IDAWG: The Immunogenomics Data-Analysis Working Group

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

ABSTRACT

The goal of the immunogenomics data analysis working group 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.

The goal of the working group is to develop methods, standards, tools and recommendations intended to;

1) record, store and transmit immunogenomic data without obscuring the limitations of the typing method used, allow easy identification of allelic equivalency under successive nomenclatures, make data both human-readable (e.g., flat-text file) and machine-readable (e.g., XML file), conform to extant nomenclature rules, all without the use of proprietary platforms;

2) document ambiguity reduction (AR) methods used, permit reproducible AR, and permit equivalency under different AR methods;

3) foster portability between extant analysis tools and methods for maximum access to investigators (e.g., web-based tools);

4) encourage consistent data formats in future analytical methods, promoting widespread accessibility and application;

5) foster methodological consistency in the analysis of low frequency alleles and heterogeneous data, haplotype estimation, Hardy-Weinberg testing of highly polymorphic data, the application of measures of and adjustment for linkage disequilibrium, tests for selection and measures of population differentiation, the calculation of odds ratios, relative risks, etc., corrections for multiple testing, mitigation of false positive readings; and

6) develop novel methods of data analysis for highly polymorphic loci in disease association and population studies (e.g., peptide and nucleotide-level analyses, multidimensional scaling analyses, and neural network analyses).

We envision a collaborative effort by investigators particularly interested in issues of immunogenomics data management and analysis, with the goal of presenting our recommendations on these topics at the 16th IHIWC followed by the publication of a reference manual.

Immunogenomics Data-Analysis Working Group Participants
Jill A. Hollenbach *co-chair*
Steven J. Mack *co-chair*
Center for Genetics, Children's Hospital Oakland Research Institute, Oakland, CA, USA
Henry Erlich
Department of Human Genetics, Roche molecular Systems, Pleasanton, CA, USA
Michael Feolo
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
Marcelo Fernandez-Vina
M. D. Anderson Cancer Center, University of Texas, Houston, TX, USA
Wolfgang Helmberg
Universitätsklinik für Blutgruppenserologie und Transfusionmedizin, Medizinische Universität Graz, Graz, Austria
Uma Kanga
Department of Transplant Immunology and Immunogenetics, All India Institute of Medical Sciences, Ansari Nagar, New Delhi, India
Pawinee Kupatawintu
Thai National Stem Cell Donor Registry, National Blood Centre, Thai Red Cross Society, Bangkok, Thailand
Alex Lancaster
Department of Ecology and Evolutionary Biology, University of Arizona, Tuscon, AZ, USA
Martin Maiers
National Marrow Donor Program, Minneapolis, MN, USA
Hazael Maldonado-Torres
Anthony Nolan Research Institute, Royal Free Hospital, London, UK
Steven G.E. Marsh
Anthony Nolan Research Institute, Royal Free Hospital, London, UK
Diogo Meyer
Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de Sao Paulo, Sao Paulo, Brazil
Derek Middleton *ASHI liason*
Transplant Immunology Laboratory, Royal Liverpool University Hospital, Liverpool, UK
Carlheinz R. Müller
Zentrales Knochenmarkspender-Register fuer die Bundesrepublik Deutschland GmbH, Ulm, Germany
Oytip Nathalang
Thai National Stem Cell Donor Registry, National Blood Centre, Thai Red Cross Society, Bangkok, Thailand
Myoung Hee Park
Department of Laboratory Medicine, Seoul National University Hospital, Seoul, Korea
Richard M. Single
Department of Medical Biostatistics, University of Vermont, Burlington, USA
Brian Tait
The Australian Red Cross Blood Service, Melbourne, Victoria, Australia
Glenys Thomson
Department of Integrative Biology, University of California, Berkeley, CA, USA
Ana Maria Valdes
Twin Research Unit, King's College London, St Thomas' Hospital, London, UK
Mike Varney
The Australian Red Cross Blood Service, Melbourne, Victoria, Australia

Immunogenomics Data-Analysis Working Group Project Rationale

Immunogenetics is an international field with a history that spans more than fifty years. Immunogenetic genotyping data-generation and data-analysis methods have proliferated in this time, in many cases in inconsistent ways. We have identified five areas where (often unavoidable) inconsistency can be introduced into genotype data; these inconsistencies in turn contribute to variation in the analysis of immunogenomic data, and the interpretation of such analyses.

Challenges To Consistency In Data Management And Analysis

1) Variation in Typing Methodology
The large variety of typing methods in use results in datasets being generated with varying levels of resolution, so that a particular allele (or set of equivalent alleles) may be identified in a sample using one method, while a slightly different allele (or set of equivalent alleles) may be identified in the same sample, using a different method.

2) Changes in Nomenclature
The continual evolution of the nomenclature conventions over time has resulted in a progression of allele and locus identifiers that are not always easily inter-related. It is usually the case that the nomenclature used to identify the alleles in a dataset is "frozen" at the time when that dataset is generated and published, which can make comparisons with datasets generated under successive nomenclature conventions difficult.

