Using cloud infrastructure to facilitate data collection and conversion of HLA diagnostic data for the 18th International HLA and Immunogenetics Workshop

Matern, B.M.; Niemann, M.; Nemparis, I.; Schimanski, A.; Peereboom, E.T.M.; Kramer, C.S.M.; ... ; Spierings, E.

Received: 1 November 2022 | Revised: 30 December 2022 | Accepted: 27 January 2023

DOI: 10.1111/tan.14989

ORIGINAL ARTICLE

HLA Immune Response Genetics WILEY

Using cloud infrastructure to facilitate data collection and conversion of HLA diagnostic data for the 18th International HLA and Immunogenetics Workshop

Benedict M. Matern[1] | **Matthias Niemann**[2] | **Ioannis Nemparis**[3] | **Andreas Schimanski**[2] | **Emma T. M. Peereboom**[1] | **Cynthia S. M. Kramer**[4] | **Sebastiaan Heidt**[4] | **Eric Spierings**[1,5]

[1]Center for Translational Immunology, UMC Utrecht, Utrecht, The Netherlands

[2]PIRCHE AG, Berlin, Germany

[3]GenDx, Genome Diagnostics B.V., Utrecht, The Netherlands

[4]Department of Immunology, Leiden University Medical Center, Leiden, The Netherlands

[5]Central Diagnostics Laboratory, UMC Utrecht, Utrecht, The Netherlands

Correspondence
Benedict M. Matern, Center for Translational Immunology, UMC Utrecht, Utrecht, The Netherlands.
Email: b.m.matern@umcutrecht.nl

Funding information
International HLA & Immunogenetics Workshop Foundation; Universitair Medisch Centrum Utrecht

The International HLA and Immunogenetics Workshop (IHIW) is a recurring gathering of researchers, technologists and clinicians where participants contribute to collaborative projects with a variety of goals, and come to consensus on definitions and standards for representing HLA and immunogenic determinants. The collaborative and international nature of these workshops, combined with the multifaceted goals of several specific workshop components, necessitates the collection and curation of a wide assortment of data, as well as an adaptable platform for export and analysis. With the aim of ensuring data quality and creation of reusable datasets, specific standards and nomenclature conventions are continuously being developed, and are an integral part of IHIW. Here we present the 18th IHIW Database, a purpose-built and extensible cloud-based file repository and web application for collecting and analyzing project-specific data. This platform is based on open-source software and uses established HLA data standards and web technologies to facilitate de-centralized data repository ownership, reduce duplicated efforts, and promote continuity for future IHIWs.

KEYWORDS
cloud, database, epitopes, HLA, IHIW, workshop

Abbreviations: API, application programming interface; AWS, Amazon Web Services; CSP, Cloud Service Provider; DASH, Data Standards Hackathon; DMA, IHIW Data Management Application; EC2, Elastic Cloud Compute; GLString, Genotype List String; HAML, HLA Antibody Markup Language; HML, Histoimmunogenetics Markup Language; IHIW, International HLA and Immunogenetics Workshop; KIR, killer cell immunoglobulin-like receptor; MFI, mean fluorescence intensity; MIRING, minimum information for reporting next generation sequencing genotype; NMDP, National Marrow Donor Program; REST, representational state transfer; S3, simple storage service; SAB, single antigen bead.

1 | INTRODUCTION

The International HLA and Immunogenetics Workshop (IHIW) series are a collaborative effort of the scientific community to share knowledge and advance the field of HLA and immunogenetics research. Since its beginning in 1964[1] the IHIW has been a valuable setting to compare studies, share reagents, contribute to global efforts, and come to consensus on identification of transplantation antigens and how HLA and non-HLA (e.g., KIR) concepts can be best represented and shared.

