hlabud: HLA genotype analysis in R

Kamil Slowikowski [1] ✉ **and Alexandra-Chloe Villani** [2]

[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

Abstract

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

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 Kennedy2017.

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 Marsh2010. 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.

Researchers have developed many software tools for calling HLA genotypes (diagram) with high accuracy from DNA-seq or RNA-seq next-generation sequencing reads Claeys2023, 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 (diagram), and most software tools produce outputs that follow this convention of reporting allele names.

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 Krishna2023. 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 Robinson2020 and automatically creates amino acid (or nucleotide) position matrices that are ready for analysis (diagram). *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.

Examples

Downloading data for a gene

Curated HLA genotype data is provided by the IMGT/HLA database at GitHub. 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.

Computing a dosage matrix

Once we have obtained a list of genotypes for each individual (e.g. '"DRB1*04:01,DRB1*05 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 **Sakaue2023** or fine-mapping (i.e. regression with each single position) (figexamples).

Visualizing alleles in two dimensions

Visualizing data in a two-dimensional embedding with algorithms like UMAP McInnes2018 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 (figexamples).

Allele frequencies in human populations

hlabud provides direct access to the allele frequencies of HLA genes in the Allele Frequency Net Database (AFND) Gonzalez-Galarza2020 (link("http://allelefrequences.net" (figexamples).

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 Wakeland1990. *hlabud* implements the Grantham divergence calculations Pierini2018 (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
```

Discussion

Our open-source R package *hlabud* gives users access to HLA data from two public databases, and implements HLA divergence calculation Pierini2018. *hlabud* downloads and caches HLA genotype data from the IMGT-HLA GitHub repository imgthla 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) Gonzalez-Galarza2020.

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 Pappas2016. HATK is set of Python scripts for processing and analyzing IMGT-HLA data Choi2020.

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.

Competing Interests

No competing interest is declared.

Author contributions statement

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