hlabud: HLA genotype analysis in R

Kamil Slowikowski [1,2,3,4], and **Alexandra-Chloe Villani** [1,2,3,4]

[1] Center for Immunology and Inflammatory Diseases, Division of Rheumatology, Allergy an Immunology, Department of Medicine, Massachusetts General Hospital, [2] Cancer Center, Massachusetts General Hospital, [3] Broad Institute, [4] Harvard 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, and the Allele Frequency Net Database (AFND) provides allele frequencies for each HLA allele across human populations. Our open-source R package *hlabud* facilitates access to HLA data from IMGT/HLA and AFND, and provides functions for HLA divergence calculations, 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, 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 diseases and phenotypes [1].

HLA nomenclature consists of allele names like *HLA*01:01* to indicate the genotype of each individual in a study. Each allele name corresponds to multiple mutations at different positions throughout the gene's sequence, so it is difficult to estimate the similarity of two alleles solely from the allele names. This ambiguity about specific amino acid positions means that allele names are not ideal for statistical analysis.

Researchers have developed software tools for calling HLA genotypes (Figure 1) with high accuracy from DNA-seq or RNA-seq next-generation sequencing reads [2], so there may be opportunities to use this type of data for HLA association studies. Most software tools report allele names, not genotypes at specific

Pre-Print

Published Nov 27, 2023

nucleotide positions. Commercial providers of HLA typing services also report genotypes with the traditional HLA allele names (i.e. *HLA*01:01*) instead of reporting alleles at specific nucleotide positions (Figure 1).

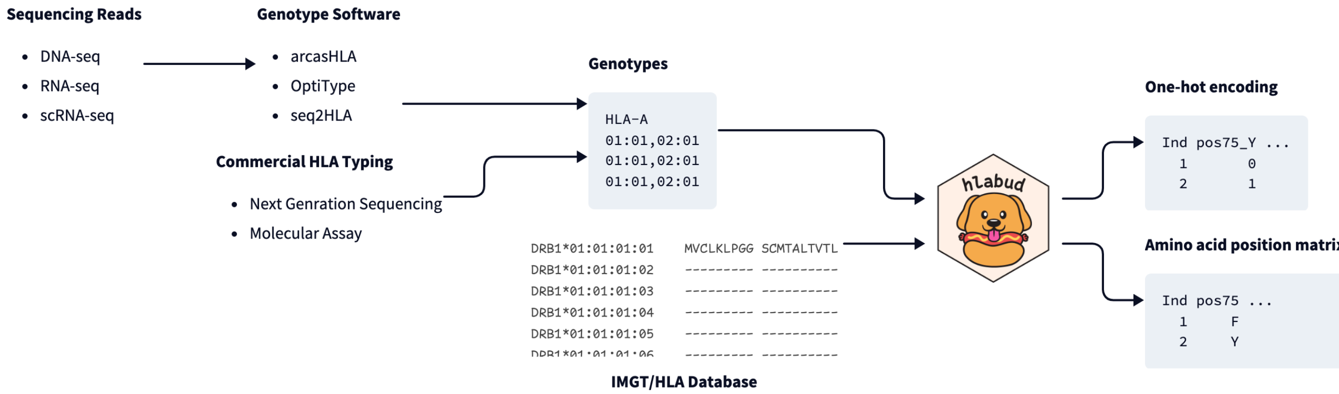

Figure 1: *hlabud* converts HLA genotypes to amino acid position matrices.

In contrast, fine-mapping analysis involves associating a phenotype with each amino acid position. Many amino acid residues at specific loci have been associated with human diseases and blood protein levels [3]. Published associations at specific amino acid positions have created opportunities for experimental validation that might advance our understanding of disease-associated mechanisms related to HLA proteins.

Fine-mapping can be more sensitive than allele-level analysis, and the results can be interpreted in the context of the protein structures that are affected by the associated amino acid positions. For example, we might have different ideas 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 analysis, we developed *hlabud*, a free and open-source R package that downloads data from the IMGT/HLA database [4] 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 simple design makes *hlabud* easy to integrate with any downstream R packages for data analysis or visualization.

2. Examples

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.

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 [5] or fine-mapping (i.e. regression with each single position) (Figure 2A).

Visualizing data in a two-dimensional embedding with algorithms like UMAP [6] 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).

hlabud provides direct access to the allele frequencies of HLA genes in the Allele Frequency Net Database (AFND) [7] (http://allelefrequences.net) (Figure 2C).

