Received: 16 March 2021 | Revised: 28 May 2021 | Accepted: 8 July 2021

DOI: 10.1111/iji.12553

SHORT COMMUNICATION

INTERNATIONAL JOURNAL OF IMMUNOGENETICS WILEY

Estimating HLA haplotype frequencies from homozygous individuals – A Technical Report

Susanne Seitz[1] | Vinzenz Lange[2] | Paul J. Norman[3] | Jürgen Sauter[1] | Alexander H. Schmidt[1,2]

[1] DKMS, Tübingen, Germany

[2] DKMS Life Science Lab, Dresden, Germany

[3] Division of Biomedical Informatics and Personalized Medicine, Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

Correspondence
Susanne Seitz, DKMS, Kressbach 1, 72072 Tübingen, Germany.
Email: seitz@dkms.de

Linked Letters: Nunes, J. M. et al. *International Journal of Immunogenetic* 2021; https://doi.org/10.1111/iji.12555 and Seitz, S., et al. *International Journal of Immunogenetic* 2021; https://doi.org/10.1111/iji.12556

Funding information
National Institutes of Health, Grant/Award Number: NIH U01 AI090905

[Correction added on 21 April 2022, after first online publication: The funding information was added.]

Abstract

We estimated HLA haplotype frequencies based on individuals homozygous for 4, 5 or 6 loci. Validation of our approach using a sample of over 3.4 million German individuals was successful. Compared to an expectation-maximization algorithm, the errors were larger. However, our approach allows the unequivocal detection of rare haplotypes.

KEYWORDS
allele frequency, donor registry, haplotype frequency, HLA, homozygosity

1 | INTRODUCTION

Population-specific human leucocyte antigen (HLA) haplotype frequencies (HF) are of particular importance in the context of stem cell donor registries in two ways. First, they allow the determination of the probability that an incompletely typed donor is a complete match for a given patient. This is used in modern donor search algorithms that sort donors based on these probabilities (Bochtler et al., 2016; Dehn et al., 2016; Steiner, 2012; Urban et al., 2020). Second, HLA HF can be used to determine the proportion of patients of given ethnicity for which a matching stem cell donor can be found in a donor pool of defined size and ethnic composition. Such analyses are of great relevance for strategic planning of donor recruitment activities (Alfraih et al., 2021; Bergstrom et al., 2009; Schmidt et al., 2014; Sonnenberg et al., 1989).

The determination of HLA haplotypes in the laboratory has been described (Z. Guo et al., 2006), but is still relatively costly and not very commonly used. In particular, it is also not currently applied for HLA typing of newly registered potential stem cell donors, which is of course highly cost-sensitive (Schmidt et al., 2020). Generally, the haplotypes of registered donors cannot be determined by pedigree analysis (Becker & Knapp, 2002; Ikeda et al., 2015) either, since donor registries usually do not have information on respective family relationships. We have attempted to determine HF from presumed family relationships of registered stem cell donors (based on address, last name and age difference) and presented corresponding preliminary results (Sauter et al., 2015). However, due to methodological difficulties, we have not further pursued that approach so far. Taken together, the HLA haplotypes of individuals enrolled in a donor registry are not known

Abbreviations: AF, allele frequency; ARD, antigen recognition domain; EM, expectation-maximization; HF, haplotype frequency; HLA, human leucocyte antigen; HSCT, hematopoietic stem cell transplantation; HWE, Hardy–Weinberg equilibrium; NGS, next-generation sequencing with certainty and therefore HF determination by simple counting is not possible.

Even if the individual HLA haplotypes are generally not known, there are methods to determine population-specific HF based on sufficiently large samples. Typically, HF from donor registry data are estimated by maximum likelihood methods via an expectation-maximization (EM) algorithm (Dempster et al., 1977; Excoffier & Slatkin, 1995). The most widely used tool for this purpose is probably the Arlequin software package (Excoffier & Lischer, 2010). Our group has developed the freely downloadable Hapl-o-Mat software (Sauter et al., 2018; Schäfer et al., 2017), which includes special adaptations to the characteristics of donor registry data (large samples, heterogeneous resolution and missing loci).

