The Journal of Open Source Software

HLAfreq: Download and combine HLA allele frequency data

David A. Wells [1] and **Michael McAuley** [2]

1 Barinthus Biotherapeutics, United Kingdom **2** School of Mathematics and Statistics, Technological University Dublin, Dublin, Ireland

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @jdrugo
- @usethedata
- @assignUser

Submitted: 27 August 2025
Published: unpublished

Summary

Human leukocyte antigen (HLA) genes encode cell-surface proteins which play an important role in immunity. Since different HLA alleles enable different immune responses, the population frequency of HLA alleles is often considered when designing vaccines (Gulukota & DeLisi, 1996). Specific HLA alleles have been linked to autoimmune disease (Simmonds & Gough, 2007) and associated with adverse drug reactions (Fan et al., 2017). Further, the success of solid organ and stem cell transplants is related to HLA matching between donor and recipient (Fürst et al., 2019; Morishima et al., 2002).

We present `HLAfreq`: a Python package which can be used to download, combine and analyse multiple HLA allele frequency datasets.

Statement of need

The Allele Frequency Net Database is a publicly available repository for human immune gene frequency data from across the world (Gonzalez-Galarza et al., 2020). However, downloading data from a large number of studies is currently manual and slow. After downloading multiple studies, combining them is hindered by different allele resolutions, missing alleles, and incomplete studies. `HLAfreq` provides functions to identify incomplete studies, handle missing alleles, harmonise allele resolution, calculate population coverage, and estimate allele frequencies and uncertainty using a Bayesian framework. Allele frequency plots can be generated to identify anomalous datasets and interesting diversity in a set of populations. To get started, see the guide and examples at github.com/BarinthusBio/HLAfreq.

Methods

Statistical methods

`HLAfreq` uses a Bayesian framework to estimate allele frequency statistics from combined datasets for a specific population. The user can select from two statistical models. The simpler 'default model' gives point estimates for allele frequencies. The more sophisticated 'compound model' gives both point estimates and credible intervals.

Default model

Let p_k be the frequency of the k-th allele of a particular gene in a given population (e.g. a country). The default model assumes that the observations from all datasets for the population are drawn independently and that the probability of being the k-th allele is p_k. In other words, each observation is drawn from a categorical distribution with parameters $(p_1, \dots, p_K)$

where K is the total number of alleles. The prior for $(p_1, \dots, p_K)$ is taken to be a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_K$. The Dirichlet distribution is a generalisation of the Beta distribution to higher dimensions; see Section 4.6.3 of (Murphy, 2022).

The Dirichlet distribution is conjugate to the categorical distribution, meaning that the posterior distribution for the default model is also Dirichlet. More precisely, if the combined datasets contain x_k observations of the k-th allele (for $k = 1, \dots, K$) then the posterior distribution is Dirichlet with parameters $\alpha_1 + x_1, \dots, \alpha_K + x_K$. The posterior mean for the frequency of allele j is then given by

$$\frac{\alpha_j + x_j}{\sum_{k=1}^{K} (\alpha_k + x_k)}.$$

By default, `HLAfreq` takes the prior parameters to be $\alpha_1 = \dots = \alpha_K = 1$. This results in a uniform prior on $(p_1, \dots, p_K)$ subject to the constraints that $p_1, \dots, p_K \geq 0$ and $p_1 + \dots + p_K = 1$. The user can specify alternative values for $\alpha_1, \dots, \alpha_K$. These parameters may be interpreted as a 'pseudocount' in the sense that choosing the prior $\alpha_1, \dots, \alpha_K$ is equivalent to taking a uniform prior and then observing a dataset with $\alpha_k - 1$ observations of the k-th allele. (Intuitively the uniform prior corresponds to one observation of each allele). This can be used as a heuristic for choosing prior parameters based on external information.

`HLAfreq` does not provide credible intervals based on the default model because they are frequently unrealistically narrow. This is because the default model does not account for variance between studies. The compound model, described below, is more complex but accounts for this variation and provides accurate credible intervals.

