



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 more associations with human diseases than any other genes, and there are thousands of different HLA alleles in the human population. Data for all known HLA genotypes are curated in the international ImmunoGeneTics (IMGT) database, and allele frequencies for each HLA allele across human populations are available in the Allele Frequency Net Database (AFND). Our open-source R package *hlabud* accesses HLA data from IMGT and AFND, and supports further analysis such as HLA divergence calculation, fine-mapping analysis of amino acid (or nucleotide) positions, and low-dimensional embedding.

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 enable cells to display antigens to other cells, which is one mechanism for immune recognition of 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 diseases and phenotypes [1]. HLA genes have a greater number of disease associations than any other genes.

HLA nomenclature consists of allele names like *HLA*01:01* and *HLA*02:01* to indicate the genotype of an individual in a study [2]. Each allele name corresponds to a haplotype that contains multiple mutations at different positions throughout the entire length of the gene sequence. It is difficult to estimate the similarity of two alleles solely from the allele names: any two alleles might differ by one or more nucleotide or amino acid residues. Any encoding of genotype data that is ambiguous regarding nucleotide or amino acid positions is not ideal for statistical analysis, because some positions might contain more information than others.

PRE-PRINT

Published Nov 28, 2023

Researchers have developed many software tools for calling HLA genotypes (Figure 1) with high accuracy from DNA-seq or RNA-seq next-generation sequencing reads [3], so there are opportunities to use this type of data for HLA association studies. Providers of HLA typing services often report genotypes with the traditional HLA allele names (i.e. *HLA*01:01*) instead of reporting alleles at specific nucleotide positions (Figure 1), and most software tools produce outputs that follow this convention of reporting allele names.

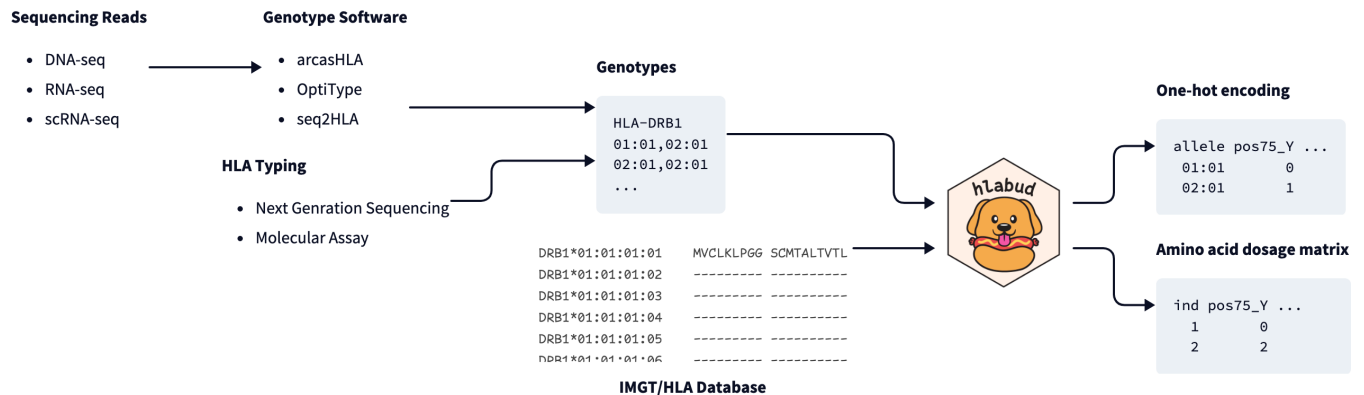


Figure 1: *hlabud* converts HLA genotypes to amino acid position matrices.

In contrast to allele-level analysis, fine-mapping analysis associates a phenotype with each amino acid (or nucleotide) at each position. Many amino acid residues at specific loci have been associated with human diseases and blood protein levels [4]. Published amino acid associations represent opportunities for experimental validation that could advance understanding of the disease-associated mechanisms related to HLA proteins.

Results from fine-mapping analysis can be interpreted in the context of the protein structures that are affected by the associated amino acid positions. We might have different hypotheses about the function of a mutation in the peptide binding groove than a mutation in the interior region of the protein.

To facilitate HLA fine-mapping, we developed *hlabud*, a free and open-source R package that downloads data from the IMGT/HLA database [5] and automatically creates amino acid (or nucleotide) position matrices that are ready for analysis (Figure 1). *hlabud* functions return simple lists, where each item in the list is a matrix or a data frame. This design makes it easy to integrate *hlabud* with any downstream R packages for data analysis or visualization.

2. EXAMPLES

2.1. Downloading data for a gene

Curated HLA genotype data is provided by the IMGT/HLA database at GitHub (github.com/ANHIG/IMGTHLA). In the example below, we use *hlabud* to download the sequence alignment data for *HLA-DRB1*, read it into R, and encode it as a one-hot matrix:

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

With one line of code, *hlabud* will:

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

2.2. Computing a dosage matrix

Once we have obtained a list of genotypes for each individual (e.g. "DRB1*04:01,DRB1*05:01"), we can use *hlabud* to prepare data for fine-mapping regression analysis that will reveal which amino acid positions are associated with a phenotype in a sample of individuals. To calculate the number of copies of each amino acid at each position for each individual, we can run:

```
dosage(genotypes, a$onehot)
```

where *genotypes* is a vector of *HLA-DRB1* genotypes and *a\$onehot* is a one-hot matrix representation of *HLA-DRB1* alleles. The dosage matrix can then be used for omnibus regression [6] or fine-mapping (i.e. regression with each single position) (Figure 2A).