Despite these options, we have been looking for a way to also determine HF in a direct way, that is, by simple counting if possible. The large number of donors registered with DKMS and typed in high resolution gives us the chance to do this. It makes it possible to obtain quite large samples of completely HLA homozygous individuals. Since the two identical haplotypes of these donors are known, the haplotypes of the homozygous subset can simply be counted. The HF of the underlying larger sample are then obtained by a simple mathematical transformation. The aim of the present study was to find out whether useful HF can be obtained with this simple method and whether they may even be superior to the HF obtained with an EM algorithm.

2 | METHODS

Our data set comprised $N = 3{,}456{,}066$ unrelated donors with self-reported German descent who were recruited by DKMS Germany from 2013 to 2019. Upon registration, donors provided informed consent including the processing of donor data for scientific studies. HLA typing of new registrants was performed for the loci HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1 and HLA-DPB1 at the EFI- and ASHI-accredited DKMS Life Science Lab (Dresden, Germany). The amplicon-based next-generation sequencing (NGS) workflow using Illumina MiSeq, HiSeq or NovaSeq devices (Lange et al., 2014; Schöfl et al., 2017) targeted the exons coding for the antigen recognition domain (ARD; exons 2 and 3 for class I and exon 2 for class II genes). For the analyses in this work, alleles with identical or synonymous DNA sequences with regard to these exons (coding for the ARD) were grouped together ('g' nomenclature; Schmidt et al., 2009). In contrast to the P-nomenclature, this includes null alleles defined by mutations outside the ARD. Only donors with unambiguous g-level alleles were considered for analyses.

The number of fully homozygous donors depended on the HLA loci taken into account. When considering four loci, HLA-A, -B, -C and -DRB1, $n_4 = 26{,}311$ donors were completely homozygous (0.76% of the original sample size). Addition of HLA-DQB1 reduced the number of fully homozygous donors only slightly to $n_5 = 25{,}433$ (0.74%). In the 6-locus scenario including HLA-DPB1, only $n_6 = 9{,}217$ donors were completely homozygous (0.27%). The much smaller sample size in that scenario apparently results from the weak linkage disequilibrium between HLA-DPB1 and the other HLA loci considered (Sanchez-Mazas et al., 2000).

To estimate HF with the method based on completely homozygous donors, we first determined the HF in the subset of homozygous donors: The frequency f_i^* of a haplotype i in a homozygous donor subset can be obtained simply by counting all individuals homozygous for haplotype i and division by the subset size, that is, by the number of all individuals homozygous for any haplotype. We then estimated the HF f_i of the original sample ($N = 3{,}456{,}066$) from the subset frequencies f_i^* previously obtained. In doing so, we made use of the fact that, under the assumption of Hardy–Weinberg Equilibrium (HWE), $f_i^* = f_i^2 / \sum_j f_j^2$ holds. This equation again includes the aforementioned quotient of individuals homozygous for a particular haplotype i and all homozygous individuals; only now we do not consider the observed numbers of the homozygous subset but the unknown frequencies of the original sample. We then obtain $f_i = \sqrt{\sum_j f_j^2}\sqrt{f_i^*}$ by elementary transformations. As the frequencies f_i^* have already been determined, the HF f_i we are looking for are now known except for a constant factor. (The factor $C = \sqrt{\sum_j f_j^2}$ is constant because all frequencies f_j are unknown but fixed. For other samples, of course, other constant factors would be obtained). As a last step, we normalize the frequencies f_i so that $\sum_i f_i = 1$ holds.

If the original sample includes a specific haplotype k but no homozygous individual with that haplotype, then our approach leads to the estimation $f_k = 0$ because the observed subset frequency f_k^* also equals 0. As this will normally happen for many haplotypes of a given sample, one may argue that the condition $\sum_i f_i = 1$ is not reasonable as cumulated frequency is inevitably 'lost.' However, it should also be noted that we considerably overestimate the frequency of a rare haplotype, for which there accidentally happens to be one or more homozygous individuals in the original sample. This is also likely to occur occasionally and contributes to the rather distinct step structure of the HF curve generated by our approach that will be discussed later. It is not clear a priori which of the two described opposing effects leading to frequency under- or overestimation of rare haplotypes dominates and to which extent. Therefore, it seems justified to use the common frequency normalization given by $\sum_i f_i = 1$. We will see in Section 3 that the good accordance of the HF obtained from the new method based on homozygous individuals with the results from the EM algorithm empirically confirms this decision, at least for a broad range of homozygous subset sizes.