Compound model

The default model assumes that all observations are sampled from a homogeneous population; however, observations within a single study are more likely to be similar e.g. they may be sampled at the same time or place. To account for this, `HLAfreq` provides a 'compound model' which accounts for the grouping of observations within studies and allows the allele frequencies of study populations to differ from each other. The additional uncertainty results in wider but more accurate credible intervals. This falls within the general class of hierarchical Bayesian models: see Chapter 5 (Gelman et al., 2014) for further details and background.

The compound model makes the following assumptions. As before, p_k denotes the frequency of the k-th allele in the population and the prior distribution for $p_1, \dots, p_K$ is Dirichlet with parameters $\alpha_1, \dots, \alpha_K$. A concentration parameter $\gamma \geq 0$ is given with a standard log-normal prior distribution. For the j-th data source, a vector $\beta^{(j)} = (\beta_1^{(j)}, \dots, \beta_K^{(j)})$ is sampled independently from a Dirichlet distribution with parameters $\gamma p_1, \dots, \gamma p_K$. Observations from the j-th data source are then sampled from a categorical distribution with parameters $\beta_1^{(j)}, \dots, \beta_K^{(j)}$. (Equivalently, the j-th data source as a whole is sampled from a multinomial distribution.)

Idiosyncratic sampling biases are captured by the different values of $\beta^{(j)}$, which result in different probabilities of sampling particular alleles for each data source. If γ is large, then $\beta^{(j)}$ is likely to concentrate around $(p_1, \dots, p_K)$ which means that different studies tend to have similar allele frequencies.

The posterior distributions of $p_1, \dots, p_K$ and γ do not have a closed form and so are estimated numerically using `PyMC` (Salvatier et al., 2016). The `HLAfreq` function `AFhdi` outputs posterior means and credible intervals for allele frequencies.

Research Impact Statement

`HLAfreq` has been used in the design of several vaccines and immunotherapies by Barinthus Biotherapeutics.

Software Design

`HLAfreq` was written in `python` rather than `R` to take advantage of `requests` and `bs4` for AFND's recommended "automated access". After downloading, the data are return in pandas dataframes rather than a custom class for familiarity and in line with Scientific-Python recommendations.

AI usage disclosure

No generative AI tools were used in the development of this software, the writing of this manuscript, or the preparation of supporting materials.

Acknowledgements

MM was supported by the European Research Council (ERC) Advanced Grant QFPROBA (grant number 741487). DW is employed by Barinthus Biotherapeutics (UK) Ltd.

References

Fan, W.-L., Shiao, M.-S., Hui, R. C.-Y., Su, S.-C., Wang, C.-W., Chang, Y.-C., & Chung, W.-H. (2017). HLA association with drug-induced adverse reactions. *Journal of Immunology Research*, *2017*.

Fürst, D., Neuchel, C., Tsamadou, C., Schrezenmeier, H., & Mytilineos, J. (2019). HLA matching in unrelated stem cell transplantation up to date. *Transfusion Medicine and Hemotherapy*, *46*(5), 326–336. https://doi.org/10.1159/000502263

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third, p. xiv+661). CRC Press, Boca Raton, FL. ISBN: 978-1-4398-4095-5

Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. D., Jones, J., Takeshita, L., Ortega-Rivera, N. D., Cid-Pavon, G. M. D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., Middleton, D., & Jones, A. R. (2020). Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, *48*(D1), D783–D788. https://doi.org/10.1093/nar/gkz1029

Gulukota, K., & DeLisi, C. (1996). HLA allele selection for designing peptide vaccines. *Genetic Analysis: Biomolecular Engineering*, *13*(3), 81–86. https://doi.org/10.1016/1050-3862(95)00156-5

Morishima, Y., Sasazuki, T., Inoko, H., Juji, T., Akaza, T., Yamamoto, K., Ishikawa, Y., Kato, S., Sao, H., Sakamaki, H., & others. (2002). The clinical significance of human leukocyte antigen (HLA) allele compatibility in patients receiving a marrow transplant from serologically HLA-a, HLA-b, and HLA-DR matched unrelated donors. *Blood, The Journal of the American Society of Hematology*, *99*(11), 4200–4206. https://doi.org/10.1182/blood.V99.11.4200

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT press.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Simmonds, M., & Gough, S. (2007). The HLA region and autoimmune disease: Associations and mechanisms of action. *Current Genomics*, *8*(7), 453–465. https://doi.org/10.2174/138920207783591690