Allele Frequencies Database

Louise Y.C. Takeshita^a Andrew R. Jones^a Faviel F. Gonzalez-Galarza^a Derek Middleton^b

^aInstitute of Integrative Biology, University of Liverpool, Liverpool, UK,

^bRoyal Liverpool University Hospital, Liverpool, UK

Keywords

Data management system · HLA · Immunogenetics · KIR

Summary

This review describes a database for the collection, archiving, sorting, searching and display of gene and allele frequencies for immunogenetic genes.

Introduction

The website Allele Frequencies Net Database (AFND) (www.allelefrequencies.net) was conceived and implemented in 2003, initially to collect and show frequencies of HLA alleles/allelic lineages, (thereafter in this publication called alleles), expressed as decimals, and the frequencies of individuals who carried the alleles, expressed as a percentage, in worldwide populations. Since then, data from other polymorphic immunogenetic loci, namely killer-cell immunoglobulin-like receptors (KIR) as well as MHC class I chain-related (MIC) and several cytokine polymorphisms, have been added [1]. Thus the database is highly relevant for those determining the alleles of all these immunogenetic genes, in solid organ and stem cell transplantation and in anthropology.

Demographics

To date AFND holds 1,324 populations differentiated by polymorphic region (table 1). In total data are available for 4,539,670 individuals. AFND provides an online repository with a set of querying tools to provide searching mechanisms. Data is collected in two aspects; demographic data to give de-

tails of the population and frequency data. The demographic data includes the number of individuals tested, the testing method used, the source of individuals e.g. anthropology study, the ethnicity, whether grandparents/parents came from same region and publication details, although data does not have to be published. Populations are given a name that best describes them; usually this consists of the name of country, region of country and if specified, ethnicity. If a new population cannot be differentiated by name from an existing population a number will be added, e.g. 'Japan 2'. If needed, 'Notes' are added for clarification. For example, if ambiguous data is added where a contributor has not been able to differentiate some alleles, a note is made to the fact that frequency has been added under the first allele of the ambiguity.

Submission of Data

The original intent of obtaining data was that individuals would directly, via an online submission system, add the demographic data via drop down boxes to the website. Thereafter, a file would be sent to them showing the HLA alleles to which the frequency data would be added (alleles from new releases of IMGT/HLA (www.ebi.ac.uk/ipd/imgt/hla/) are automatically added to AFND, as are alleles that have their name changed). Unfortunately this mechanism of data entry has had limited success. Thus, more than 80% of the data has been added by us manually from peer-reviewed publications. As such AFND must be missing a lot of data which, although of good quality, is not published.

At present, we are working with the editor of *Human Immunology* to set up a procedure whereby frequency data can be published in a set format along with demographic data in a brief communication section of the journal. We believe this will aid our attempts to obtain more data directly from community submissions. In the future, with online data sets in-

Table 1. Populations on AFND

Polymorphic region	Population studies	Gene/allele data	Haplotype data	Genotype data
HLA	925	910	365	–
KIR	225	224	–	143
Cytokine	114	114	–	–
MIC	60	60	21	–
Total	1,324	1,308	386	143

Fig. 1. Example of an allele frequency in world-wide populations.

creasingly replacing publications as a unit of citation (citation to a data set on a website acting as a peer-reviewed reference), an increase in the number of populations should ensue. In the intended format in *Human Immunology*, individuals will have to submit HLA types of the individuals making up their study. This will enable the data to be subjected to quality control aspects such as Hardy-Weinberg equilibrium testing and enabling meta-analysis to be run. It will also mean that existing data sets can be compared with other data sets at different time periods with regards to degree of resolution, new alleles being found, etc. It is envisaged that data sets submitted via this mechanism will have to fulfil minimum reporting requirements describing the method used for typing, the alleles tested for, how the allele interpretation was arrived at, etc.

AFND was set up so that alleles that could be determined but were not found were given a frequency of zero, to distinguish them from alleles that could not be determined in which case the frequency was left blank. If properly implemented, this would have meant that the website would show for a population which alleles had not been tested for, including those that would only be found and sequenced in later years. AFND was able to do this for data submitted directly but unfortunately most peer-reviewed publications do not describe which alleles could be determined and do not even give the release number from the IMGT/HLA database of the version of alleles available at the time of testing.

