



hlabud: HLA genotype analysis in R

Kamil Slowikowski^{1,2,3,4} , and Alexandra-Chloe Villani^{1,2,3,4} 

¹Center for Immunology and Inflammatory Diseases, Division of Rheumatology, Allergy and Immunology, Department of Medicine, Massachusetts General Hospital, ²Cancer Center, Massachusetts General Hospital, ³Broad Institute, ⁴Harvard Medical School

Summary

The human leukocyte antigen (HLA) genes have thousands of different alleles in the human population, and have more associations with human diseases than any other genes. Data for all known HLA genotypes are curated in the international ImmunoGeneTics (IMGT) database in versioned releases on [GitHub](#). Here, we introduce *hlabud*, an R package that provides access to data from the IMGT/HLA database and the Allele Frequency Net Database (AFND), functions to encode the data in different formats, and tutorials for association analysis, embedding, and HLA divergence.

Availability

Source code and documentation are available at github.com/slowkow/hlabud

Contact

kslowikowski@mgh.harvard.edu

Keywords immunoinformatics, genetics, immunology, HLA

1. INTRODUCTION

Human leukocyte antigen (HLA) genes encode the proteins that display antigens so the immune system can recognize pathogens such as bacteria and viruses. Geneticists have identified thousands of variants (e.g. single nucleotide polymorphisms) in the human genome that are associated with hundreds of different disease and phenotypes [1].

The HLA genes encode a protein complex that presents antigens to other cells.

To facilitate HLA genotype analysis, we developed *hlabud*, a free and open-source software package that downloads information from the IMGT/HLA database of HLA genotypes and sequence alignments [2] directly in the R programming environment. The *hlabud* package provides functions that return convenient lists of items, where each item is either a matrix or a data frame. The simple design makes *hlabud* easy to integrate with any downstream R packages for data analysis or visualization.

hlabud downloads HLA genotype data from the IMGT-HLA GitHub repository [3] and automatically caches it in a user-configurable folder. Functionality includes

PRE-PRINT

Published Nov 27, 2023

parsing the custom IMGT/HLA file format for multiple sequence alignments, converting sequence alignments to a one-hot matrix, and calculating the Grantham divergence between HLA alleles [4].

The documentation includes tutorials for analysis of the one-hot encoding of amino acid positions, including association analysis with logistic regression and low-dimensional embedding with UMAP [5]. *hlabud* also provides direct access to the allele frequencies for all HLA genes from the Allele Frequency Net Database (AFND) [6].

2. DESCRIPTION

Comprehensive HLA genotype data is curated in the IMGT/HLA database, and the data is archived in a GitHub repository (github.com/ANHIG/IMGT_HLA). We can use *hlabud* to download the sequence alignment data, read it into R, and automatically encode the data as a one-hot matrix like this:

```
a <- hla_alignments("DRB1")
```

When the user runs this line of code, *hlabud* will:

- Download data from the IMGT/HLA Github repository.
- Cache data files in a local folder that supports multiple releases of the data.
- Read the data into data frames and matrices for downstream analysis.
- Create a one-hot encoding of the multiple sequence alignment data.

Many amino acid residues at specific loci have been associated with human diseases and blood protein levels [7]. Researchers have developed software tools for calling HLA genotypes with high accuracy from DNA-seq or RNA-seq next-generation sequencing reads [8], so there are opportunities to use that data for association studies.

Once we have a list of genotypes for each individual (e.g. "DRB1*04:01,DRB1*05:01"), we can use *hlabud* to prepare data for regression analysis to find which amino acid positions are associated with a phenotype in a sample of individuals. We call `dosage(genotypes, a$onehot)` where `genotypes` is a vector of genotypes and `a$onehot` is a one-hot matrix representation of HLA alleles (from the example above). The `dosage()` function returns the number of copies of each amino acid at each position for each individual, which can then be used for omnibus regression [9] or single-position testing (Figure 1A).

UMAP accepts the one-hot matrix of HLA alleles as input, and it can be used to visualize the dataset in a latent space with reduced dimensionality (Figure 1B).

hlabud provides direct access to the allele frequencies HLA genes reported in the Allele Frequency Net Database (AFND) (<http://allelefrequencies.net>) (Figure 1C).

