Human leukocyte antigen (HLA) A-C-B-DRB1-DQB1 haplotype segregation analysis among 2152 families in China and the comparison to expectation-maximization algorithm result

Yang Li[1,2], Lu-Yao Chen[1,2], Teng-Teng Zhang[1,2], Xiao-Ni Yuan[1,2], Xiao-Jing Bao[1,2], Jun He[1,2]

[1]Department of HLA Laboratory, Jiangsu Institute of Hematology, First Affiliated Hospital of Soochow University, Suzhou, Jiangsu 215031, China;
[2]Collaborative Innovation Center of Hematology, Soochow University, Suzhou, Jiangsu 215006, China.

Human leukocyte antigen (HLA) is the major histocompatibility complex for humans. Previous studies have shown that high-resolution HLA matching can reduce graft *vs*. host disease and improve the outcome of hematopoietic stem cell transplantation (HSCT). Unrelated donor HSCT is very important when patients have no HLA-identical sibling donors. However, the chance of finding an available unrelated donor is not equal for every patient. Previous studies in Western countries have found that the HLA haplotype frequency (HF) could help to predict the probability of identifying HLA allele-matched unrelated donors.[1] However, the HLA system shows ethnic diversity and regional disparity. Therefore, if we want to use the HLA haplotype tool in China, a reliable HLA haplotype database for the Chinese population needs to be established.

A previous study in a Japanese population showed that unrelated individual data were similar to family data.[2] However, family studies are a reliable way to identify rare haplotypes.[3] This study was a family study of five-locus HLA A-C-B-DRB1-DQB1 high-resolution haplotypes, with a very large sample size in China to date. In addition, we verified the accordance between expected HFs calculated by the expectation-maximization (EM) algorithm from unrelated individuals and observed HFs by segregation analysis (direct counting) in families to find a suitable method for haplotype database setup.

A total of 2152 families, all four haplotypes (a, b, c, and d) presenting and confirmed by descent, were included in this study. They were divided into three groups: (1) Families with parents (n=1531); (2) Families with one parent and one or more siblings (n=175); and (3) Families without parents, but the haplotypes (a, b, c, and d) were checked by two or more siblings (n=446). Among these families, 1907, 173, and 72 were from the East China, Central China, and South China areas, respectively. According to doctors' typing applications and laboratory standard operation procedure, sequence-based typing plus sequence-specific oligonucleotide probe methods were performed for high-resolution HLA-A, B, C, DRB1, and DQB1 typing for all patients. Additional tests were performed to resolve ambiguities. Genomic DNA was extracted from peripheral blood. The study was reviewed by the ethics committee of our hospital (No. 2020-322). Informed consent was obtained from all the participants when they submitted a sample for HLA typing.

First, observed HFs for the 2152 families were calculated by segregation analysis by using Arlequin software version 3.5.2.2 (http://www.cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html). Only four haplotypes (a, b, c, and d) were counted for each family to avoid repetition. In Supplementary Table 1, http://links.lww.com/CM9/A520, a total of 3274 five-locus A-C-B-DRB1-DQB1 haplotypes were observed. Only 285 haplotypes were common, and most were less common or even rare. The thresholds of common HLA haplotypes were not agreed upon in different studies,[4] so we defined HF ≥0.1% as a common HLA haplotype, referring to the acknowledged threshold of the common HLA allele.[5] In Table 1, the top 20 segregation analysis results in our study were compared to unrelated individual results (EM algorithm) reported,[6] and no statistically significant differences were found (P values were all >0.5). Our data were very similar to the results of East China reported[6] because a majority of the

---

Correspondence to: Dr. Jun He, Department of HLA Laboratory, Jiangsu Institute of Hematology, First Affiliated Hospital of Soochow University, 13/f (West), Hospital Comprehensive Building, 899 Ping Hai Road, Suzhou 215031, Jiangsu, China
E-Mail: junhe1964@163.com

Table 1: Comparison of the top 20 haplotypes between our family study and an unrelated individual study reported in China.

