Experimental Results (2022), 3, e24, 1–10
doi:10.1017/exp.2022.17

CAMBRIDGE
UNIVERSITY PRESS

LIFE SCIENCE AND BIOMEDICINE
NEGATIVE-RESULT

Human leukocyte antigen distributions do not share a copula across sub-populations

Dan Schellhas[1] and Robert C. Green II[2]

[1]Data Science Program, Bowling Green State University, Bowling Green, OH 43402, USA, and [2]Department of Computer Science, Bowling Green State University, Bowling Green, OH 43402, USA
Corresponding author: Email: greenr@bgsu.edu

(Received 09 May 2022; Revised 08 September 2022; Accepted 09 September 2022)

Abstract

The distribution of human leukocyte antigens in the population assists in matching solid organ donors and recipients when the typing methods used do not provide sufficiently precise information. This is made possible by linkage disequilibrium (LD), where alleles co-occur more often than random chance would suggest. There is a trade-off between the high bias and low variance of a broad sample from the population and the low bias but high variance of a focused sample. Some of this trade-off could be alleviated if sub-populations shared LD despite having different allele frequencies. These experiments show that Bayesian estimation can balance bias and variance by tuning the effective sample size of the reference panel, but the LD as represented by an additive or multiplicative copula is not shared.

Key words: copula; estimation; human leukocyte antigens; imputation; organ transplantation

Introduction

Human leukocyte antigens

When performing a solid organ transplant, it is important to not only procure a high-quality organ (Massie et al., 2014), but also ensure the compatibility of the donor and recipient (Doxiadis, 2012). Recent methods for organ allocation match on the eplets that comprise the human leukocyte antigens (HLA) and these require a high-resolution typing of both individuals (Kosmoliaptsis et al., 2008). This level of typing is typically only available for transplants that absolutely require a perfect match, but is too slow and expensive for transplants from deceased donors. To bridge the gap between the low- and mid-resolution typing that are available and the high-resolution that is required, statistical methods are used to estimate the high from the low using programs like HLA Matchmaker (Duquesnoy and Askar, 2007).

These programs use a set of reference panels to estimate the high-resolution HLA from a patient's low-resolution HLA and self-reported ethnicity or race. Ethnic and racial groupings are inadequate proxies for genetic information on their own (Bamshad et al., 2004), but do provide some insight in this setting. Estimates are typically derived from a single large reference panel for that ethnicity. Through an application of Bayesian estimation (explained in the Bayesian Estimation subsection), it is also reasonable to combine a small sample from a group very similar to the patient with a less precise, but larger, sample. The seven currently recognized HLA loci (genetic locations with a recognizable function) are given the alphanumeric labels A, B, C, DRB1, DRB3/4/5, DQB1, and DPB1. However, there are situations in which

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/exp.2022.17 Published online by Cambridge University Press