For comparison, we also estimated the HF using the EM algorithm implemented in the Hapl-o-Mat software. The threshold for stopping the algorithm was set to $\varepsilon = 10^{-8}$.

For HF estimation with the EM algorithm, the underlying data set is required to be in HWE. Deviations from HWE were calculated per locus by determining deviations between observed and expected homozygosity and by performing an exact test using a Markov chain (S. W. Guo & Thompson, 1992) implemented in Arlequin 3.5.2.2 (Excoffier & Lischer, 2010). Parameters were set to 10^6 Markov chain steps and 10^5 dememorization steps. For the exact test, p-values <0.05 were regarded as indicative for a significant deviation from HWE.

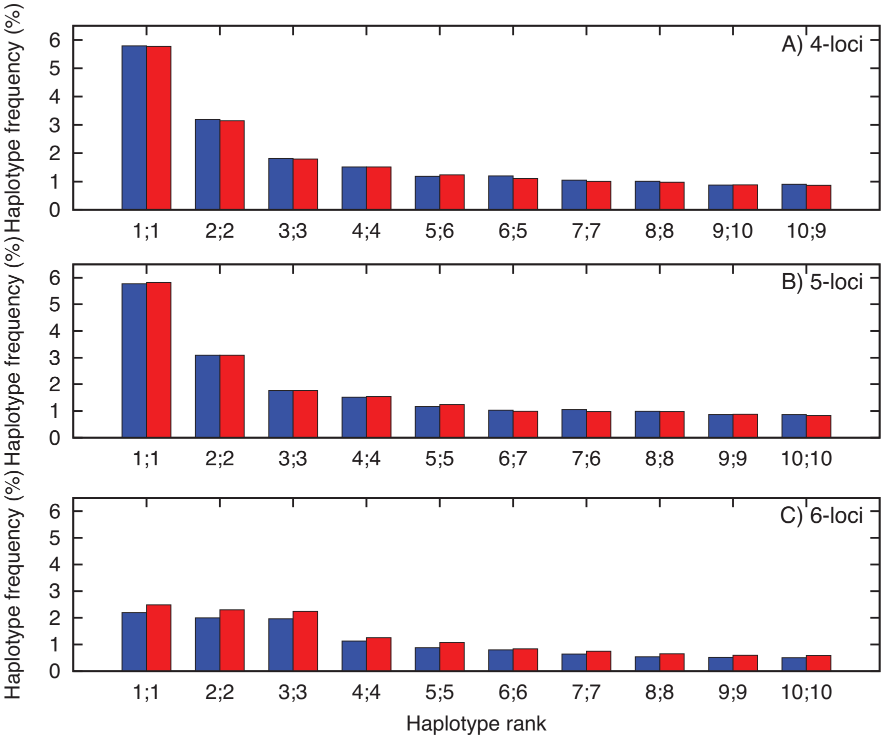

FIGURE 1 Frequencies of the 10 most frequent 4-, 5- and 6-locus haplotypes. Blue: Hapl-o-Mat; $N = 3{,}456{,}066$. Red: homozygous donors. (a) 4 loci (HLA-A, -B, -C, -DRB1); size of homozygous subset: $n_4 = 26{,}311$. (b) 5 loci (HLA-A, -B, -C, -DRB1, -DQB1); $n_5 = 25{,}433$. (c) 6 loci (HLA-A, -B, -C, -DRB1, -DQB1, -DPB1); $n_6 = 9{,}217$

3 | RESULTS

We detected no significant deviations from HWE (Table 1), which allowed the use of the EM algorithm and the approach based on homozygous donors. For common haplotypes, the frequencies determined by both methods showed a good accordance. The 10 most common haplotypes were identical in all scenarios (4, 5 or 6 loci; Figure 1, Table 2). The order of these top 10 haplotypes was the same for both HF estimation methods when 6-locus haplotypes were considered. The 5-locus haplotypes showed one rank discrepancy (ranks 6 and 7 switched), and the 4-locus haplotypes showed two discrepancies (ranks 5/6 and 9/10 switched). Among the 10 most frequent haplotypes, the mean deviation between the HF derived from the homozygous sample and those estimated with Hapl-o-Mat was 2.9% in the 4-locus scenario. The corresponding values for the 5- and 6-locus scenarios were 2.7% and 15.1%, respectively. Not only did the 6-locus scenario have by far the highest mean deviation, it was also the only scenario in which the HF obtained with one method (homozygous donors) were consistently higher than the HF obtained with the other method (Hapl-o-Mat). We consider the much smaller sample size in the 6-locus scenario to be the most likely reason for the larger deviation from the results obtained with the EM algorithm. This assumption is supported by Figure S1 where we downsized the subset of the 5-locus scenario ($n_5 = 25{,}433$) to match the sample size of the 6-locus homozygous subset ($n_{5\text{Small}} = n_6 = 9{,}217$). The mean deviation from the 10 highest HF obtained using Hapl-o-Mat was 21.9%, compared to 2.7% in the original 5-locus scenario, with all HF obtained from the downsized subset being higher than those from the EM algorithm.