HLA A-C-B-DRB1-DQB1 haplotype					Rank and Haplotype frequency					
					Our family study (4n = 8608)		The unrelated study reported[6]			
							China		East China	
A*	C*	B*	DRB1*	DQB1*	Rank[†]	HF (%)	Rank	HF (%)	Rank	HF (%)
30:01	06:02	13:02	07:01	02:02	1	4.98	1	3.70	1	4.50
02:07	01:02	46:01	09:01	03:03	2	3.18	2	2.46	3	2.56
33:03	03:02	58:01	03:01	02:01	3	2.83	3	2.40	2	2.89
33:03	03:02	58:01	13:02	06:09	4	1.45	5	1.06	4	1.43
11:01	08:01	15:02	12:02	03:01	5	1.30	4	1.13	5	1.01
02:07	01:02	46:01	08:03	06:01	6	0.94	6	0.93	7	0.92
33:03	14:03	44:03	13:02	06:04	6	0.94	7	0.74	6	0.95
11:01	04:01	15:01	04:06	03:02	8	0.86	12	0.56	8	0.60
02:01	03:04	13:01	12:02	03:01	9	0.60	10	0.58	9	0.58
02:01	03:03	15:11	09:01	03:03	9	0.60	22	0.36	18	0.40
11:01	03:04	13:01	15:01	06:01	11	0.58	9	0.64	11	0.56
01:01	06:02	57:01	07:01	03:03	12	0.53	14	0.45	16	0.46
24:02	01:02	54:01	04:05	04:01	13	0.51	16	0.44	14	0.47
11:01	07:02	40:01	08:03	06:01	14	0.48	17	0.42	13	0.49
24:02	14:02	51:01	09:01	03:03	14	0.48	26	0.31	19	0.35
11:01	01:02	46:01	09:01	03:03	16	0.46	11	0.58	10	0.56
11:01	07:02	40:01	09:01	03:03	17	0.45	21	0.36	15	0.46
33:03	07:06	44:03	07:01	02:02	18	0.43	13	0.47	17	0.45
01:01	06:02	37:01	10:01	05:01	19	0.39	8	0.66	12	0.56
11:01	14:02	51:01	09:01	03:03	19	0.39	25	0.31	20	0.34
24:02	03:04	13:01	12:02	03:01	19	0.39	>20[‡]	–	>20[‡]	–

[†] The same ranks were assigned to different haplotypes if their HFs were equal. [‡] >20 means that the ranks and frequencies of some haplotypes outside of the top 20 list of all the areas showed in study reported.[6] HF: Haplotype frequency; –: No data.

families in our study were living in this area. The seven most common haplotypes showed the best accordance.

Second, the same data of the first part were used for allele frequency (AF) estimation and pairwise linkage disequilibrium (LD) by using Arlequin. HFs were compared to AFs, and they were not always completely positively associated. Supplementary Table 2, http://links.lww.com/CM9/A520 shows the top 20 haplotypes and the comparison with alleles. For example, A*30:01-C*06:02-B*13:02-DRB1*07:01-DQB1*02:02 was the most frequent haplotype, but the ranks of the A, C, B, DRB1, and DQB1 alleles were 6, 3, 3, 3, and 4, respectively. AFs were not as frequent as HF because of the strong positive association between each allele, and the D′ values of the LD test were 0.88, 0.87, 0.71, 0.71, 0.98, 0.83, 0.82, 0.70, 0.73, and 1.00 for A-C, A-B, A-DRB1, A-DQB1, B-C, B-DRB1, B-DQB1, C-DRB1, C-DQB1, and DR-DQB1, respectively. In another example, the ranks of A*11:01-C*01:02-B*46:01-DRB1*09:01-DQB1*03:03 and A*11:01-C*07:02-B*40:01-DRB1*09:01-DQB1*03:03 were 16 and 17, respectively. However, the ranks were 1 or 2 for each allele. This is because the positive association was not very strong, with even a negative association for some two-locus haplotypes. In Supplementary Table 3, http://links.lww.com/CM9/A520, 11 A-B, 5 A-C, 27 B-C, 23 DRB1-DQB1, 3 A-DRB1, 5 B-DRB1, 4 C-DRB1, 2 A-DQB1, and 4 C-DQB1 two-locus haplotypes show strong positive associations (HF ≥ 0.1%, D′ > 0.5, r^2 > 0.1). A*30:01-C*06:02-B*13:02, A*02:07-C*01:02-B*46:01, A*33:03-

C*03:02-B*58:01, A*29:01-C*15:05-B*07:05, and A*69:01-C*12:02-B*52:01 three-locus haplotypes show very strong linkages, and they accounted for proportions of 6.02%, 5.63%, 5.30%, 0.49%, and 0.34%, respectively.

Third, the patients' typing results were used as phase-known and phase-unknown data to obtain observed and expected HFs, respectively, by the direct counting EM algorithm (only HFs $> 1 \times 10^{-5}$ were outputted by Arlequin). In Supplementary Table 4, http://links.lww.com/CM9/A520, a total of 2050 observed and 1852 expected haplotypes were obtained, and 1228 haplotypes overlapped. The remaining 822 observed and 624 expected haplotypes did not overlap, and their HFs were all less than 0.1%. Among the 1228 overlapping haplotypes, less-common haplotypes were more common, but the numbers of common and less-common haplotypes were not equal between the observed and expected groups. Because 17 commonly observed haplotypes were less common in the expected group, 41 less commonly observed haplotypes were common in the expected group. Therefore, observed HF was the common and less-common threshold for the chi-square test. The Chi-square test for trend was carried out for the overlapping haplotypes by using GraphPad Prism 6 software. The P values of the Chi-square test showed that there were no statistically significant differences between observed and expected haplotypes, not only in total overlapping data ($P = 0.2424$) and common data (HF ≥ 0.1%, $P = 0.3698$) but also in less-common data (HF < 0.1%, $P = 0.1582$).