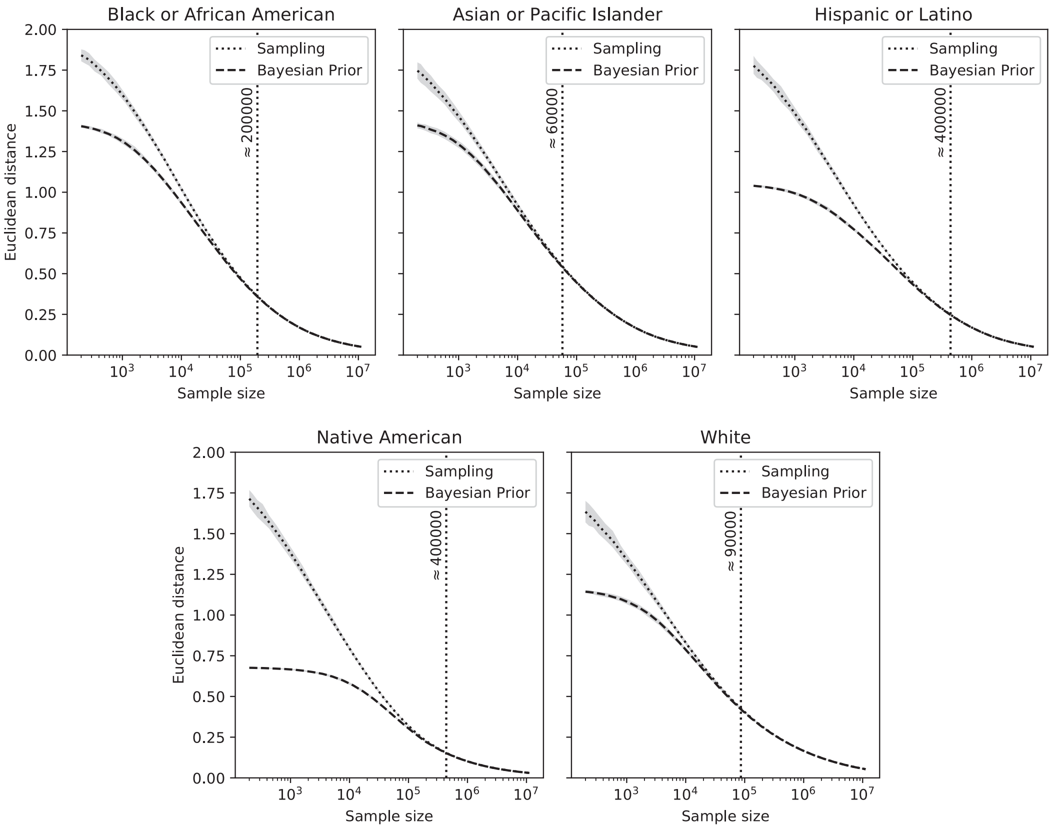

Figure 1. The benefit of Bayesian estimation varies significantly by group, but is clearly useful for reference panels with less than 80,000 samples.

only three or four of these loci are available for clinical or budgetary reasons. In the three-locus setting, the loci used are typically A, B, and DRB1 and the fourth that is typically added is DQB1 (Bekbolsynov, 2018).

The primary feature of the human genome that makes this estimation possible is linkage disequilibrium (LD), where certain alleles (genetic information at a given locus) co-occur with others more often than independent randomness would suggest. As such, it would seem reasonable to suspect that this LD could be shared across ethnicities despite different distributions of fundamental genetic materials. To this end, Bayesian estimation can be applied using the allele frequencies (AF) of the small sample and the copula (relationship among alleles) of the large sample instead of the full reference distribution.

Given a sufficiently large reference panel, one can directly estimate the distribution of high-resolution haplotypes (alleles at a set of loci) given a low-resolution haplotype by conditioning on the low-resolution: $P(\text{high}|\text{low}) = P(\text{high})/P(\text{low})$. However, there is always a trade-off between statistical bias and variance. Having a large reference panel reduces the uncertainty about that panel, and thus the variance of its estimates. Unfortunately, no patient is the average member of that panel, so larger panels provide biased estimates for that individual. To re-balance the trade-off, we can prioritize a sub-population that is more similar to the patient. This sub-population can lower the bias of the estimate, but would increase the variance. Consequently, we would need to find the appropriate balance between the two.

Bayesian estimation

To find a balance between bias and variance we can use Bayesian estimation. Haplotype frequencies follow a multinomial distribution where the probability of each haplotype occurring is its own parameter

https://doi.org/10.1017/exp.2022.17 Published online by Cambridge University Press