TABLE 1 Deviations from HWE per locus for the full sample ($N = 3{,}456{,}066$)

HLA locus	*p*-value (exact test)	Observed homozygosity	Expected homozygosity
A	0.945	0.147	0.145
B	0.125	0.064	0.059
C	0.734	0.095	0.090
DRB1	0.711	0.082	0.079
DQB1	0.371	0.132	0.130
DPB1	0.440	0.232	0.230

Abbreviation: HLA, human leucocyte antigen.

The EM algorithm also struggles with decreasing sample sizes, as shown in Figure S2. Here, we downsized the Hapl-o-Mat input data for 6 loci from $N = 3{,}456{,}066$ to match the sample size of the 6-locus homozygous donors ($n_6 = 9{,}217$). The mean deviation from the 10 highest HF obtained using the original dataset ($N = 3{,}456{,}066$) was 20.6% and thus even larger than with the approach based on homozygous individuals (15.1%). In the further course of the HF curve, however, the characteristic step formation (see below) occurs significantly later than in the approach using homozygous donors.

Albeit we observed good accordance for high HF, accuracy decreased with lower frequency haplotypes (Figure 2). This is due to the limited number of different haplotypes in the homozygous samples,

TABLE 2 10 most frequent haplotypes in the 4-locus, 5-locus and 6-locus scenarios with the Hapl-o-Mat software and the approach based on homozygous donors