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 [8]. *hlabud* implements the Grantham divergence calculations [9] (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
```

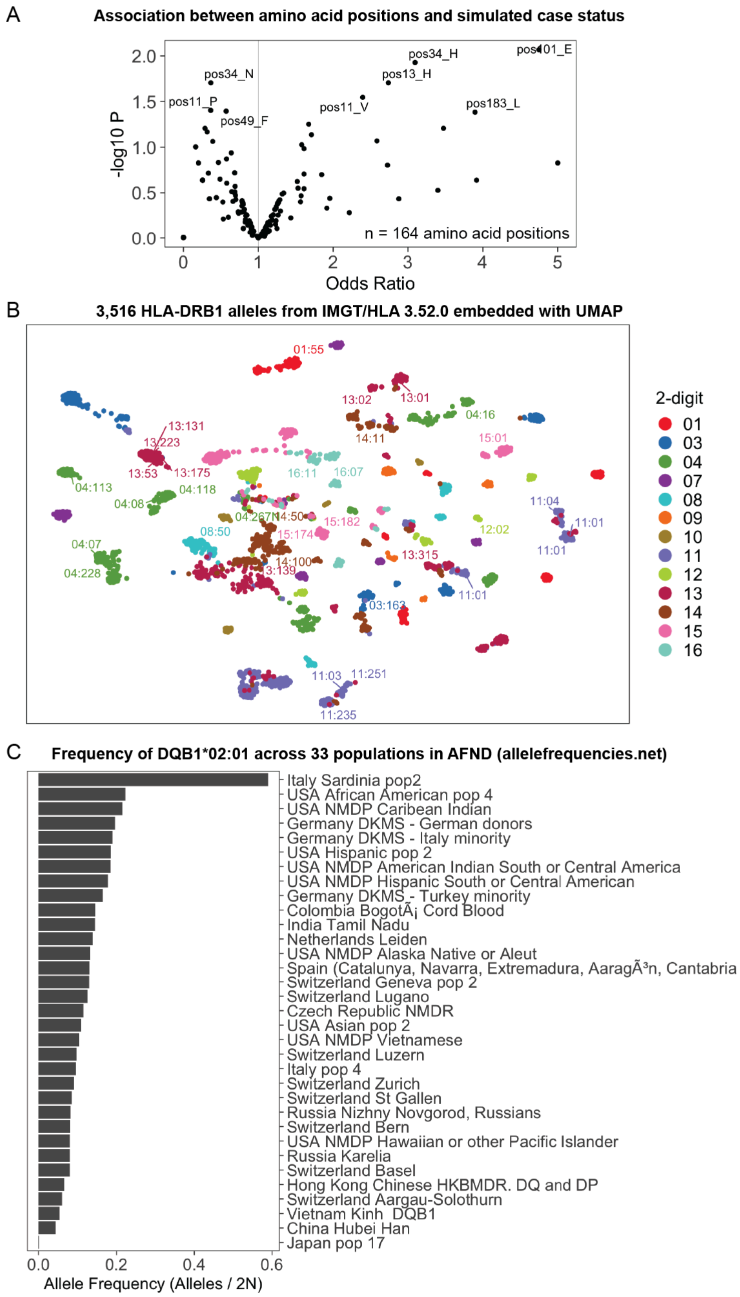

Figure 2: (**A**) Association between amino acid positions and simulated case-control status. The x-axis represents the odds ratio and the y-axis represents $-\log_{10} P$ from a logistic regression analysis in R. (**B**) 3,516 HLA-DRB1 alleles represented as dots in a two-dimensional embedding computed 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* enables easy access to HLA data from two public databases and provides functions for HLA divergence calculations, amino acid or nucleotide fine-mapping analysis, and low-dimensional embedding. *hlabud* downloads HLA genotype data from the IMGT-HLA GitHub repository [10], caches it in a user-configurable folder, and prepares the data for downstream analysis in R.

We provide tutorials for HLA divergence calculation, fine-mapping association analysis with logistic regression, and embedding with UMAP. *hlabud* also provides direct access to the allele frequencies for all HLA genes from the Allele Frequency Net Database (AFND) [7].

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.

[2] 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.

[3] C. Krishna *et al.*, "The influence of HLA genetic variation on plasma protein expression", Jul. 2023, doi: 10.1101/2023.07.24.550394.

[4] 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.

[5] 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.

[6] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". [Online]. Available: https://arxiv.org/abs/1802.03426

[7] 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.

[8] 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.

[9] 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.

[10] 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

[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.

[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.