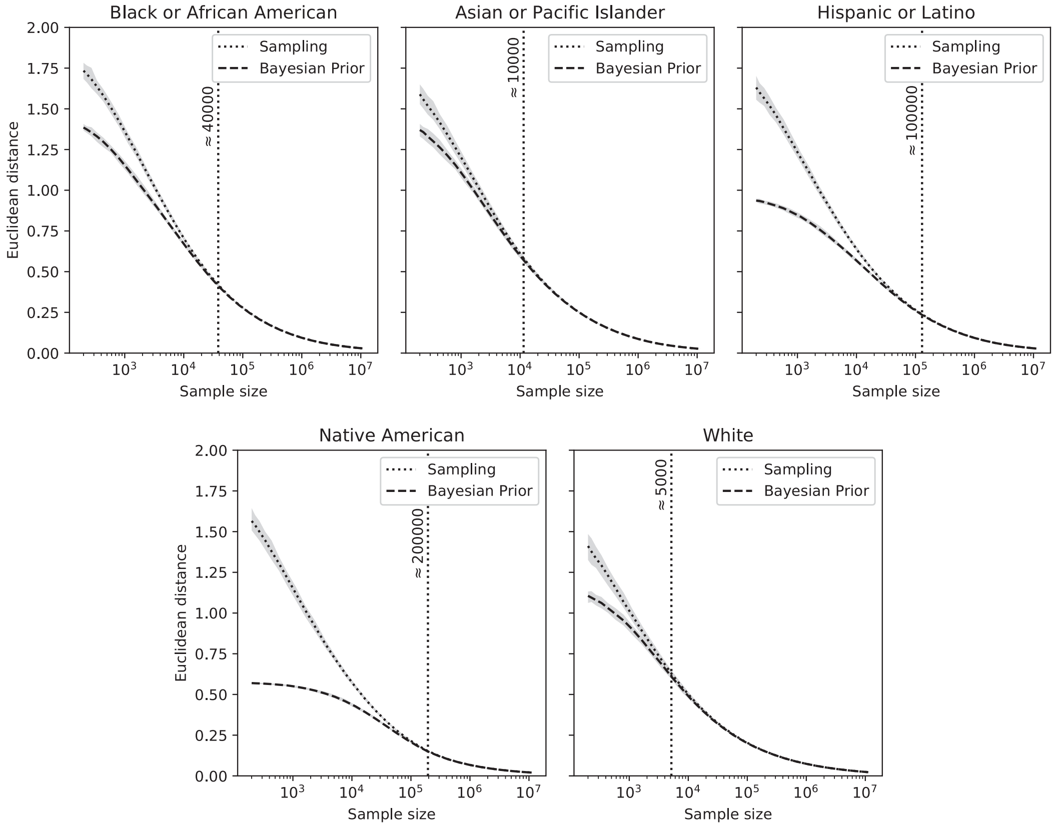

Figure 2. When the loci are reduced to only A, B, DRB1, and DQB1, the sampling error is decreased but the results of Bayesian estimation are nearly identical.

of the model. The likelihood of these parameters is simply the proportion of the small panel that has that haplotype. To accommodate the larger panel, we treat it as a prior belief regarding the nature of the smaller. The combination of the two acts as a posterior belief where the prior belief has been updated by the evidence (likelihood).

$$P(\theta|\, X) \propto \mathscr{L}(X|\, \theta)P(\theta). \tag{1}$$

The probability of the true frequency given the observed data is $P(\theta|X)$ whereas $\mathscr{L}(X|\theta)$ is the likelihood of observing the data given a certain true frequency. That leaves $P(\theta)$, which is the probability of a certain frequency being the true one.

The distribution of the prior belief most appropriate for this application would be the Dirichlet distribution. Similar to the multinomial, each haplotype is its own parameter. However, instead of each being a proportion bounded by $[0, 1]$, they are unbounded above: $[0, \infty)$. This unboundedness provides an easily interpretable result in an effective sample size (the sum of all parameters) for the prior belief that can be tuned to balance the effects of the large panel versus the small. Conveniently, this also has the desirable property that the *posterior* is then also a Dirichlet distribution, so the *maximum a posteriori* estimate is easily computed (Carlin and Louis, 2008). However, one limitation of this prior is that the unboundedness can cause the simulation to fail to converge in cases where no improvement is possible. In practice, one must choose an effective sample size upper limit that is considered effectively infinite such that a result is always returned.