Searching the Data

Multiple filter schemas performed in each of the frequency searches allow users to optimise their searches and obtain matching results that best fit their requirements. Normally, the data can be searched according to the following filters: population, country, geographical region (e.g. West Europe) as well as level of resolution and source of data. Data can also be filtered on the sample size of population, by level of resolution and year of test. Apart from looking at all data for a specific locus, the user can search for specific alleles at that locus. Validation of data ensures that correct and current nomenclature is used. High resolution data is shown when low resolution alleles have been selected. For example, a search for HLA-A*01: 01 will also show allele frequencies, if available, at higher resolution that start with HLA-A*01: 01. The data can be displayed one population at a time, or one allele at a time in all populations and each of these categories can be sorted by highest to lowest frequency. Data can also be displayed on worldwide maps (fig. 1). The results of searches can be accessed programmatically via a URL. For example, the results of a search for global population frequencies for HLA-B*57: 01 can be accessed via the following address: www.allelefrequencies.net/hla6006a.asp?hla_locus_type=Classical&hla_locus=B&hla_allele1=B*57%3A01&hla_allele2=B*57%3A01. This mechanism thus enables other databases to

link to specific search results provided by AFND. AFND has also several links to other websites; for example, if an individual has performed a search and is interested in specific alleles, he/she can be directed specifically to details of that allele, including frequency, on IMGT/HLA.

HLA haplotype information is also provided. To date frequencies from 44,210 HLA haplotypes and 272 MICA-HLA-B associations from 3,838,492 individuals are available. When users wish to search for haplotype frequencies they may enter an allele at one or more loci, leaving other loci as 'any' or 'not' according to their needs. They will be provided with haplotype frequencies according to their search.

One of suggestions made to AFND has been to have more explicit reasoning behind how the data is displayed as some data could be open to misinterpretation. For example, in populations with 4-digit typing, AFND would, in addition to showing the 4-digit data, show 2-digit allele lineage data. Thus, it would appear that frequency of alleles added to more than one. In addition some data would not add up to 1, as not all allele frequencies were necessarily published. Thus, we are in process of adding an 'Explanation of Data' section whereby these points are elucidated and the populations they refer to are listed. Another example is that sometimes the numbers typed in the same population study for different loci are different. Previously, this was explained in the demographics but we also have thought it was worthwhile to show this in the 'Explanation of Data' section.

Archiving HLA Types

Initially HLA individual types making up a population study were not collected. However, such data is now being collected as part of an FP7 European grant, EUROSTAM. This grant seeks to assist individuals who, because of very high levels of sensitisation to HLA antigens, find it virtually impossible to receive a kidney transplant. The HLA laboratory at Eurotransplant led by F. Claas have shown that individuals with these high levels of antibodies (reacting with greater than 90% of the population) can receive a successful transplant by taking into account acceptable mismatches, which of course includes their own type. Despite this innovative idea, many patients still languish on their own country's transplant waiting list. The purpose of EUROSTAM is to ascertain in which other European country these individuals might have success in obtaining a transplant. Thus, AFND is now revisiting those populations that are already on the website along with new populations to obtain raw data, i.e. HLA types of individuals in these populations. Thus, it can be determined if some individuals would benefit from being on the transplant list of another European country.

Rare Alleles

The increase in number of HLA alleles, especially since molecular techniques came to the fore, is vast and continues to rise, even more so after Next Generation Sequencing is widely applied. This increase makes it very difficult for laboratories and vendors to keep their techniques current to determine these new alleles. At the last release from IMGT/HLA (release 3.15.0, January 2014) there are 10,533 known HLA alleles. A section in AFND has been introduced in an attempt to quantify the incidence of these alleles. Each allele is shown with information on whether the initial sequence submitted to IMGT/HLA has been confirmed and, if so, in how many individuals and in addition whether this allele has been found in individuals typed for the National Marrow Donor Programme (NMDP) or in individual laboratories. Most of the data from individual laboratories comes from projects conducted under auspices of the 15th and 16th International Histocompatibility and Immunogenetic Workshops [2, 3]. An individual searcher on AFND can determine their own criteria for what they want to use to define rarity of an allele. Previous reports during the International Workshops have shown that around 40% of HLA alleles have never been reported in another individual, after the report of the initially sequenced sample. Thus, laboratories can use this information in estimating what allele is present when they are faced with an ambiguous combination in the HLA type. At present AFND provides information on the country and ethnicity in which the allele is found, but the search is not able to show rare alleles filtered by each country or even in each geographic region, or filtered by ethnicity. This is one area where improvements are needed. But this depends totally on obtaining data from laboratories.

KIR Genes

Although the HLA section is probably the most used aspect of AFND, as mentioned previously data is also collected from other polymorphic immunogenetic gene systems. Of these KIR is the next most popular section of the database. Some additional data is supplied for KIR compared to HLA. Notably, frequencies of genotypes of KIR individuals (presence/absence of *KIR* gene) in different populations are available for 225 populations. To date 492 different genotypes are present on AFND, varying in their frequency; some being present in all populations studied, whereas others have only been found in one population, or indeed, in isolated cases, in one individual. When the genotype is found in only one population, it is conceivable that the data determined may not be accurate. Thus AFND provides a mechanism whereby the closest genotype to the rare genotype is displayed. This can indicate to the provider of the data, which gene could have been tested false-positive or false-negative in the unique gen-

otype. Each of the genotypes has been given a number, and it is gratifying for continuity purposes that in many publications authors list their genotypes according to that number.

