Toward a Community Standard for Immunogenetics Data Reporting and Analysis

Jill A. Hollenbach[1], Henry Erlich[2], Michael Feolo[3], Marcelo Fernandez-Vina[4], Pierre-Antoine Gourraud[5,6], Wolfgang Helmberg[7], Uma Kanga[8], Pawinee Kupatawintu[9], Alex Lancaster[10], Martin Maiers[11], Hazael Maldonado-Torres[12], Steven G.E. Marsh[12], Diogo Meyer[13], Derek Middleton[14], Carlheinz R. Müller[15], Oytip Nathalang[9], Myoung Hee Park[16], Richard M. Single[17], Brian Tait[18], Glenys Thomson[19], Ana Maria Valdes[20], Mike Varney[18] and Steven J. Mack[1]

[1]Center for Genetics, Children's Hospital Oakland Research Institute, Oakland, CA, USA,[2]Department of Human Genetics, Roche molecular Systems, Pleasanton, CA, USA ,[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA ,[4]M. D. Anderson Cancer Center, University of Texas, Houston, TX, USA ,[5]Registre France Greffe de Moelle, Agence de La Biomédecine, Paris, France,[6]Department of Neurology, University of California, San Francisco, CA, USA ,[7]Universitätsklinik für Blutgruppenserologie und Transfusionmedizin, Medizinische Universität Graz, Graz, Austria ,[8]Department of Transplant Immunology and Immunogenetics, All India Institute of Medical Sciences, Ansari Nagar, New Delhi, India ,[9]Thai National Stem Cell Donor Registry, National Blood Centre, Thai Red Cross Society, Bangkok,Thailand ,[10]Department of Ecology and Evolutionary Biology, University of Arizona, Tuscon, AZ, USA,[11]National Marrow Donor Program, Minneapolis, MN, USA ,[12]Anthony Nolan Research Institute, Royal Free Hospital, London, UK,[13]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de Sao Paulo, Sao Paulo, Brazil ,[14]Transplant Immunology Laboratory, Royal Liverpool University Hospital, Liverpool, UK ,[15]Zentrales Knochenmarkspender-Register fuer die Bundesrepublik Deutschland GmbH, Ulm, Germany ,[16]Department of Laboratory Medicine, Seoul National University Hospital, Seoul, Korea , [17]Department of Medical Biostatistics, University of Vermont, Burlington, VT, USA ,[18]The Australian Red Cross Blood Service, Melbourne, Victoria, Australia , [19]Department of Integrative Biology, University of California, Berkeley, CA, USA and [20]Twin Research Unit, King's College London, St Thomas' Hospital, London, UK**

Introduction

In genomics research a consensus is emerging regarding the need for community data-reporting and analysis standards in genetic disease association studies. The recent STREGA (Strengthening the Reporting of Genetic Association studies) statement is an important advance in these efforts, defining specific areas in which adoption of community standards can improve the consistent interpretation of genetic studies, particularly for genome-wide association studies. While many data-reporting issues described in STREGA pertain to immunogenomic studies, the high level of polymorphism associated with these data requires specific, and additional, standards and recommendations. The wide-spread use of high-throughput genotyping methods, which generate data at differing levels of resolution, and the ongoing identification of additional allelic diversity require consistent approaches to managing and analyzing immunogenomic data that will permit synthesis across datasets generated at different times and using different methods. The Immunogenomic Data Analysis Working Group (IDAWG) is comprised of investigators from five continents whose collective experiences make clear the need for community-wide immunogenomic data reporting and analysis standards. We propose that community data-management standards are needed to promote transparency in the recording and dissemination of HLA and KIR genotype data and associated ambiguities, and for maintaining the long-term utility of genotype data. In addition, analytical standards for the treatment of rare alleles, haplotype estimation, and donor-recipient matching algorithms would greatly benefit the immunogenomics community. A STREGA-like set of standards, specifically tailored to immunogenomic data, will facilitate donor-recipient matching, the use of combined controls, pooled data, cross-population analyses, replication studies and meta-analyses with greater power and efficiency, and increase the utility of these important data resources.

- **Challenges To Consistency In Immunogenomic Data Management And Analysis**
  - Variation in Typing Methodology
  - Changes in Nomenclature
  - Variation in Data Management Standards
  - Variation in Ambiguity Reduction Methods
  - High Polymorphism

- **Proposed Solutions for Immunogenomic Data Management and Analysis**
  - Develop data equivalency standards intended to foster consistency in the use of extant and future analytical methods
  - Develop novel statistical and computational methodologies for the analysis of highly polymorphic loci
  - Determine the impact of various standards and methods for data mangement on downstream data-analyses
  - Produce recommendations for consistency in the analysis of highly polymorphic datasets
  - Promote widespread accessibility and application of these novel data equivalency and analytical tools using web-based and multi-platform approaches