https://doi.org/10.1017/exp.2022.17 Published online by Cambridge University Press

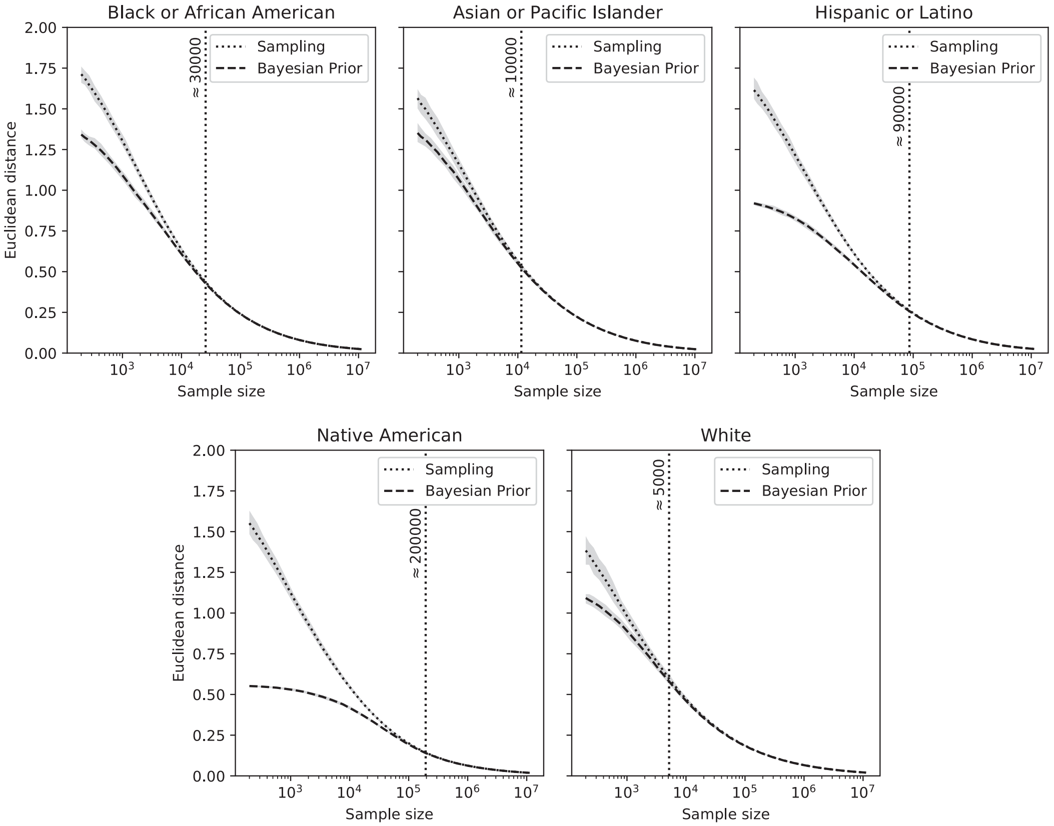

Figure 3. When the loci are reduced to A, B, and DRB1, the results are nearly indistinguishable from the 4-locus version.

To include the shared copula hypothesis, we then treat the prior in the same manner. The AF of the small reference panel becomes the multinomial likelihood for the *prior* and the AF of the large panel becomes a Dirichlet *hyperprior* (a *prior* of a *prior*). This *posterior* for the *prior* is then composed with the copula to regain the *prior* distribution.

$$P(\theta|X) \propto \mathscr{L}(X|\theta)\mathscr{L}(\theta|\phi)P(\phi). \tag{2}$$