		Hapl-o-Mat		Homozygous donors	
	Rank	Haplotype	Frequency (%)	Haplotype	Frequency (%)
4 loci	1	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g	5.795	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g	5.769
	2	A*03:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g	3.189	A*03:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g	3.145
	3	A*02:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g	1.813	A*02:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g	1.794
	4	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g	1.517	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g	1.515
	5	A*02:01 g~B*44:02 g~C*05:01 g~DRB1*04:01 g	1.200	A*02:01 g~B*15:01 g~C*03:04 g~DRB1*04:01 g	1.237
	6	A*02:01 g~B*15:01 g~C*03:04 g~DRB1*04:01 g	1.182	A*02:01 g~B*44:02 g~C*05:01 g~DRB1*04:01 g	1.103
	7	A*29:02 g~B*44:03 g~C*16:01 g~DRB1*07:01 g	1.052	A*29:02 g~B*44:03 g~C*16:01 g~DRB1*07:01 g	1.003
	8	A*02:01 g~B*40:01 g~C*03:04 g~DRB1*13:02 g	1.006	A*02:01 g~B*40:01 g~C*03:04 g~DRB1*13:02 g	0.975
	9	A*01:01 g~B*57:01 g~C*06:02 g~DRB1*07:01 g	0.902	A*02:01 g~B*13:02 g~C*06:02 g~DRB1*07:01 g	0.884
	10	A*02:01 g~B*13:02 g~C*06:02 g~DRB1*07:01 g	0.876	A*01:01 g~B*57:01 g~C*06:02 g~DRB1*07:01 g	0.865
5 loci	1	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g~DQB1*02:01 g	5.769	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g~DQB1*02:01 g	5.813
	2	A*03:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g	3.096	A*03:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g	3.095
	3	A*02:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g	1.765	A*02:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g	1.775
	4	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g~DQB1*05:01 g	1.518	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g~DQB1*05:01 g	1.534
	5	A*02:01 g~B*15:01 g~C*03:04 g~DRB1*04:01 g~DQB1*03:02 g	1.165	A*02:01 g~B*15:01 g~C*03:04 g~DRB1*04:01 g~DQB1*03:02 g	1.235
	6	A*02:01 g~B*44:02 g~C*05:01 g~DRB1*04:01 g~DQB1*03:01 g	1.051	A*29:02 g~B*44:03 g~C*16:01 g~DRB1*07:01 g~DQB1*02:01 g	0.991
	7	A*29:02 g~B*44:03 g~C*16:01 g~DRB1*07:01 g~DQB1*02:01 g	1.031	A*02:01 g~B*44:02 g~C*05:01 g~DRB1*04:01 g~DQB1*03:01 g	0.976
	8	A*02:01 g~B*40:01 g~C*03:04 g~DRB1*13:02 g~DQB1*06:04 g	0.993	A*02:01 g~B*40:01 g~C*03:04 g~DRB1*13:02 g~DQB1*06:04 g	0.975
	9	A*02:01 g~B*13:02 g~C*06:02 g~DRB1*07:01 g~DQB1*02:01 g	0.864	A*02:01 g~B*13:02 g~C*06:02 g~DRB1*07:01 g~DQB1*02:01 g	0.883
	10	A*01:01 g~B*57:01 g~C*06:02 g~DRB1*07:01 g~DQB1*03:03 g	0.858	A*01:01 g~B*57:01 g~C*06:02 g~DRB1*07:01 g~DQB1*03:03 g	0.828
6 loci	1	A*03:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g~DPB1*04:01 g	2.199	A*03:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g~DPB1*04:01 g	2.484
	2	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g~DQB1*02:01 g~DPB1*01:01 g	1.995	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g~DQB1*02:01 g~DPB1*01:01 g	2.302
	3	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g~DQB1*02:01 g~DPB1*04:01 g	1.965	A*01:01 g~B*08:01 g~C*07:01 g~DRB1*03:01 g~DQB1*02:01 g~DPB1*04:01 g	2.245
	4	A*02:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g~DPB1*04:01 g	1.128	A*02:01 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g~DPB1*04:01 g	1.257
	5	A*02:01 g~B*15:01 g~C*03:04 g~DRB1*04:01 g~DQB1*03:02 g~DPB1*04:01 g	0.880	A*02:01 g~B*15:01 g~C*03:04 g~DRB1*04:01 g~DQB1*03:02 g~DPB1*04:01 g	1.076
	6	A*02:01 g~B*44:02 g~C*05:01 g~DRB1*04:01 g~DQB1*03:01 g~DPB1*04:01 g	0.794	A*02:01 g~B*44:02 g~C*05:01 g~DRB1*04:01 g~DQB1*03:01 g~DPB1*04:01 g	0.838
	7	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g~DQB1*05:01 g~DPB1*04:02 g	0.637	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g~DQB1*05:01 g~DPB1*04:02 g	0.747
	8	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g~DQB1*05:01 g~DPB1*04:01 g	0.540	A*03:01 g~B*35:01 g~C*04:01 g~DRB1*01:01 g~DQB1*05:01 g~DPB1*04:01 g	0.650
	9	A*02:01 g~B*40:01 g~C*03:04 g~DRB1*13:02 g~DQB1*06:04 g~DPB1*03:01 g	0.515	A*02:01 g~B*40:01 g~C*03:04 g~DRB1*13:02 g~DQB1*06:04 g~DPB1*03:01 g	0.593
	10	A*24:02 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g~DPB1*04:01 g	0.506	A*24:02 g~B*07:02 g~C*07:02 g~DRB1*15:01 g~DQB1*06:02 g~DPB1*04:01 g	0.590