1.1 | Immunogenic analysis of HLA

Early workshops defined the names of the major histocompatibility antigens based on serological compatibility groups, providing essential foundations for matching in transplantation. These antigen groups have been further defined and categorized[2,3] by examining variations in an individual's HLA type, such as single nucleotide polymorphisms (SNPs), insertions, deletions, and recombinations in HLA genetic sequence. HLA genotypes can be further extended to define B-cell[4,5] and T-cell[6] epitopes to allow for studies on immunogenicity, and establish haplotype patterns[7] across the MHC, which in turn facilitates family inheritance studies.[8,9] HLA patterns are also collected across population groups, which sheds light on allele[10] and haplotype[11] frequencies, and variations in individuals from world populations.[12] Studies have also defined and cataloged HLA based on expression levels,[13,14] non-coding or intronic polymorphism,[15,16] as well as observed immunological compatibility.[17]

In addition to characterization of HLA antigens and sequence, studies on immunization events such as transplantations, transfusions or pregnancies, produce longitudinal data on the development of anti-HLA antibodies. These antibody studies are highly influenced by genotyping and matching of patient or donor's HLA phenotype, and while these studies provide important insights, they require the analysis of highly complex and heterogeneous data. In addition to exploring the links of donor-specific antibodies with transplantation outcomes,[18,19] these data can zoom in on which epitopes are likely to lead to an immune response.[20]

As diagnostic methods advance and data collection capabilities expand, the use of computer analysis and bioinformatic methods become crucial. Bioinformatics not only facilitates the cataloging of the hyperpolymorphic nucleotide and peptide HLA sequences, but provides additional lenses through which HLA can be studied. Advanced data structures such as Histoimmunogenetics Markup Language (HML)[21] and Genotype List Strings (GLStrings),[22] with well-defined guidelines such as MIRING[23] are needed to effectively represent and communicate this complex data, and complex computational tools are necessary to aid in data analysis.

1.2 | IHIW aims

The 18th iteration of the IHIW featured a wide range of projects and goals under the umbrella of three major components: Antigenicity & Immunogenicity, Immunogenetics, and Bioinformatics. These projects include, for example, definition of epitope immunogenicity,[24] sequencing of full-length sequences,[25] and linking HLA with SARS-Cov2.[26] However, as the IHIW is a continuous effort with evolving goals, new collaborations and projects are constantly being established and repurposed. Data requirements change not only between workshops, but even during a workshop. Individual projects are unique, and may have their own specialized data analysis requirements, which often forces researchers to set up and maintain purpose-built compute environments, data analysis pipelines and tools from scratch. Each workshop requires a considerable amount of preparation. Valuable efforts have been made in previous workshops[27] to establish data repositories and platforms for analysis, which have been expanded and built upon in the 18th IHIW.

This manuscript describes the preparation and implementation of the data collection filesystem and analysis platform created for the 18th IHIW. We describe the IT infrastructure and cloud-based analysis platform, and how these efforts create a usable, extensible and transferable framework for data analysis. This database has been used in the 18th workshop, but since accumulated data is owned by the workshop as opposed to a single entity, it is created to be carried forward to future workshops without moving infrastructure or cloning resources, in order to reduce the amount of duplicate work introduced by infrastructural tasks.

2 | MATERIALS AND METHODS

The design of the data repository was focused on facilitating data collection, transformation, and analysis in a transferable and customizable infrastructure. Open source tools and software were used whenever possible, and infrastructure was designed using established and popular platforms, such as the full-stack JHipster platform. The website's front-end was created to be flexible and generic; it allows standard administrative functionality such as user, laboratory, and project registration, as well as data sharing functionality such as file upload and management. Most of the project-specific software tools are organized in small components and analysis scripts

