IDAWG: The Immunogenomics Data-Analysis Working Group

Steven J. Mack , Henry A. Erlich, Michael Feolo, Marcelo Fernnandez-Vina, Pierre-Antoine Gourraud, Wolfgang Helmberg, Uma Kanga, Martin Maiers, Hazael Maldonado-Torres, Steven G.E. Marsh, Diogo, Meyer, Derek Middleton, Carlheinz R. Müller, Myoung Hee Park, Richard M. Single, Brian Tait, Glenys Thomson, Ana Maria Valdes, Michael Varney and Jill A. Hollenbach

INTRODUCTION

The goal of the immunogenomics data-analysis working group (IDAWG) is to foster consistent analytical interpretation of immunogenetic data by the immunogenomics and larger genomics communities. Comprised of investigators from five continents, the working group aims to develop a set of community standards intended to facilitate the sharing of these data (HLA, KIR, etc.) and analyses, as well as develop novel methods and tools for immunogenomic data management and analysis.

While a wealth of immunogenomic data resources are available, obstacles to the synthesis of information across datasets have limited the majority of these data to individual studies. As a result, the research community is not able to realize the full potential of these immunogenomic data resources for disease association, hematopeotic stem cell transplantation, and population genetics studies; these topics could be investigated with much greater power and efficiency through meta-analyses, replication studies, data pooling, and combined controls.

Current barriers to the synthesis of data resources include varying levels of typing resolution between data sets, temporal nomenclature variation, and a lack of standards with regard to data capture and processing. Additional statistical and computational challenges exist for consistent analytical interpretation due to the high levels of polymorphism associated with immunogenomic data.

The immunogenomic data-analysis working group (IDAWG) aims to foster the expanded use of new and extant immunogenomic resources through the development of methods and standards for the integration of data generated at different levels of typing resolution, using different methodologies and under different nomenclature paradigms, and to facilitate the consistent application of analytical methods to highly polymorphic datasets through the refinement of extant methods and the development of novel approaches to immunogenomic analysis.

SOFTWARE TOOLS

In keeping with the goals of the working group we have developed the Allele Name Translation Tool (ANTT) and Update NomenCLature (UNCL), software tools designed to translate allele names recorded using naming conventions first described as part of the 2002 Nomenclature for Factors of the HLA System(1) to allele names recorded under the impending April 2010 nomenclature and naming conventions (http://hla.alleles.org/announcement.html; Table 1) in an automated fashion. The new HLA allele naming conventions will incorporate colons to explicitly define the domains that specify polymorphisms among serological antigens, protein sequences, synonymous coding nucleotide sequences, and non-coding nucleotide sequences; at the same time, the nomenclature for many allele names will change in non-obvious ways (e.g., DPB1*0502 will change to DPB1*104:01) (http://hla.alleles.org/announcement.html), and the locus identifier for HLA-C locus alleles will change from Cw* to C*.

Table 1. Example of April 2010 nomenclature change for DPB1.

April 2010 HLA-DPB1 Allele Naming Conventions

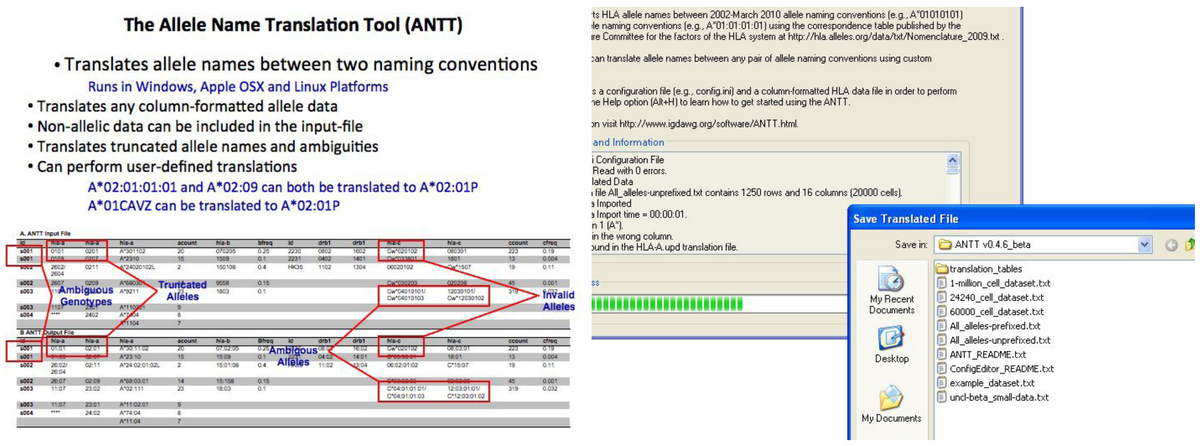
DPB1 alleles names will be renamed to reflect the order in which they were identified (and to avoid the addition of leading zeroes):

DPB1*0101xx	becomes DPB1*01:01:xx
DPB1*0102	becomes DPB1*100:01
DPB1*0201xx	becomes DPB1*02:01:xx
DPB1*0202	becomes DPB1*02:02
DPB1*0203	becomes DPB1*101:01
DPB1*0301xx	becomes DPB1*03:01:xx
DPB1*0302	becomes DPB1*102:01
DPB1*0401xx	becomes DPB1*04:01:xx
DPB1*0402	becomes DPB1*04:02
DPB1*0403	becomes DPB1*103:01

These tools consistently translate allele names to the new nomenclature to ensure that extant HLA data are updated easily and consistently across entire data sets, reducing the likelihood of analytical and reporting errors related to allele-name variation within and between data sets. The ANTT and UNCL translate allele names rapidly, and have been tested on datasets of up to one-million alleles. IDAWG members representing research groups, clinical laboratories and national registries are currently beta-testing the software.

