Human Immunology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Human Immunology

journal homepage: www.elsevier.com/locate/humimm

Collection and storage of HLA NGS genotyping data for the 17th International HLA and Immunogenetics Workshop

Chia-Jung Chang[a], Kazutoyo Osoegawa[b], Robert P. Milius[c], Martin Maiers[c], Wenzhong Xiao[a,d], Marcelo Fernandez-Viña[b,e], Steven J. Mack[f,*]

[a] Stanford Genome Technology Center, Palo Alto, CA, USA
[b] Histocompatibility, Immunogenetics & Disease Profiling Laboratory, Stanford Blood Center, Palo Alto, CA, USA
[c] Bioinformatics Research, National Marrow Donor Program, Minneapolis, MN, USA
[d] Massachusetts General Hospital and Shriners Hospital for Children, Boston, MA, USA
[e] Department of Pathology, Stanford University Medical Center, Stanford, CA, USA
[f] Center for Genetics, Children's Hospital Oakland Research Institute, Oakland, CA, USA

ARTICLE INFO

Keywords:
International Workshop
17th IHIW
Next generation sequencing
HLA
Database
Data management
XML
HML

ABSTRACT

For over 50 years, the International HLA and Immunogenetics Workshops (IHIW) have advanced the fields of histocompatibility and immunogenetics (H&I) via community sharing of technology, experience and reagents, and the establishment of ongoing collaborative projects. Held in the fall of 2017, the 17th IHIW focused on the application of next generation sequencing (NGS) technologies for clinical and research goals in the H&I fields. NGS technologies have the potential to allow dramatic insights and advances in these fields, but the scope and sheer quantity of data associated with NGS raise challenges for their analysis, collection, exchange and storage. The 17th IHIW adopted a centralized approach to these issues, and we developed the tools, services and systems to create an effective system for capturing and managing these NGS data. We worked with NGS platform and software developers to define a set of distinct but equivalent NGS typing reports that record NGS data in a uniform fashion. The 17th IHIW database applied our standards, tools and services to collect, validate and store those structured, multi-platform data in an automated fashion. We have created community resources to enable exploration of the vast store of curated sequence and allele-name data in the IPD-IMGT/HLA Database, with the goal of creating a long-term community resource that integrates these curated data with new NGS sequence and polymorphism data, for advanced analyses and applications.

1. Introduction

1.1. The Histocompatibility Workshops

Since their introduction in 1964, the Histocompatibility Workshops have been forums for the exchange of community knowledge and experience, allowing histocompatibility and immunogenetics (H&I) researchers, clinicians and technologists to evaluate new methods and technologies, establish standards and advance ongoing collaborative projects. Sixteen International HLA and Immunogenetics Workshop (IHIW) meetings have been held on five continents over the last half-century [1–16], and the 17th IHIW was held in northern California in the fall of 2017, continuing many long-standing workshop projects.

The advent of next-generation sequencing (NGS) based genotyping technologies has allowed new insights and innovations for the fields of histocompatibility, immunogenetics and immunogenomics. The 17th IHIW's ultimate goals were to advance H&I basic research and clinical efforts through the application and evaluation of NGS HLA and KIR genotyping technologies, and to foster the development of NGS technologies tailored to meet the H&I community's needs, building on the technological and scientific momentum of the previous sixteen workshops.

Abbreviations: CSV, Comma-Separated Values; GFE, Gene Feature Enumeration; GL, Genotype List; HLA, Human Leukocyte Antigen; HML, Histoimmunogenetics Markup Language; H&I, Histocompatibility and Immunogenetics; IHIW, International HLA and Immunogenetics Workshop; IMGT, ImMunoGeneTics; IPD, ImmunoPolymorphism Database; IUPAC, International Union of Pure and Applied Chemistry; KIR, Killer-cell Immunoglobulin-like Receptor; MIRING, Minimum Information for Reporting Immunogenomic NGS Genotyping; NGS, Next Generation Sequencing; PI, Principal Investigator; RMAN, Recovery Manager; RSCA, Reference Strand Conformation Analysis; rSSO, Reverse Sequence-Specific Oligo; SBT, Sequence-Based Typing; sFTP, secure File Transfer Protocol; SS, Sequence-Specific; SSO, Sequence-Specific Oligo; SSP, Sequence-Specific Priming; WMDA, World Marrow Donor Association; WS, Workshop; XML, eXtensible Markup Language

* Corresponding author.
E-mail address: sjmack@chori.org (S.J. Mack).

https://doi.org/10.1016/j.humimm.2017.12.004
Received 18 July 2017; Received in revised form 12 November 2017; Accepted 8 December 2017

Toward those ends, we developed systems, standards and tools for the collection, storage and management of NGS HLA genotyping data (the HLA genotype and associated consensus sequences) generated for 17th IHIW projects. The goals of this effort were to build on the data-collection and -storage experiences of previous workshops, and produce NGS data-managing tools that will support IHIW efforts and persist as public resources after the 17th IHIW. Here, we provide a brief overview of the challenges faced in organizing coordinated data-generation and -collection efforts, the strategies we applied, and the tools, standards and services we developed to address these challenges.

1.2. The challenges of coordinated data collection

The collection, storage and analysis of data have been key issues of all workshops. Many workshops have used centralized databases [17–21], while in several of the more recent workshops, individual components and projects were responsible for collecting, managing and analyzing data [22–32]. Centralized data-management requires close communication between workshop participants and leaders, instrument and software vendors, and database developers to achieve consensus regarding required data content, data formats, reporting guidelines and quality standards. Sufficient time is also required for all parties involved to develop both the systems and tools to manage data, and the preliminary data on which to test the tools.