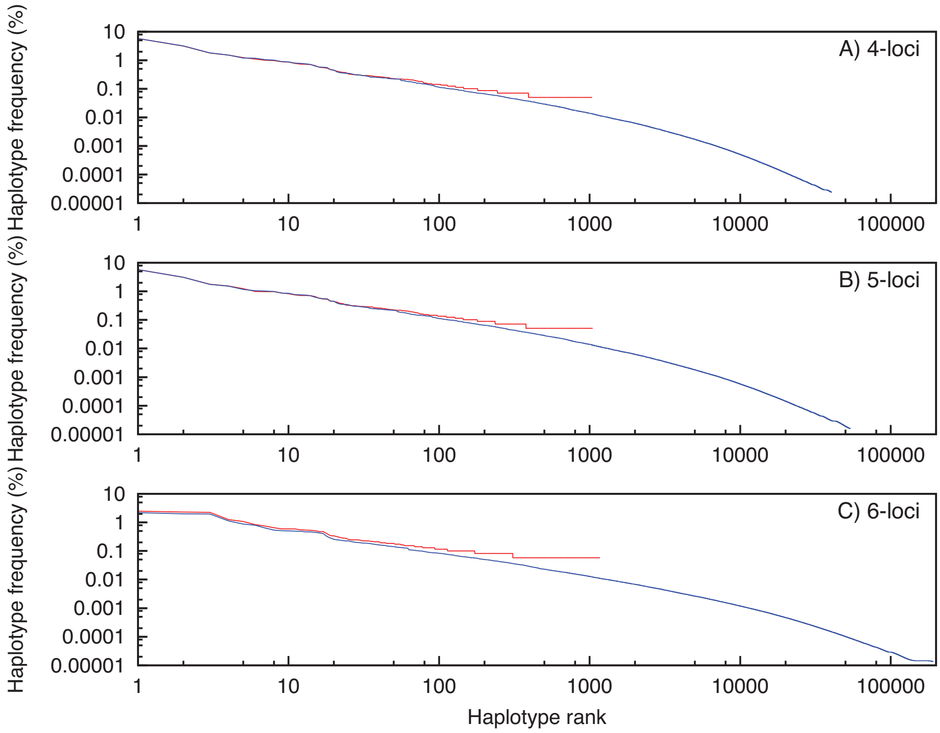

FIGURE 2 Haplotype frequencies. Blue: Hapl-o-Mat; $N = 3{,}456{,}066$. Red: homozygous donors. (a) 4 loci (HLAA, -B, -C, -DRB1); size of homozygous subset: $n_4 = 26{,}311$. (b) 5 loci (HLA-A, -B, -C, -DRB1, -DQB1); $n_5 = 25{,}433$. (c) 6 loci (HLA-A, -B,-C, -DRB1, -DQB1, -DPB1); $n_6 = 9{,}217$

ranging from 1,039 (4 loci), 1,048 (5 loci) to 1,169 (6 loci). The number of haplotypes occurring more than three times was between 113 (six loci) and 180 (4 loci). Therefore, already in the range between haplotype ranks 100 and 1000, a clear step structure of the HF distribution curve became apparent. The individual steps corresponded to the n-fold occurrence of a haplotype in the relatively small homozygous samples, with the lowest step representing haplotypes that occurred once, the next higher step haplotypes that occurred twice and so forth. Due to these pronounced step artifacts, the accuracy of HF estimated from homozygous donors is lower than that of HF obtained with the EM algorithm. However, HF estimation from homozygous donors has the advantage that nonzero HF as calculated for, for example, 1039 4-locus haplotypes prove the existence of the corresponding haplotypes.

A complete list of the estimated HF for both methods and derived allele frequency (AF) is available in Table S1. We will also make these data available on allelefrequencies.net.

4 | DISCUSSION

In this work, we investigated the possibility of determining HLA HF of a large sample of individuals by examining only the subsets fully homozygous for the considered HLA loci. To our knowledge, this approach is new for HLA, although it has already been applied in the analysis of single nucleotide polymorphism data (Yamaguchi-Kabata et al., 2010).

The main appeal of our approach is that the HF can be obtained by simple counting. Thus, it is not possible to obtain frequencies different from zero for haplotypes that are not included in the sample, as it may happen when using the EM algorithm. However, our results are less accurate than those from HF estimation using the EM algorithm. The large reduction in sample size of between 99.2% (4 loci) and 99.7% (6 loci) compared to the full sample of more than 3.4 million individuals means that artifacts associated with sample size are very apparent even at relatively low haplotype ranks between 100 and 1000. Similar step-like artifacts are also known from HF distributions determined with the EM algorithm (Pappas et al., 2015; Schmidt et al., 2020) and also appear in our Hapl-o-Mat results in this work, but much less pronounced and at orders of magnitude lower frequencies or higher haplotype ranks. In practice, the differences in accuracy between the two methods will in most cases be even greater than in this work as one is likely to have much smaller homozygous subsets. This leads to more inaccurate results even for the common haplotypes, as it was the case in our study for the 6-locus haplotypes. As reliable software packages for HF determination via the EM algorithm are available, it is therefore reasonable to rely primarily on these. If questions of the actual existence of certain haplotypes are relevant, our approach based on homozygous individuals may be a valuable addition. It may also be helpful in the validation process of new software implementations of the EM algorithm.