Example: Challenges in Consistent Ambiguity Reduction

Most analysis tools require only two allele assignments per locus, per individual, but in the case of HLA data, the typing laboratory must resolve allele and genotype ambiguity to make these assignments. Allele ambiguity results when polymorphisms that distinguish alleles fall outside of assessed regions, while genotype ambiguity results from an inability to establish phase between identified polymorphisms (see Table 1). Although both types of ambiguity can occur in all typing systems (including sequence based typing (SBT) methods), "high resolution" typing systems, generate less ambiguous HLA data. There are currently no standards for making these allelic assignments; they are based on the individual investigator's accumulated and empirical knowledge of the data under study, linkage disequilibrium (LD) patterns, etc. Thus, allele assignments for the same specimen may vary between laboratories.

Table 1. Ambiguous and Unambiguous HLA Allele and Genotype Data

Sample	SSOP-determined Genotypes		Ambiguity-reduced Genotypes[c]		SBT-determined Genotypes[d]	
	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2
001[a]	3501, 3507, 3511, 3523[b]	400101, 400102[b]	3501	4001	3501	400101, 400102[b]
001[a]	3520	4007				
002[a]	1525	390101, 390103, 3905[b]	1525	3901	1525	390101, 390103[b]
002[a]	1521	390201, 390202[b]				
003[a]	3501, 3507, 3511, 3523[b]	4002	3501	4002	3501	4002
003[a]	3502, 3504, 350901, 350902[b]	4003				
003[a]	3515	4005				
004[a]	150101, 1526N, 1533[b]	4601	1512	4601	1512, 1519[b]	4601
004[a]	1512, 1519[b]	4601				
004[a]	1532, 1535[b]	4601				
005[a]	1521	390101, 390103, 3905[b]	1521	3901	1521	390101, 390103[b]
005[a]	1521	3910				

a. Ambiguous genotypes are presented as multiple rows for a given sample.
b. Ambiguous alleles are presented as comma-separated strings of allele names.
c. SSOP-determined genotypes have been reduced to two allele assignments per specimen using our ambiguity reduction (AR) method.
d. SBT methods were used to confirm the AR method, but still yield allelic ambiguity.

Summary

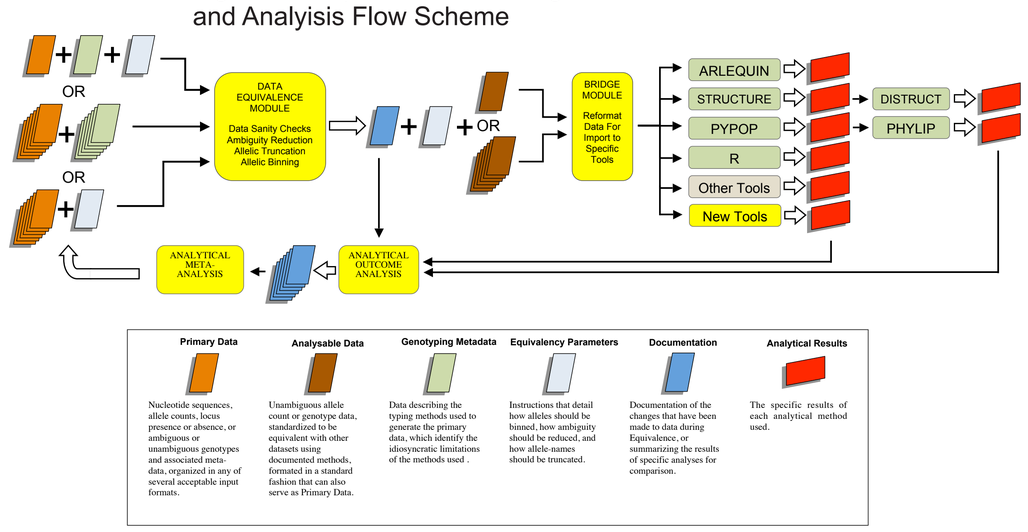
We propose development of data equivalency standards intended to foster consistency in the use of extant and future analytical methods, and to develop novel statistical and computational methodologies for the analysis of highly polymorphic loci.

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

The goal is to promote widespread accessibility and application of these novel data equivalency and analytical tools by making them available to the community using web-based and multi-platform approaches.

A schematic depicting an integrated data management and analysis strategy is shown in Figure 1.

Figure 1. Proposed Data Management and Analyisis Flow Scheme



Reference

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J, Scheet P, Gwinn M, Williamson RE, Zou GY, Hutchings K, Johnson CY, Tait V, Wiens M, Golding J, van Duijn C, McLaughlin J, Paterson A, Wells G, Fortier I, Freedman M, Zecevic M, King R, Infante-Rivard C, Stewart A, Birkett N. (2009) STrengthening the REporting of Genetic Association studies (STREGA)--an extension of