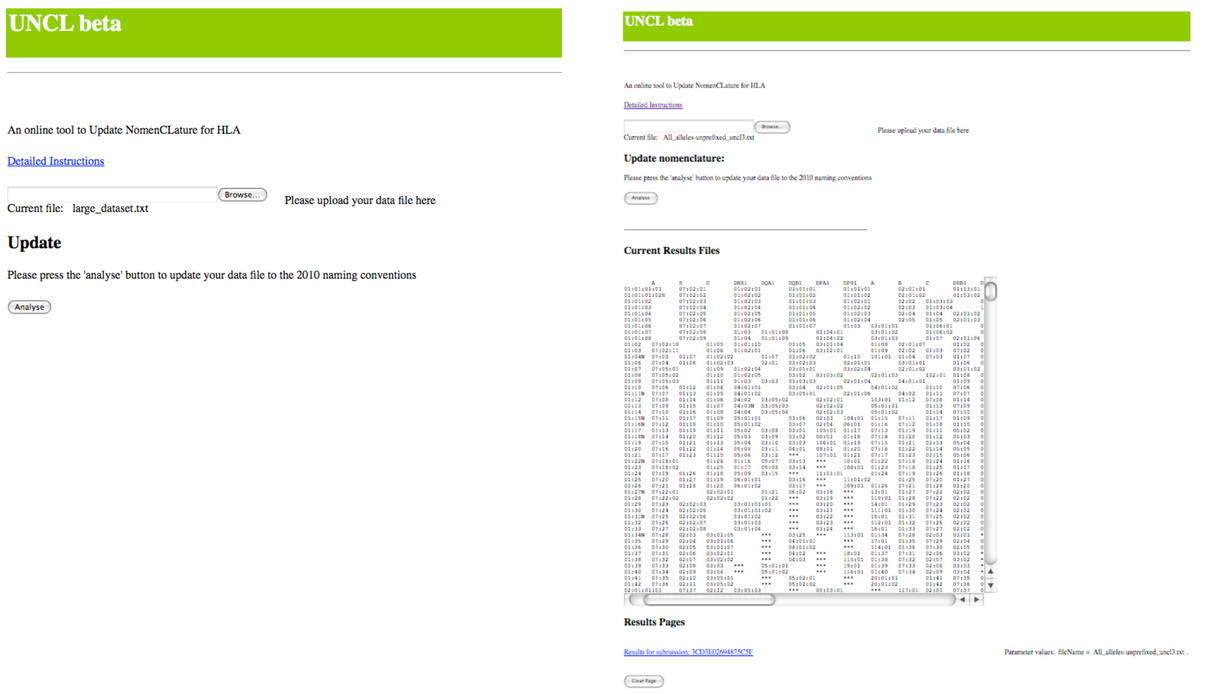
The ANTT is a Microsoft .NET application (2) that runs in Windows, Apple OSX and Linux operating systems; in addition to translating current HLA allele names to 2010 naming conventions, it can translate between pairs of user-defined allele name conventions, e.g., translating pre-2002 allele names to April 2010 names sans intermediate steps.

Figure 1. ANTT translation scheme and screenshot.



UNCL is a platform-independent, web-based tool scripted in R, the open-source language and environment for statistical computing (3). UNCL is designed for speed, utility and accessibility, and requires only a modern web-browser and internet-access to function. UNCL's R-script will be made available for public use. Functionality allowing UNCL to be used as a more flexible user-defined translation tool, as with the ANTT, is under development.

Figore 2. UNCL screenshots. Before and after file translation.



Both tools can translate genotype, allele-count, and allele-frequency data, and will translate files including any non-allelic data, requiring minimal data reformatting. Both tools will be adapted to accommodate impending changes in the KIR nomenclature, recently recommended by consensus in the KIR community.

SUMMARY

The IDAWG's collaborative work began with an inaugural meeting at the 2009 European Federation of Immunogenetics (EFI) meeting in Ulm, Germany, and continued in conjunction with the 2009 International KIR Polymorphism Workshop, in Berkeley, CA. The IDAWG is an ASHI Scientific and Clinical Affairs subcommittee, and convened as such during the pre-conference meeting sessions. The group will continue to meet at the annual EFI, ASHI, and Australian and South East Asian Tissue Typing Association (ASEATTA) meetings each year, to review progress and coordinate ongoing efforts. Finally, the IDAWG is registered as a HuGENet collaborator, allowing our work to be closely aligned with that of the larger genomics community.

We envision a continued collaborative effort by investigators particularly interested in issues of immunogenomics data management and analysis, with participation in specific projects open to the community as part of the 16th International Workshop, and the goal of presenting our recommendations on these topics at the 16th IHWC, followed by the publication of a reference manual and ongoing development of software tools for immunogenomics research. Please visit our website at www.igdawg.org.

REFERENCES

1. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Geraghty DE, Hansen JA, Mach B, Mayr WR, Parham P, Petersdorf EW, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI (2002) Nomenclature for factors of the HLA system, 2002. Tissue Antigens 60:407-464

2. Balena F (2002) Programming Microsoft Visual Basic .NET. Microsoft Press Redmond, WA USA

3. R Development Team (2008) R: A language and environment for statistical computing.