Received: 23 August 2021 | Accepted: 29 August 2021

DOI: 10.1111/iji.12556

INTERNATIONAL JOURNAL OF IMMUNOGENETICS WILEY

LETTER TO THE EDITOR

Should results of HLA haplotype frequency estimations be normalized?

Dear Editor,

regarding the comment by Nunes (Nunes, 2021) on our publication 'Estimating HLA haplotype frequencies from homozygous individuals' (Seitz et al., 2021):

The only difference between the approach preferred by Nunes and our analysis is that we normalized the estimated haplotype frequencies (HF), that is, we multiplied each frequency by a constant factor chosen so that the frequency sum equals 1. So, the question is whether it is appropriate to normalize an HF set obtained from a corresponding estimation procedure.

We think there may be no universal answer to this question, but that it depends on what the frequencies are intended to be used for. As we mentioned in the introduction of our original paper, we are particularly interested in questions in the context of stem cell donor registries such as what proportion of patients of given ethnicity will find an HLA-matched donor in a registry of defined size and ethnic composition. This question is usually (Beatty et al., 1995; Müller et al., 2003; Schmidt et al., 2014) answered via a two-step procedure: First, one estimates population-specific HF from appropriate samples of HLA-genotyped individuals. Then, the HF obtained are used as input for the determination of matching probabilities (MP) by registry size. In the simplest scenario (all donors and patients are from the same population), this is done using the formula $p\,(n) = \sum_i g_i[1-(1-g_i)^n]$ (Müller et al., 2003). Here, p is the MP, n is the registry size, and the g_i are the genotype frequencies (GF) of the population under consideration that are derived from the HF determined in step 1 under the assumption of Hardy–Weinberg equilibrium (HWE).

We will now analyze the implications of using non-normalized HF sets for MP estimation with the help of the frequency sets from our original paper: For the sums s_i of the estimated HF without normalization, we obtain $s_4 = 1.066$, $s_5 = 1.051$, and $s_6 = 0.926$ for the 4-, 5-, and 6-locus scenarios, respectively. (These results can be easily calculated from data given in the Supplementary Information of our original paper.) It is straightforward to deduce that $p\,(n \to \infty) = s_i^2$. In our three scenarios, we have: $s_4^2 = 1.136$, $s_5^2 = 1.105$, and $s_6^2 = 0.857$.

This means that even in a setting with identical donor and patient populations and arbitrary registry growth, one can never achieve an MP greater than 0.857 in the 6-locus scenario. On the other hand, in the other two scenarios one achieves MP well above 1. These unreasonable results provide, in our view, a strong argument that normalized HF sets are the appropriate outcome of HF estimation for our purposes. As stated above, one may reach different conclusions in other contexts although it might generally be difficult to interpret a frequency from a non-normalized HF set with a frequency sum that deviates considerably from 1.

It should be noted that the question of HF set normalization arises generally, not only in HF estimation based on homozygous individuals. When analyzing the original data set ($n = 3{,}456{,}066$) with the expectation-maximization (EM) algorithm (Excoffier & Slatkin, 1995) using our Hapl-o-Mat software (Sauter et al., 2018; Schäfer et al., 2017), the sum of all HF $\geq 1/(2n)$ (corresponding to a unique occurrence in the sample) ranged from 0.993 (6-locus scenario) to 0.997 (4-locus scenario). The question of whether to normalize such an HF set is obviously less pressing than for the significant deviations of the HF sums from 1 that we obtained without normalization when estimating HF from homozygous individuals. This is another piece of evidence for the general superiority of the EM algorithm over the HF estimation from homozygous donors, which we had already clearly stated in our original paper.

For much smaller – and probably more common – sample sizes, however, the question if estimated HF sets should be normalized becomes more relevant also for the EM algorithm. To demonstrate this, we determined HF from a random sample ($n = 10{,}000$) of the original sample using the Hapl-o-Mat software. The sum of all frequencies corresponding to at least one occurrence in the sample ranged from 0.772 (6-locus scenario) to 0.905 (4-locus scenario). Thus, if one wants to use such an HF set as input for MP estimation and to avoid unreasonable results like above, one has the choice to (a) include frequencies in the calculation whose underlying haplotypes are presumably not included in the sample at all; (b) normalize the estimated HFs; or (c) perform a combination of these two approaches. Indeed, the latter is what we have done in the past (Schmidt et al., 2020). We included, starting with the largest HF, all estimated frequencies – including those $< 1/(2n)$ – up to a cumulative frequency of 0.995, and then normalized this HF set to 1. However, this is a merely pragmatic approach. To our knowledge, there is no standard way to generate input to the MP calculation from the output of the EM algorithm, let alone a mathematically proven optimal approach. We think it would be a worthwhile, though probably non-trivial, scientific effort to define one.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

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, and Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA*

Correspondence
Alexander Schmidt, DKMS, Kressbach 1, 72072 Tübingen, Germany.
Email: schmidt@dkms.de

Linked articles: Seitz, S., et al. *International Journal of Immunogenetic* 2021; https://doi.org/10.1111/iji.12553 and Nunes, J. M. et al. *International Journal of Immunogenetic* 2021; https://doi.org/10.1111/iji.12555

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

Beatty, P. G., Mori, M., & Milford, E. (1995). Impact of racial genetic polymorphism on the probability of finding an HLA-matched donor. *Transplantation*, *60*(8), 778–783. https://doi.org/10.1097/00007890-199510270-00003

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

Müller, C. R., Ehninger, G., & Goldmann, S. F. (2003). Gene and haplotype frequencies for the loci HLA-A, HLA-B, and HLA-DR based on over 13,000 German blood donors. *Human Immunology*, *64*(1), 137–151. https://doi.org/10.1016/S0198-8859(02)00706-1

Nunes, J. M. (2021). A comment on estimating HLA haplotype frequencies from homozygous individuals. *International Journal of Immunogenetics*, *48*(6), 496–497.

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., 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., Daiss, J., Keller, A., Klussmeier, A., T. Mengling, G. Rall, T. Riethmüller, G. Schöfl, U. V. Solloch, T. Torosian, D. Means, H. Kelly, L. Jagannathan, P. Paul, A. S. Giani, S. Hildebrand, S. Schumacher, … 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

Seitz, S., Lange, V., Norman, P. J., Sauter, J., & Schmidt, A. H. (2021). Estimating HLA haplotype frequencies from homozygous individuals – A Technical Report. *International Journal of Immunogenetics*, *48*(6), 490–495.