2.3. Visualizing alleles in two dimensions

Visualizing data in a two-dimensional embedding with algorithms like UMAP [7] can help to build intuition about the relationship between all objects in a dataset. UMAP accepts the one-hot matrix of HLA alleles as input, and the resulting embedding can be used to visualize the dataset for exploratory data analysis (Figure 2B).

2.4. Allele frequencies in human populations

hlabud provides direct access to the allele frequencies of HLA genes in the Allele Frequency Net Database (AFND) [8] (<http://allelefrequencies.net>) (Figure 2C).

2.5. HLA divergence

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 [9]. *hlabud* implements the Grantham divergence calculations [10] (based on the original Perl code) to estimate which individuals can bind a greater number of peptides (higher Grantham divergence):

```
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
```

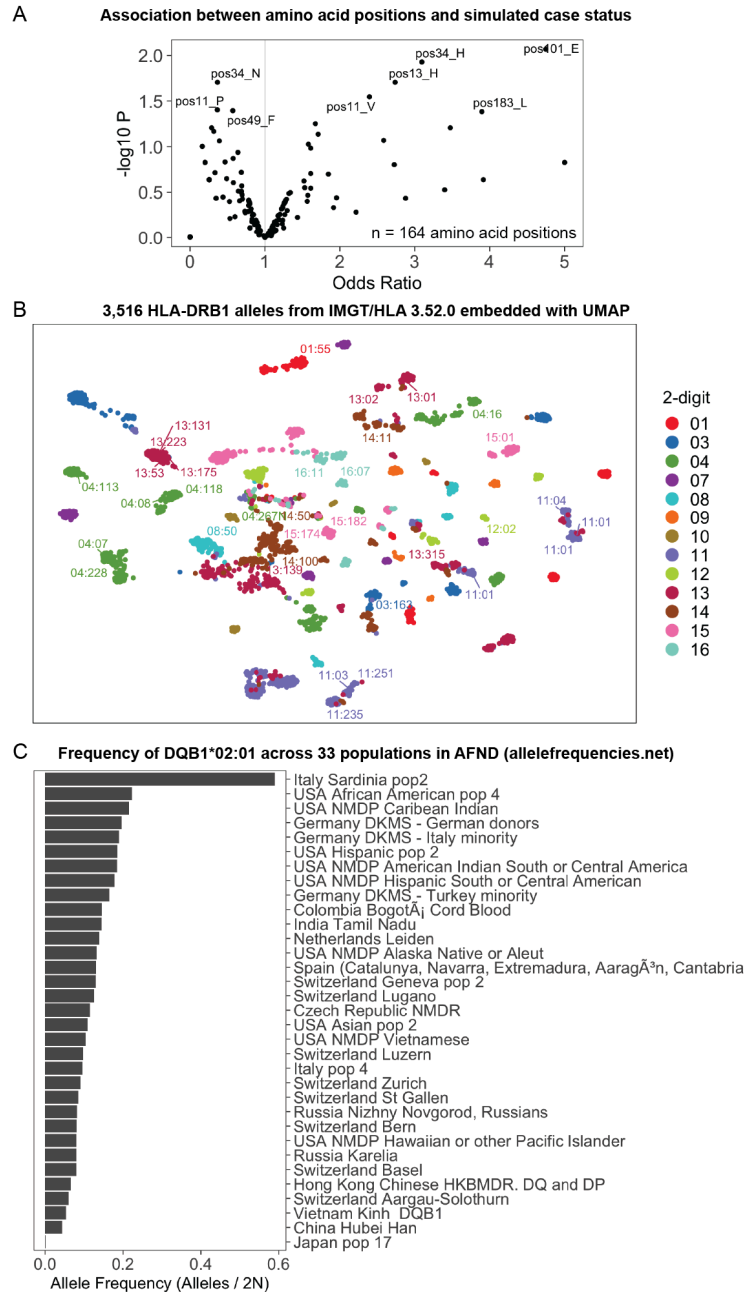


Figure 2: **(A)** Association between amino acid positions and simulated case-control status. The x-axis represents the odds ratio and the y-axis represents $-\log_{10} P$ from a logistic regression analysis in R. **(B)** 3,516 HLA-DRB1 alleles represented as dots in a two-dimensional embedding computed 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

The easiest way to install *hlabud* is to run this command in an R session:

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

The complete manual is available 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* gives users access to HLA data from two public databases, and implements HLA divergence calculation [10]. *hlabud* downloads and caches HLA genotype data from the IMGT-HLA GitHub repository [11] and prepares the data for downstream analysis in R.

We provide [tutorials](#) for HLA divergence, fine-mapping association analysis with logistic regression, embedding with UMAP, and visualizing allele frequencies from the Allele Frequency Net Database (AFND) [8].

5. 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 [12]. HATK is set of Python scripts for processing and analyzing IMGT-HLA data [13].

6. 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. Thanks to Jean Fan for creating the logo and discussing the paper.

7. COMPETING INTERESTS

No competing interest is declared.

8. AUTHOR CONTRIBUTIONS STATEMENT

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

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] S. G. E. Marsh *et al.*, "Nomenclature for factors of the HLA system, 2010", *Tissue Antigens*, vol. 75, no. 4, p. 291, Mar. 2010, doi: [10.1111/j.1399-0039.2010.01466.x](https://doi.org/10.1111/j.1399-0039.2010.01466.x).
- [3] 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).

- [4] 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).
- [5] 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).
- [6] 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).
- [7] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". [Online]. Available: <https://arxiv.org/abs/1802.03426>
- [8] 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).
- [9] 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).
- [10] 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).
- [11] 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>
- [12] 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).
- [13] 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).