Each HLA allele binds a specific set of peptides. So, an individual with two highly dissimilar alleles can bind a greater number of different peptides than a

homozygous individual [10]. *hlabud* implements the Grantham divergence calculations based on the original Perl code [4]:

```
my_genos <- c("A*23:01:12,A*24:550", "A*25:12N,A*11:27", "A*24:381,A*33:85")
hla_divergence(my_genos, method = "grantham")
#> A*23:01:12,A*24:550      A*25:12N,A*11:27      A*24:381,A*33:85
#>      0.4924242      3.3333333      4.9015152
```

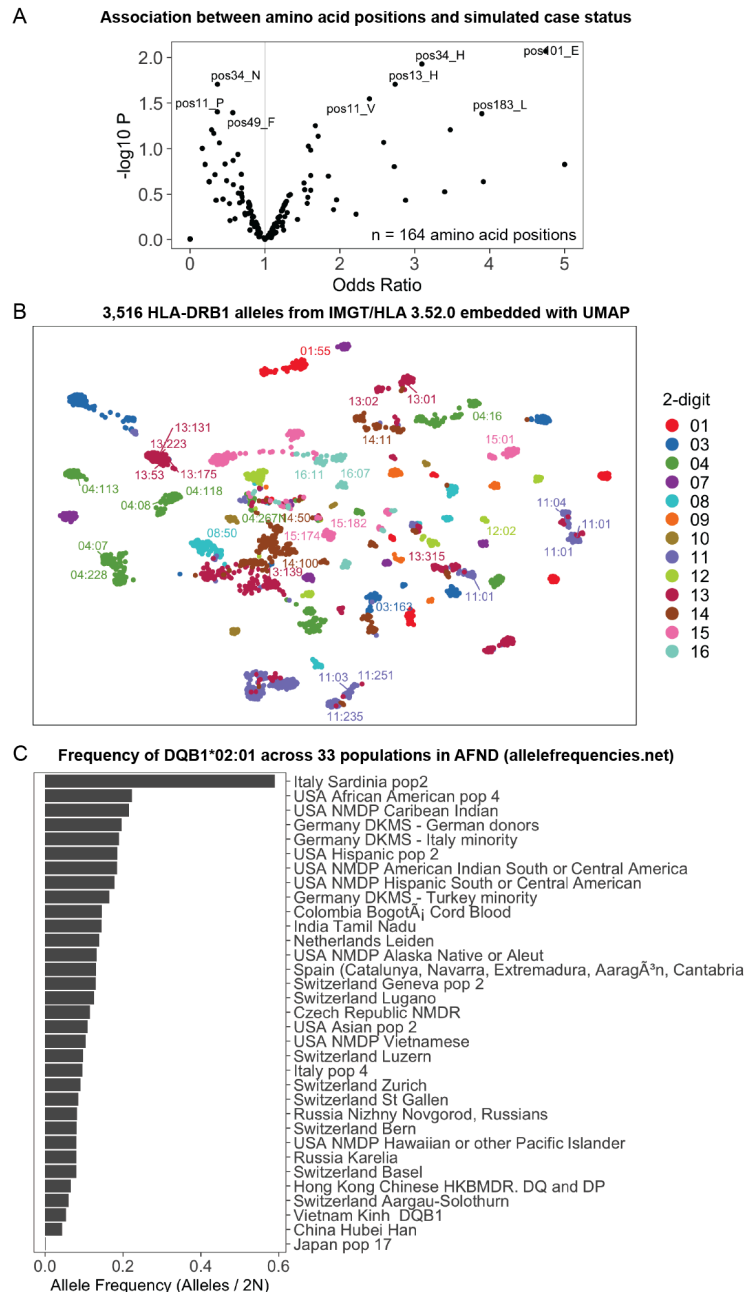


Figure 1: **(A)** Association between amino acid positions and simulated case-control status. The x-axis represents the odds ratio and the y-axis represents $-\log_{10} P$ from a logistic regression analysis in R. **(B)** 3,516 HLA-DRB1 alleles represented as dots in a two-dimensional embedding computed by UMAP from a one-hot encoding of amino acids. **(C)** Allele frequencies for HLA-DQB1*02:01 in the AFND.

3. INSTALLATION AND DOCUMENTATION

hlabud can be installed in an R session with:

```
remotes::install_github("slowkow/hlabud")
```

Each function is documented extensively, and the complete manual can be viewed on the *hlabud* website at <https://slowkow.github.io/hlabud>. *hlabud* has been tested on Linux/Unix, Mac OS (Darwin) and Windows.