3) Variation in Data Management Standards
The lack of clear standards with regard to the manner in which immunogenetic data are recorded and stored often makes it difficult to integrate datasets for meta-analyses, or to even share data between research groups. This becomes especially problematic with very large datasets, as the process of reformatting data is often accomplished by hand.

4) Variation in Ambiguity Reduction Methods
When polymorphisms that distinguish alleles are not assessed (e.g., because they are in an exon that is not interrogated by the typing method employed), the result is allelic ambiguity, where exact identity of one or both of the alleles present in a given sample at a given locus cannot be known. When it is not possible to establish phase between key polymorphisms common to many alleles, the result is genotypic ambiguity, where multiple genotypes are possible for a given sample. The choice of one allele over another, or of one genotype over another, is the process of ambiguity reduction. Currently, there is no standard method for reducing these ambiguities, and different research groups may apply different methods to the same set of ambiguities, resulting in different alleles and genotypes being chosen for the same typing result.

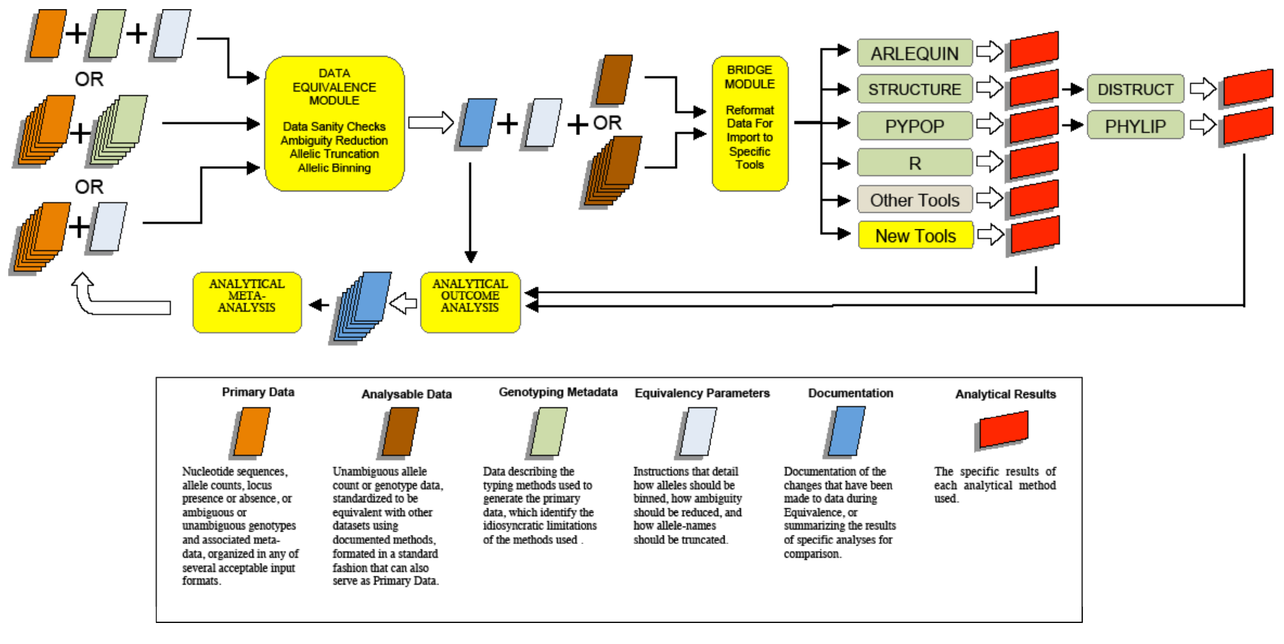
5) High Polymorphism
The high level of polymorphism associated with these genetic systems presents particular bioinformatic, statistical and computational challenges that have yet to be addressed in a standardized manner. For example, haplotype estimation, case-control association studies and test of fit to HWE are all subject to biases introduced by large numbers of low-frequency alleles (aka, sparse cells). Guidelines based on theoretical and empirical considerations are required for each method in order to insure consistency in the application and interpretation of these analyses.

Solutions for Immunogenomic Data Management and Analysis
Given these challenges to consistent data-analysis, the Immunogenomics Data-Analysis Working Group proposes to develop data equivalency standards intended to foster consistency in the use of extant and future analytical methods, and to develop novel statistical and computational methodologies for the analysis of highly polymorphic loci.

In addition, we will determine the impact of various standards and methods for data management on downstream data-analyses, comparing them to extant immunogenetic data analysis systems, and producing recommendations for consistency in the analysis of highly polymorphic datasets.

Finally, we will promote the widespread accessibility and application of novel data equivalency and analytical tools by making them available to the community using web-based and multi-platform approaches.

Figure 1. Proposed Data Management and Analyisis Flow Scheme



For more information, visit *www.igdawg.org*.