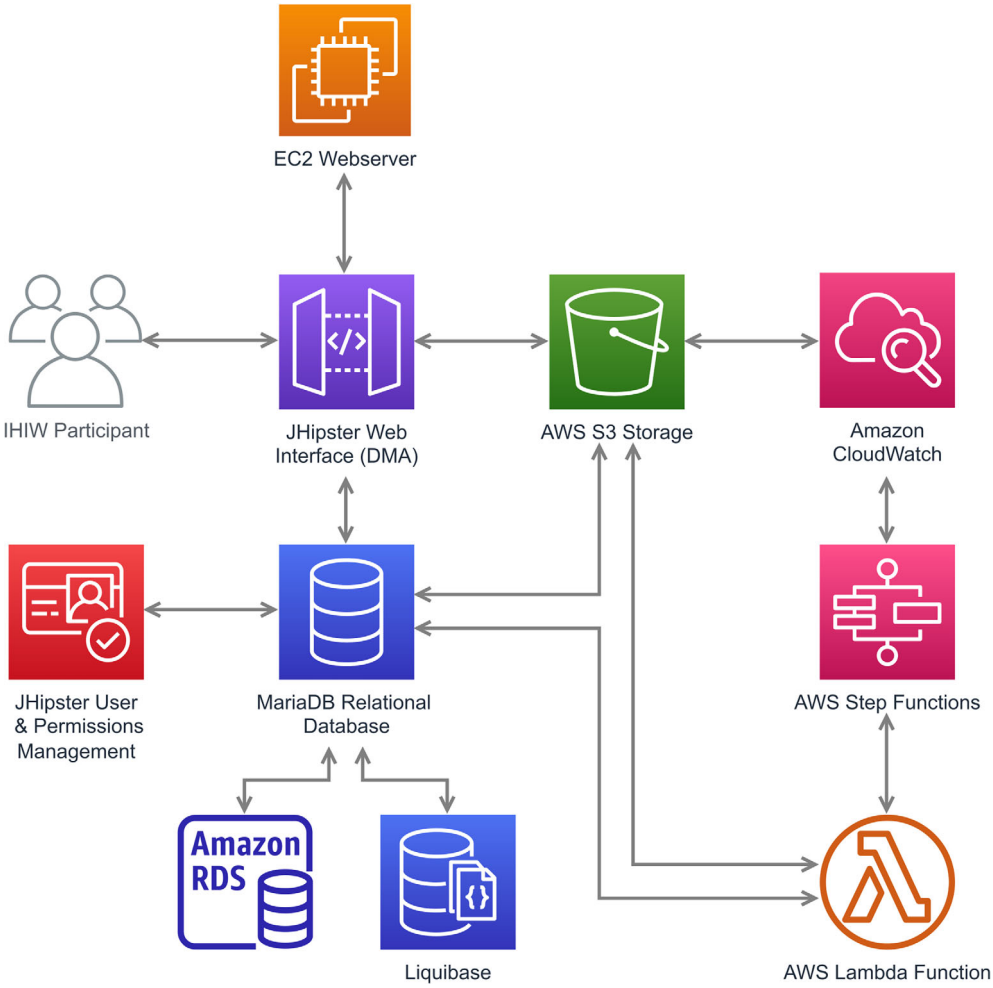

FIGURE 1 IHIW database cloud architecture. Users interact with the Data Management Application (front-end) which is a JHipster Web Application. The MariaDB Relational Database manages user credentials and information about uploaded files. Uploaded files are stored within a S3 bucket, and uploading a file triggers Step Functions using the CloudWatch interface. The Lambda Functions that are managed by a Step Function perform processing on uploaded files, and can send results back to the S3 storage or the MariaDB database. All code is open-source and made available in IHIW github repositories.

within the website's back-end. Separating the general features from the project-specific features allows the design to remain extensible and customizable for specific projects. All components were designed in a cloud-based structure to reduce the dependence on a given physical location and promote transferability.

2.1 | Cloud-based infrastructure

In light of the suggestions and expertise of the participants of the IHIW Database team, the cloud-based web architecture was implemented using Amazon Web Services (AWS) [Amazon Web Services Inc., Seattle, WA, USA], which is a widely-used Cloud Service Provider (CSP). The main website interface, referred to as the Data Management Application (DMA) is hosted on an AWS Elastic Compute Cloud (EC2) instance, which is a flexible cloud-based server. Uploaded data files are stored in an AWS Simple Storage Services (S3) bucket, a scalable and effectively unlimited cloud-based encrypted data storage system. The relational database which stores login, participant, laboratory, project, and upload metadata information exists within a MariaDB database, hosted in an AWS Relational Database Service (RDS). This cloud-based architecture (Figure 1) is scalable and containerized, and although it is currently hosted in European cloud networks, it is portable and can be accessed and modified worldwide.