The section for KIR has been augmented with a recently developed sub-section that shows KIR associations with disease [4]. It was thought that this was logistically practical as KIR determination in disease is a relatively new application, compared to HLA associations in disease, which have been reported for 50 years and would be too difficult to capture. One of the main purposes of this KIR/disease section is to collect non-published studies, which due to publication bias, are in the main studies that show no association.

Structural and Sequence-Based Analysis Tools

Another addition to AFND is the Amino Acid Section. Frequency data inputted at 4-digit level is converted to the frequency of individual amino acids at polymorphic positions in HLA alleles. In a disease association study, allele(s) can be shown to be more or less frequent in the disease cohort than the control population. However, in a way this is somewhat artificial, as it is not the allele as whole that is associated, but rather individual functional aspects of the allele. Thus, to examine the frequency of particular amino acids in disease and control cohorts is more meaningful. This section is a forerunner to the new section on HLA epitopes, which was stimulated by the collection of data from the EUROSTAM project. In transplantation, HLA epitope data is starting to change the current view of HLA matching, from allele matching to structural matching where epitopes are patches of polymorphic residues that can stimulate production of specific anti-HLA antibodies, a concept especially important in preventing high sensitization of transplant patients [5]. The new HLA epitopes part of the site uses the nomenclature released through the HLA Epitope Registry [6], which indicates the mapping from HLA allele-level nomenclature to confirmed or predicted epitopes. In AFND the epitope sec-

tion shows the frequencies of these epitopes in different populations, with various ways for querying and visualisation. Initial data in worldwide populations is being delivered from 4-digit HLA data, either from raw data or by statistical means from haplotype data.

Conclusion

AFND now contains the most extensive archive of immune gene/allele frequencies. The success can be judged by the 300 hits received on average every day. AFND has been used in many applications. Although not comparable to actually searching on stem cell registries, it can be used to give some indication of which country a potential donor might be found, particularly if one uses the haplotype information available. It has also been used extensively in clinical laboratories to aid in determining HLA types and in anthropology research units for comparative purposes. However, for AFND to be a continued success we rely on help from colleagues. AFND is very willing to listen to feedback, suggestions and criticisms of the website, in order for improvement to ensue. One area in which the readers of this review might be interested is in whether there should be a section on blood group frequencies in worldwide populations. To do this, we would need to liaise and obtain assistance from someone familiar to this field. This might be suitable for e.g. a Master's project for anyone interested.

In the future it is our intention to have available statistical packages on AFND so that these can be used by individuals to examine their data, to compare them with other populations or to perform analysis on existing populations of their choice.

Disclosure Statement

None of the authors have a conflict of interest.

References

- 1 Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR: Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res* 2011; 39:D913–919.
- 2 Gonzalez-Galarza FF, Mack SJ, Hollenbach J, Fernandez-Vina M, Setterholm M, Kempenich J, Marsh SG, Jones AR, Middleton D: HLA Rare Allele Consortium: 16th IHIW: extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles. *Int J Immunogenet* 2013; 40: 60–65.
- 3 Middleton D, Gonzalez F, Fernandez-Vina M, Tiercy JM, Marsh SG, Aubrey M, Bicalho MG, Cannossi A, Carter V, Cate S, Guerini FR, Loiseau P, Martinetti M, Moraes ME, Morales V, Perasaari J, Setterholm M, Sprague M, Tavoularis S, Torres M, Vidal S, Witt C, Wohlwend G, Yang KL: A bioinformatics approach to ascertaining the rarity of HLA alleles. *Tissue Antigens* 2009; 74: 480–485.
- 4 Takeshita LY, Gonzalez-Galarza FF, dos Santos EJ, Maia MH, Rahman MM, Zain SM, Middleton D, Jones AR: A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations. *Database (Oxford)* 2013; 2013:bat021.
- 5 Duquesnoy RJ: Antibody-reactive epitope determination with HLAMatchmaker and its clinical applications. *Tissue Antigens* 2011; 77: 525–534.
- 6 Duquesnoy RJ, Marrari M, da M Sousa LC, de M Barroso JR, de S U Aita KM, da Silva AS, do Monte SJ: 16th IHIW: a website for antibody-defined HLA epitope Registry. *Int J Immunogenet* 2013; 40: 54–59.