In summary, we presented a new HLA HF estimation method based on homozygous donors. While the results match acceptably with the results of the EM algorithm, especially for frequent haplotypes, they show considerably stronger artifacts. However, our approach allows the identification of definitely existing haplotypes. It should not go unmentioned that the HF and AF of the German population, which were obtained with the EM algorithm from a very large sample of over 3.4 million individuals, represent a relevant data set, even if they were not the focus of this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Susanne Seitz https://orcid.org/0000-0003-4420-0728

Jürgen Sauter https://orcid.org/0000-0001-8485-2945

Alexander H. Schmidt https://orcid.org/0000-0003-0979-5914

REFERENCES

Alfraih, F., Alawwami, M., Aljurf, M., Alhumaidan, H., Alsaedi, H., El Fakih, R., Alotaibi, B., Rasheed, W., Bernas, S. N., Massalski, C., Heidl, A., Sauter, J., Lange, V., & Schmidt, A. H. (2021). High-resolution HLA allele and haplotype frequencies of the Saudi Arabian population based on 45,457 individuals and corresponding stem cell donor matching probabilities. *Human Immunology*, *82*(2), 97–102. https://doi.org/10.1016/j.humimm.2020.12.006

Becker, T., & Knapp, M. (2002). Efficiency of haplotype frequency estimation when nuclear family information is included. *Human Heredity*, *54*(1), 45–53. https://doi.org/10.1159/000066692

Bergstrom, T. C., Garratt, R. J., & Sheehan-Connor, D. (2009). One chance in a million: Altruism and the bone marrow registry. *The American Economic Review*, *99*(4), 1309–1334. https://doi.org/10.1257/aer.99.4.1309

Bochtler, W., Gragert, L., Patel, Z. I., Robinson, J., Steiner, D., Hofmann, J. A., Pingel, J., Baouz, A., Melis, A., Schneider, J., Eberhard, H.-P., Oudshoorn, M., Marsh, S. G. E., Maiers, M., & Müller, C. R. (2016). A comparative reference study for the validation of HLA-matching algorithms in the search for allogeneic hematopoietic stem cell donors and cord blood units. *HLA*, *87*(6), 439–448. https://doi.org/10.1111/tan.12817

Dehn, J., Setterholm, M., Buck, K., Kempenich, J., Beduhn, B., Gragert, L., Madbouly, A., Fingerson, S., & Maiers, M. (2016). HapLogic: A predictive human leukocyte antigen-matching algorithm to enhance rapid identification of the optimal unrelated hematopoietic stem cell sources for transplantation. *Biology of Blood and Marrow Transplantation*, *22*(11), 2038–2046. https://doi.org/10.1016/j.bbmt.2016.07.022

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, *12*(5), 921–927. https://doi.org/10.1093/oxfordjournals.molbev.a040269

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Guo, S. W., & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, *48*(2), 361–372. https://doi.org/10.2307/2532296

Guo, Z., Hood, L., Malkki, M., & Petersdorf, E. W. (2006). Long-range multilocus haplotype phasing of the MHC. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(18), 6964–6969. https://doi.org/10.1073/pnas.0602286103

Ikeda, N., Kojima, H., Nishikawa, M., Hayashi, K., Futagami, T., Tsujino, T., Kusunoki, Y., Fujii, N., Suegami, S., Miyazaki, Y., Middleton, D., Tanaka, H., & Saji, H. (2015). Determination of HLA-A, -C, -B, -DRB1 allele and haplotype frequency in Japanese population based on family study. *Tissue Antigens*, *85*(4), 252–259. https://doi.org/10.1111/tan.12536

Lange, V., Böhme, I., Hofmann, J., Lang, K., Sauter, J., Schöne, B., Paul, P., Albrecht, V., Andreas, J. M., Baier, D. M., Nething, J., Ehninger, U., Schwarzelt, C., Pingel, J., Ehninger, G., & Schmidt, A. H. (2014). Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics*, *15*, 63. https://doi.org/10.1186/1471-2164-15-63

Pappas, D. J., Tomich, A., Garnier, F., Marry, E., & Gourraud, P. A. (2015). Comparison of high-resolution human leukocyte antigen haplotype frequencies in different ethnic groups: Consequences of sampling fluctuation and haplotype frequency distribution tail truncation. *Human Immunology*, *76*(5), 374–380. https://doi.org/10.1016/j.humimm.2015.01.029