4. DISCUSSION

Our open-source R package *hlabud* enables easy access to HLA data from two public databases, and provides functions to enable HLA divergence calculations, regression analysis, and low-dimensional embedding. We hope that *hlabud* will raise awareness of the IMGT/HLA and AFND databases and influence other developers to share more open-source tools for HLA analysis. We envision that *hlabud* will be used by biomedical researchers, and also by teachers and students who study genetics and bioinformatics.

5. ACKNOWLEDGMENTS

This work was supported by a NIAID grant T32AR007258 (to K.S.) and the National Institute of Health Director's New Innovator Award (DP2CA247831; to A.C.V.) Thanks to Sreekar Mantena for reporting issues with the code.

6. COMPETING INTERESTS

No competing interest is declared.

7. AUTHOR CONTRIBUTIONS STATEMENT

K.S. wrote the software and the manuscript. A.C.V. reviewed the manuscript.

8. RELATED WORK

BIGDAWG is an R package that provides functions for chi-squared Hardy-Weinberg and case-control association tests of highly polymorphic genetic data like HLA genotypes [11]. HATK is set of Python scripts for processing and analyzing IMGT-HLA data [12].

BIBLIOGRAPHY

- [1] A. E. Kennedy, U. Ozbek, and M. T. Dorak, "What has GWAS done for HLA and disease associations?", *International journal of immunogenetics*, vol. 44, no. 5, pp. 195–211, Oct. 2017, doi: [10.1111/iji.12332](https://doi.org/10.1111/iji.12332).
- [2] J. Robinson, D. J. Barker, X. Georgiou, M. A. Cooper, P. Flicek, and S. G. E. Marsh, "IPD-IMGT/HLA Database", *Nucleic acids research*, vol. 48, no. D1, p. D948–D955, Jan. 2020, doi: [10.1093/nar/gkz950](https://doi.org/10.1093/nar/gkz950).

- [3] J. Robinson, D. Barker, X. Georgiou, and M. Cooper, "A GitHub repository with files currently published in the IPD-IMGT/HLA FTP Directory hosted at the European Bioinformatics Institute". [Online]. Available: <https://github.com/ANHIG/IMGTHLA>
- [4] F. Pierini and T. L. Lenz, "Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection", *Molecular Biology and Evolution*, vol. 35, no. 9, p. 2145, Jun. 2018, doi: [10.1093/molbev/msy116](https://doi.org/10.1093/molbev/msy116).
- [5] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". [Online]. Available: <https://arxiv.org/abs/1802.03426>
- [6] F. F. Gonzalez-Galarza *et al.*, "Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools", *Nucleic acids research*, vol. 48, no. D1, p. D783--D788, Jan. 2020, doi: [10.1093/nar/gkz1029](https://doi.org/10.1093/nar/gkz1029).
- [7] C. Krishna *et al.*, "The influence of HLA genetic variation on plasma protein expression", Jul. 2023, doi: [10.1101/2023.07.24.550394](https://doi.org/10.1101/2023.07.24.550394).
- [8] A. Claeys, P. Merseburger, J. Staut, K. Marchal, and J. Van den Eynden, "Benchmark of tools for in silico prediction of MHC class I and class II genotypes from NGS data", *BMC Genomics*, vol. 24, no. 1, May 2023, doi: [10.1186/s12864-023-09351-z](https://doi.org/10.1186/s12864-023-09351-z).
- [9] S. Sakaue *et al.*, "Tutorial: a statistical genetics guide to identifying HLA alleles driving complex disease", *Nature Protocols*, vol. 18, no. 9, p. 2625, Jul. 2023, doi: [10.1038/s41596-023-00853-4](https://doi.org/10.1038/s41596-023-00853-4).
- [10] E. K. Wakeland *et al.*, "Ancestral polymorphisms of MHC class II genes: Divergent allele advantage", *Immunologic Research*, vol. 9, no. 2, p. 115, Jun. 1990, doi: [10.1007/bf02918202](https://doi.org/10.1007/bf02918202).
- [11] D. J. Pappas, W. Marin, J. A. Hollenbach, and S. J. Mack, "Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline", *Human immunology*, vol. 77, no. 3, pp. 283–287, Mar. 2016, doi: [10.1016/j.humimm.2015.12.006](https://doi.org/10.1016/j.humimm.2015.12.006).
- [12] W. Choi, Y. Luo, S. Raychaudhuri, and B. Han, "HATK: HLA analysis toolkit", *Bioinformatics*, vol. 37, no. 3, p. 416, Jul. 2020, doi: [10.1093/bioinformatics/btaa684](https://doi.org/10.1093/bioinformatics/btaa684).