Methods

The data for this study comes from the National Marrow Donor Program and contain haplotype frequencies for 21 sub-populations and five standard groupings thereof (Maiers and Gragert, 2020). To simulate having small reference panels with known distribution, haplotypes are sampled with replacement from the five standard groupings: Black or African American; Asian or Pacific Islander; Hispanic or Latino; Native American; and White (BeTheMatch.org, 2022). These groups have sample sizes 1,742,191; 1,965,495; 2,714,930; 228,006; and 11,226,174 respectively. They have 130,972; 149,171; 166,397; 38,244; and 312,928 different 7-locus haplotypes with 642,477 unique haplotypes among them. The number of possible 7-locus haplotypes given known alleles is of order 10^{18}, so LD has reduced this amount dramatically. The large reference panel is then composed of a combination of the remaining four groupings.

This simulation is done at 30 exponentially spaced sample sizes of order 10^2 to 10^7 for adequate accuracy at multiple potential scales across the three haplotype sizes: 7-locus, 4-locus, and 3-locus. Given the sampling described above, the squared error of the frequencies estimated by the Bayesian model is computed by taking advantage of the conjugate nature of the Dirichlet prior and posterior. The posterior estimate involves adding the prior estimated observations of each haplotype (the large reference panel scaled by the effective sample size) to the actual observations and then normalizing this set of parameters to sum to one such that it is comparable to the true frequencies. The bisection method is then used to optimize the effective sample size of the *prior* in the first experiment to minimize the error. The second experiment optimizes the effective sample size of both the *prior* and *hyperprior*.

The objective function of the optimizations is the Euclidean distance between the *maximum a posteriori* estimate of the haplotype distribution and the observed full data for each group summed across all sample sizes. This will find the effective sample size that provides the greatest benefit across all settings as well as the largest small reference panel size that still has Bayesian performance outside the Frequentist 95% confidence interval. Each step of the optimization uses 100 replications of the sampling and when the optimal effective sample size(s) are found another 200 replications are performed for additional precision of the presented results.

The second experiment uses two types of copula: multiplicative and additive. A multiplicative copula assumes that the distribution of haplotypes is proportional to a constant multiple of the haplotype distribution without LD as in (3). The constant is shared, but each group has its own haplotype distribution without LD. An additive copula is similar except its constant is added instead of multiplied as in (4).

$$P(A_1, \ldots, A_7) \propto \alpha P_{\perp}(A_1, \ldots, A_7), \tag{3}$$

$$P(A_1, \ldots, A_7) = \beta + P_{\perp}(A_1, \ldots, A_7). \tag{4}$$

The program is written in Python 3.9.7 and the source code is available to the public (Schellhas, 2022).

Results

The first experiment shows a wide range of optimal effective sample sizes and reference panel sizes that would benefit from Bayesian estimation that have similar relationships across the different haplotype sizes. The details are available in Table 1. Complete details for all sample sizes are available in Figures 1–3. With such a wide variance in such a small sample, no conclusion can be drawn regarding the best effective sample size to use with a new sub-population. It would be reasonable to say that it would be useful for panels of size less than 80,000 with an effective sample size of less than 1,000 with a complete 7-locus haplotype. These numbers drop to about 10,000 and 300 respectively for 4- and 3-locus haplotypes. For less conservative practitioners, the effective sample size could be rather safely increased to 2,000 (or 600) without increasing the bias much. This experiment also found that the Native American panel ($N = 228{,}006$) would benefit from Bayesian estimation with the remaining four sub-populations as a *prior* since such a prior was found to be useful up to a panel size of about 400,000 for 7-locus haplotypes.

The second experiment failed to converge to an optimal effective sample size for the *hyperprior* across all haplotype sizes. Since the parameters of the Dirichlet distribution are unbounded, the search algorithm could continue increasing them indefinitely. This implies that the large panel AF should have an infinite effective sample size which means that using the small panel AF with the copula estimated from the large panel is not useful. That makes the result of experiment two identical to that of experiment one. Thus, there is no evidence for the LD to be the same across sub-populations relative to their AF.