Sanchez-Mazas, A., Djoulah, S., Busson, M., Le Monnier De Gouville, I., Poirier, J.-C., Dehay, C., Charron, D., Excoffier, L., Schneider, S., Langaney, A., Dausset, J., & Hors, J. (2000). A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. *European Journal of Human Genetics*, *8*(1), 33–41. https://doi.org/10.1038/sj.ejhg.5200391

Sauter, J., Hernandez, C., Hardtmann, M., & Schmidt, A. H. (2015). German high-resolution haplotype frequencies based on family pedigrees. *Tissue Antigens*, *85*(5), 301–431. https://doi.org/10.1111/tan.12557

Sauter, J., Schäfer, C., & Schmidt, A. H. (2018). HLA haplotype frequency estimation from real-life data with the Hapl-o-Mat software. *Methods in Molecular Biology*, *1802*, 275–284. https://doi.org/10.1007/978-1-4939-8546-3_19

Schäfer, C., Schmidt, A. H., & Sauter, J. (2017). Hapl-o-Mat: Open-source software for HLA haplotype frequency estimation from ambiguous and heterogeneous data. *BMC Bioinformatics*, *18*(1), 284. https://doi.org/10.1186/s12859-017-1692-y

Schmidt, A. H., Baier, D., Solloch, U. V., Stahr, A., Cereb, N., Wassmuth, R., Ehninger, G., & Rutt, C. (2009). Estimation of high-resolution HLA-A, -B, -C, -DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning. *Human Immunology*, *70*(11), 895–902. https://doi.org/10.1016/j.humimm.2009.08.006

Schmidt, A. H., Sauter, J., Pingel, J., & Ehninger, G. (2014). Toward an optimal global stem cell donor recruitment strategy. *Plos One*, *9*(1), e86605. https://doi.org/10.1371/journal.pone.0086605

Schmidt, A. H., Sauter, J., Baier, D. M., Daiss, J., Keller, A., Klussmeier, A., Mengling, T., Rall, G., Riethmüller, T., Schöfl, G., Solloch, U. V., Torosian, T., Means, D., Kelly, H., Jagannathan, L., Paul, P., Giani, A. S., Hildebrand, S., Schumacher, S., … Schetelig, J. (2020). Immunogenetics in stem cell donor registry work: The DKMS example (Part 1). *International Journal of Immunogenetics*, *47*(1), 13–23. https://doi.org/10.1111/iji.12471

Schöfl, G., Lang, K., Quenzel, P., Böhme, I., Sauter, J., Hofmann, J. A., Pingel, J., Schmidt, A. H., & Lange, V. (2017). 2.7 million samples genotyped for HLA by next generation sequencing: Lessons learned. *BMC Genomics*, *18*(1), 161. https://doi.org/10.1186/s12864-017-3575-z

Sonnenberg, F. A., Eckman, M. H., & Pauker, S. G. (1989). Bone marrow donor registries: The relation between registry size and probability of finding complete and partial matches. *Blood*, *74*(7), 2569–2578. https://doi.org/10.1182/blood.V74.7.2569.2569

Steiner, D. (2012). Computer algorithms in the search for unrelated stem cell donors. *Bone Marrow Research*, *2012*, 175419. https://doi.org/10.1155/2012/175419

Urban, C., Schmidt, A. H., & Hofmann, J. A. (2020). Hap-E Search 2.0: Improving the performance of a probabilistic donor-recipient matching algorithm based on haplotype frequencies. *Frontiers in Medicine*, *7*, 32. https://doi.org/10.3389/fmed.2020.00032

Yamaguchi-Kabata, Y., Tsunoda, T., Takahashi, A., Hosono, N., Kubo, M., Nakamura, Y., & Kamatani, N. (2010). Making a haplotype catalog with estimated frequencies based on SNP homozygotes. *Journal of Human Genetics*, *55*(8), 500–506. https://doi.org/10.1038/jhg.2010.56

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Seitz, S., Lange, V., Norman, P. J., Sauter, J., & Schmidt, A. H. (2021). Estimating HLA haplotype frequencies from homozygous individuals – A Technical Report. *Int J Immunogenet*, *48*, 490–495. https://doi.org/10.1111/iji.12553