Therefore, the tendencies of observed and expected haplotypes are coincident, and the tendency concordance of common data is the best; then, the total overlapping data last the less common data. However, the family segregation analysis found 822 haplotypes that were missed by EM, and more importantly, 624 haplotypes were incorrectly built using EM. These 822 haplotypes found from the family segregation analysis are real because these are observed haplotypes from the segregation. AFs and pairwise LDs are two important factors for these phenomenon, which can be proved by Supplementary Tables 2 and 5, http://links.lww.com/CM9/A520. For example, A*02:07-C*03:04-B*40:01-DRB1*10:01-DQB1*05:01 was the incorrect haplotype built by EM. All the constituent alleles were common and should be observed easily. However, it had eight negative pairwise LDs, so it would be hard to present.

Unrelated data are easier to obtain than family data, and good consistency of tendency between expected and observed haplotypes was the basis using unrelated related study (EM) for HLA haplotype database setup. However, EM could miss less-common real haplotypes but built incorrectly less-common haplotypes. Therefore, identifying less common haplotypes from segregation analysis must be used for checking and as supplements to cover the shortage of EM.

The HLA haplotype tool will be very useful in both unrelated-HSCT and haplo-HSCT fields.

In unrelated HSCT, it can be helpful to predict the possibility of finding an HLA allele-matching unrelated donor and the possible mismatching alleles. Patients with common haplotypes can find unrelated HLA matching donors more easily. The chance for patients will decline along with a decline in HFs. Some patients with common HLA alleles also have difficulty finding HLA allele-matching unrelated donors because of the less common haplotypes caused by strong negative LD. In addition, if patients or donors only have A, B, and DRB1 typing results, it can help to predict C and DQB1 results. The prediction can help clinicians choose several suitable donors at the HLA confirmatory typing stage.

In haplo-HSCT, for some families with patients and only one sibling, the donor may be 5/10 allele match, but he/she may not be real 1-haplo-match; for some families with patients and one parent/child, the donor is 10/10 allele match, but he/she is still 1-haplo-match; for some families

with patients and one sibling, the donor is 10/10 allele match, and he/she may not be a sib-match but is still 1-haplo-match. Before complete family data are obtained, these data can help clinicians to predict whether the 10/10 or 5/10 allele match sibling is truly 2-haplo-match or 1-haplo-match and take appropriate treatment. Usually, the more common the haplotype is, the more reliable it is.

Conflicts of interest

None.

References

1. Buhler S, Baldomero H, Ferrari-Lacraz S, Nunes JM, Sanchez-Mazas A, Massouridi-Levrat S, et al. High-resolution HLA phased haplotype frequencies to predict the success of unrelated donor searches and clinical outcome following hematopoietic stem cell transplantation. Bone Marrow Transplant 2019;54:1701–1709. doi: 10.1038/s41409-019-0520-6.
2. Ikeda N, Kojima H, Nishikawa M, Hayashi K, Futagami T, Tsujino T, et al. Determination of HLA-A, -C, -B, -DRB1 allele and haplotype frequency in Japanese population based on family study. Tissue Antigens 2015;85:252–259. doi: 10.1111/tan.12536.
3. Osoegawa K, Mallempati KC, Gangavarapu S, Oki A, Gendzekhadze K, Marino SR, et al. HLA alleles and haplotypes observed in 263 US families. Hum Immunol 2019;80:644–660. doi: 10.1016/j.humimm.2019.05.018.
4. Pedron B, Duval M, Elbou OM, Moskwa M, Jambou M, Vilmer E, et al. Common genomic HLA haplotypes contributing to successful donor search in unrelated hematopoietic transplantation. Bone Marrow Transplant 2003;31:423–427. doi: 10.1038/sj.bmt.1703876.
5. Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, et al. Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue Antigens 2013;81:194–203. doi: 10.1111/tan.12093.
6. Zhou X-Y, Zhu F-M, Li J-P, Mao W, Zhang D-M, Liu M-L, et al. High-resolution analyses of human leukocyte antigens allele and haplotype frequencies based on 169,995 volunteers from the China Bone Marrow Donor Registry Program. PLoS One 2015;10: e0139485. doi: 10.1371/journal.pone.0139485.