Discussion

The simulations demonstrated the benefit of tempering the variance of haplotype frequency estimates using Bayesian estimation. Including the larger reference panel despite it being from other ethnicities

https://doi.org/10.1017/exp.2022.17 Published online by Cambridge University Press

Table 1. Optimal effective sample sizes vary by ethnicity from 1,000 to 12,000 and the largest panel size that benefits from additional references vary from 86,000 to 650,000. Native American is the only broad reference panel that falls within the size where the Bayesian estimation would help

	Effective sample size			Useful until		
Sub-population	7-locus	4-locus	3-locus	7-locus	4-locus	3-locus
Black or African American	2,289	694	553	≈200,000	≈40,000	≈30,000
Asian or Pacific Islander	1,260	292	275	≈60,000	≈10,000	≈10,000
Hispanic or Latino	4,617	2,151	2,162	≈400,000	≈100,000	≈90,000
Native American	12,847	8,240	7,058	≈400,000	≈200,000	≈200,000
White	2,636	649	593	≈90,000	≈5,000	≈5,000

improves performance. It seemed reasonable to expect the LD present in HLA to have similar causes across ethnicities and thus have shared influence when attempting HLA imputations. The lack of any evidence of this using these two copula implies that either (a) the sources of LD are different between groups or (b) none of the shared nature is captured by either of these copula. Both options are reasonable directions for future research.

The lack of convergence of the simulations may cast some doubt on the results, but the algorithm was allowed to run until the effective sample size exploded to a point where the shared copula results were statistically indistinguishable from the copula-free first experiment.

Conclusion

LD in HLA does not appear to have a simple relationship with the available genetic materials in a given population. This implies that the biological pressures have historically differed across sub-populations or there is a feature of HLA that is not made immediately apparent by how they are currently encoded.

Funding Information. This research received no grants from any funding agency, commercial, or not-for-profit sectors.

Authorship Contribution. Investigation: D.S.; Methodology: D.S.; Writing—original draft: D.S.; Data curation: R.C.G.; Resources: R.C.G.; Supervision: R.C.G.; Writing—review and editing: R.C.G.

Data Availability Statement. The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.4474442. The source code is available on GitLab at https://gitlab.com/dschell/hla-shared-copula/.

Conflict of Interest. The authors have no conflicts of interest to declare.

References

Bamshad, M., Wooding, S., Salisbury, B. A., & Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. *Nature Reviews Genetics*, **5**(8), 598–609.

Bekbolsynov, D. (2018). *A new concept of immunogenicity to calculate the risk stratification for kidney transplantation* [PhD thesis, University of Toledo].

BeTheMatch.org (2022). *About us.* https://bethematch.org/transplant-basics/matching-patients-with-donors/how-does-a-patients-ethnic-background-affect-matching/.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.

Doxiadis, I. I. (2012). Compatibility and kidney transplantation: the way to go. *Frontiers in Immunology*, **3**, 111.

Duquesnoy, R. J., & Askar, M. (2007). HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. V. Eplet matching for HLA-DR, HLA-DQ, and HLA-DP. *Human Immunology*, **68**(1), 12–25.

Kosmoliaptsis, V., Bradley, J. A., Sharples, L. D., Chaudhry, A., Key, T., Goodman, R. S., & Taylor, C. J. (2008). Predicting the immunogenicity of human leukocyte antigen class I alloantigens using structural epitope analysis determined by HLAMatchmaker. *Transplantation*, **85**(12), 1817–1825.

https://doi.org/10.1017/exp.2022.17 Published online by Cambridge University Press

Maiers, M., & Gragert, L. (2020). NMDP 7-locus haplotype frequencies trimmed. https://doi.org/10.5281/zenodo.4474442.

Massie, A., Luo, X., Chow, E., Alejo, J., Desai, N., & Segev, D. (2014). Survival benefit of primary deceased donor transplantation with high-KDPI kidneys. *American Journal of Transplantation*, **14**(10), 2310–2316.