1.2.1. Reference data management

The specifics of the H&I field bring additional challenges that any data-management and analysis approach, centralized or decentralized, must address [33]. The body of HLA sequence data and associated allele names curated by the IPD-IMGT/HLA Database [34] (Reference Database) increases every four months; because workshop data-generation efforts often span multiple years, the details of the pertinent Reference Database version under which each HLA genotype was generated must be collected along with the genotyping data. The collection and management of such genotyping meta-data (Table 1) can be just as important for the workshop effort as the genotyping data themselves; without them it may not be possible to determine the extent to which datasets generated years apart or using different methods are equivalent. When workshop efforts span time periods that include major changes to the nomenclature [35,36], these problems are only compounded.

1.2.2. Primary data management

The nature of the primary or "raw" data, from which all experimental data and meta-data are ultimately derived, can vary widely from method to method and from project to project. This was particularly pronounced for the molecular genotyping methods applied in the 11th through the 16th workshops, where multiple reference strand conformation analysis (RSCA), sequence-specific (SS) oligo (SSO), reverse SSO (rSSO), SS priming (SSP) and sequence-based typing (SBT) methods were in use, each with its own distinct type of primary data.

1.2.3. Allele name data management

Allele name data must be recorded and managed in a standard manner to facilitate meaningful data-analysis. For many previous workshops, the management of HLA allele names was performed by humans, and involved data recorded in paper documents or spreadsheets in a variety of different ways. Humans are adept at "figuring out" the true meaning of unusual notations and spreadsheet-initiated errors that may occur, but machines are not. For example, "HLA-A∗02:99″ and "HLA-A∗03:01:02″ are often recorded as "02:99″ or "03:01:02″ in spreadsheet columns labeled "HLA-A", "A", etc.; however, common spreadsheet applications may change "02:99″ to "0.152083333333333″ or "3:39″, and "03:01:02″ to "3:01:02″, all of which erroneously represent times instead of alleles. The range of potential human-generated transcription errors is large. Previous workshops devoted considerable manual effort to review, identify and correct errors, and standardize allele-name notations prior to analysis. However, the analysis, collection, exchange and storage of NGS genotyping data requires machines (computers) that are able to process allele name data, and the accompanying nucleotide sequence data, without the human ability to identify and correct errors.

1.2.4. Describing novel polymorphism

The description of previously unknown (novel) HLA sequence variants has been a long-standing challenge for the H&I community. Until a novel sequence is named by the World Health Organization Nomenclature Committee for Factors of the HLA System (Nomenclature Committee) [37], it is very difficult to discuss that sequence in the context of the HLA nomenclature. The common practice, associated with pre-NGS genotyping, has been to append a "novel-allele" identifier to a truncated version of a related allele name (e.g. "HLA-A∗02 V", "HLA-A∗02:NEW", "HLA-A∗02:01new", etc.). The World Marrow Donor Association guidelines for the use of HLA nomenclature (WMDA guidelines) indicate that "NEW" should be reported for alleles that have not been named by the Nomenclature Committee [38]. However, the absence of a standard for describing novel HLA alleles and associated nucleotide sequences represents a considerable challenge for the collection of NGS HLA genotyping data.

2. Meeting the challenge

The 17th IHIW adopted a centralized data-storage approach, in which all specimen-related data, reference data, genotyping data and associated meta-data were stored in a single database system. The goal of this effort was to facilitate data and analysis access for workshop participants, with these workshop products and the database itself made available to the H&I community after the workshop. The 17th IHIW focus on NGS provided an advantage for centralized data collection in that there are currently only a small number NGS platforms, which generate primary data in FASTQ [39] format, and associated genotyping software. A key 17th IHIW goal was to collect machine-generated HLA data for consumption by IHIW informatics services, with minimal human intervention. We worked with NGS software developers to define a small number of equivalent and interchangeable data reporting formats that allowed genotyping data and meta-data to be collected using a "uniform NGS data-collection" approach. This approach built on the genotype list (GL) string format [40] and the GL Service [41], the Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING) reporting guidelines and messaging standard [42], and the MIRING-compliant Histoimmunogenetics Markup Language (HML) version 1.0 messaging format [43].

2.1. Uniform NGS data collection

The 17th IHIW did not require all workshop projects or participating laboratories to use the same NGS platform, typing kit or protocol. NGS instruments manufactured by Illumina (e.g., MiSeq), One Lambda (e.g., S5XL), Pacific Biosciences (e.g., PacBio RSII) and Roche 454 (e.g., GS FLX) were used in 17th IHIW NGS genotyping efforts. The goal in uniform NGS data collection was that all NGS HLA genotyping data and associated meta-data (which together constitute a "typing report") be compatible and comparable, so that all collected data were equally interpretable, regardless of the format in which those data were exchanged. This allowed data generated by different laboratories, in different countries, using different platforms and software, to be stored in one database and made available for multiple projects.

Toward this end, the 17th IHIW accepted NGS genotyping data and meta-data in three MIRING-compliant eXtensible Markup Language (XML) [44] based typing report document formats – HML (version 1.0.1); GenDx XML, exported by GenDx NGS Engine version 2.4.0; and IHIW XML[A], a format developed specifically for the 17th IHIW