G_i a 4/6, 5/6 and 6/6 match in C_i is searched for. When all the patients in G_i have been processed we can compute the probability of finding a 4/6, 5/6 and 6/6 donor for simulation round i and for a CB bank of size n_i. For every n_i we must run several simulations (4000 in this study) and average to get a reliable estimate of the probabilities.

2.3. Uniqueness

To examine the uniqueness of the haplotypes in the populations, the measure U is defined as the number of haplotypes that are only found in the corresponding population A, divided by the total number of different haplotypes in A. More formally,

$$U_A = \frac{|A \setminus D_A|}{|A|}, \tag{1}$$

where $|\ |$ is the size (i.e., number of elements), and the operator $\setminus$ is the relative complement of D_A in A. The D_A is the set of all the haplotypes in the other populations except A. If the population A consists totally of unique haplotypes, U will be one (or 100%). If all the haplotypes in A are present in the other populations, U will be zero (0%).

2.4. G-test

The G-test is used for testing whether the sample fits a given (discrete) distribution [8,9]. The G-test is an exact likelihood-ratio test, whereas the alternative chi-square test is an approximation. Particularly in cases where $|o_i - e_i| < e_i$ (o_i is the observed and e_i the expected frequency) for any i, the G-test is recommended over the chi-square. The test statistic G is calculated as

$$G = 2\sum_i o_i \ln(o_i/e_i) \tag{2}$$

Under the null hypothesis that the observed frequencies follow the distribution of the expected frequencies, G should follow the chi-square distribution with n-p degrees of freedom where n is the number of classes i and p the number of estimated parameters for the expected distribution of e_i's. If we want to test against the discrete uniform distribution, we must choose $e_i = N/n$ for all i, where N is the total number of observations, and $p = 1$.

2.5. Similarity coefficients

Similarity coefficients are used to compare populations in, e.g., ecology or data mining applications. We applied these to the haplotype distribution comparison. Several different similarity coefficients exist in the literature of statistical methods. Suitable coefficients for our purposes include the Sørensen similarity coefficient [10,11] for categorical data, and the cosine similarity coefficient [12,13] for frequency data. The Sørensen similarity coefficient Q is defined between sets (i.e., haplotypes in population) A and B as

$$Q_{A,B} = \frac{2|A \cap B|}{|A| + |B|}, \tag{3}$$

where $\cap$ is the intersection of two sets. The Q is always between zero (or 0%, no similarity) and one (or 100%, full similarity).

The cosine similarity coefficient C between frequency vectors A and B (i.e., population haplotype frequency distributions) is defined as the cosine of the angle between the vectors:

$$C_{A,B} = \frac{A \cdot B}{\|A\|\|B\|}, \tag{4}$$

where $\cdot$ is the dot product between the vectors, and $\|\ \|$ is the norm of a vector. The C is always between -1 and 1, but with frequency vectors where all the elements are positive, as in our case, C is between 0 and 1. The value zero (or 0%) indicates dissimilarity while the value one (or 100%) indicates total similarity.

2.6. Principle component analysis

The principle component analysis (PCA) [14] is a popular method in multivariate statistics for variable reduction in large datasets.