Schellhas, D. (2022). *Source code.* https://gitlab.com/dschell/hla-shared-copula/.

Cite this article: Schellhas D, Green RC (2022). Human leukocyte antigen distributions do not share a copula across sub-populations. *Experimental Results,* 3, e24, 1–10. https://doi.org/10.1017/exp.2022.17

Peer Reviews

Reviewing editor: Dr. Vitor Francisco

Minor revisions requested.

doi:10.1017/exp.2022.17.pr1

Review 1: Human leukocyte antigen distributions do not share a copula across sub-populations

Reviewer: Dr. Michael Nevels

University of St Andrews, Biomolecular Sciences Building, Fife, United Kingdom of Great Britain and Northern Ireland, KY16 9ST

Date of review: 15 August 2022

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Conflict of interest statement. Reviewer declares none

Comment

Comments to the Author: This computational study is overall very well presented, and the conclusions appear to be justified by the data. However, I recommend that the manuscript be sent to a statistical editor for further validation since I am not a statistics expert.

Minor issues to be fixed:

1. The authors should increase the size of all figure panels. The font size in the figures is too small and the labels are difficult to read.

2. The authors provide little if any discussion of their results and do not highlight the limitations. Thus, a brief Discussion should be added, either as a separate section or as part of the Results.

3. The first sentence in the legend to Table 1 ("...and largest panel size that benefits from additional references vary...") appears to be grammatically incorrect.

Score Card

Presentation

3.9 /5

Is the article written in clear and proper English? (30%)	5/5
Is the data presented in the most useful manner? (40%)	3/5
Does the paper cite relevant and related articles appropriately? (30%)	4/5

Context

5.0 /5

Does the title suitably represent the article? (25%)	5/5
Does the abstract correctly embody the content of the article? (25%)	5/5

Does the introduction give appropriate context? (25%)	5/5
Is the objective of the experiment clearly defined? (25%)	5/5

Analysis

3.4
/5

Does the discussion adequately interpret the results presented? (40%)	3/5
Is the conclusion consistent with the results and discussion? (40%)	4/5
Are the limitations of the experiment as well as the contributions of the experiment clearly outlined? (20%)	3/5

doi:10.1017/exp.2022.17.pr2

Review 2: Human leukocyte antigen distributions do not share a copula across sub-populations

Reviewer: Dr. Hao Chen

Date of review: 23 August 2022

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Conflict of interest statement. Reviewer declares none.

Comment

Comments to the Author: The article discussed the distribution of HLA in fitve subpopulations and their application on tuning bias and variance in the high-resolution HLA estimation. Their simulations showed a larger reference panel in the Bayesian estimation would be useful to balance the bias of the small panel, but no copula of LD was shared in the subpopulations. Overall, their conclusion could benefit development of the HLA estimation algorithm in future.

However, this manuscript was lack of many detais which the authors should offer:

(1) How to define the larger and smaller panel?

(2) the details of the first and second experiment. e.g. how to perform the simulations? how to get the effective sample size from the simulations?

(3) Formula in the manuscript was not well defined. For instance, in Formula (1), what does θ,X,P,L means? They need to be explicitly defined in manuscript.

Score Card

Presentation

4.3 /5

Is the article written in clear and proper English? (30%)	4/5
Is the data presented in the most useful manner? (40%)	4/5
Does the paper cite relevant and related articles appropriately? (30%)	5/5

Context

5.0 /5

Does the title suitably represent the article? (25%)	5/5
Does the abstract correctly embody the content of the article? (25%)	5/5
Does the introduction give appropriate context? (25%)	5/5
Is the objective of the experiment clearly defined? (25%)	5/5

Analysis

4.8 /5

Does the discussion adequately interpret the results presented? (40%)	5/5
Is the conclusion consistent with the results and discussion? (40%)	5/5
Are the limitations of the experiment as well as the contributions of the experiment clearly outlined? (20%)	4/5