2.2 | Relational database implementation

The MariaDB relational database structure, shown in Figure 2, was designed for storing user credentials, and can specify permissions for authorized users and membership within participating laboratories or IHIW projects. The database also manages ownership of uploaded files, and is used to specify who can access the uploaded data through permissions. It does not contain the actual workshop data, but rather pointers to the locations of the files. During development, database changes are automatically managed using the Liquibase database version management tool, which is integrated in the JHipster architecture.

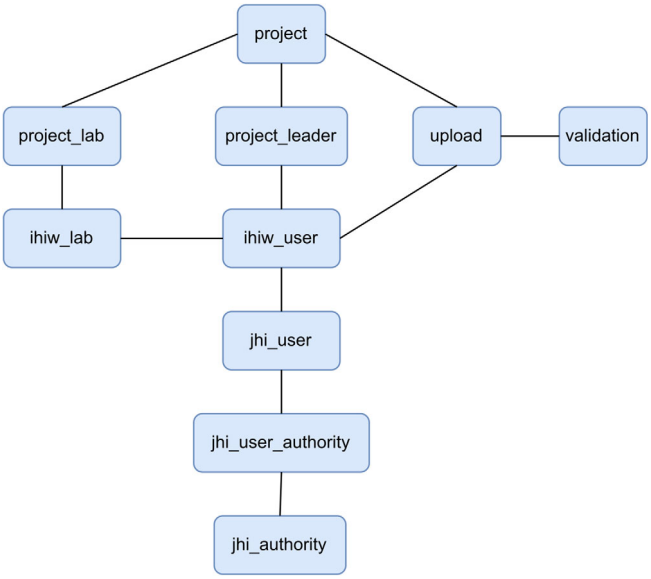

FIGURE 2 MariaDB relational database architecture. This figure shows a simplified diagram of the IHIW relational database architecture, designed to be lightweight and flexible. Each node in this chart represents a table within the database. This design establishes user management ("ihiw_user") and encrypted login data ("jhi_user"), but also IHIW-specific data, including the location of data uploads ("upload"), data validation feedback ("validation"), membership within a participating laboratory ("ihiw_lab"), and subscription of a laboratory to a project ("project_lab"). Data access permissions ("jhi_authority") are managed using a version of the Jhipster permission module which is extended to handle laboratories and projects.

2.3 | Data upload, transformation and validation

Uploaded data within the S3 bucket are encrypted, but have permission-restricted access by the uploader, project leaders, administrators, and lab members. This access can be direct, with participants accessing or modifying uploaded files using the web interface, or programmatically, where scripts within the AWS back-end, using the same authorization system, can modify, validate, upload or delete IHIW files.

The transformation and validation of data was performed using a serverless back-end, implemented using the AWS Lambda architecture. In this design, we do not specify a machine that hosts and performs all of the computation, but this design uses ad hoc on-demand compute time to run the data transformations and validations at the moment they are needed, most often triggered when an IHIW file is uploaded. Although the website front-end runs on a continuous webserver hosted on a single virtual machine, the data-intensive processes are not running during periods when no data is uploaded. This allows scalability, where the storage and processing requirements expand as they are needed.

The validation and transformation processes were encoded as Lambda functions, which are lightweight scripts that run simple logic for a specific purpose. AWS Lambda functions support a wide variety of coding and scripting languages, but for the IHIW back-end these are Python (3.8) scripts. These Lambda functions are orchestrated using AWS Step Functions, and triggered using the AWS CloudTrail and EventBridge interfaces. This Step Function architecture allows the definition of specific state machines (Figure 3) that define how an individual file is analyzed, converted, and validated. The Lambda functions within the Step Function architecture can also access upload and user information from the RDS database, and validation behavior varies depending on the type of submitted data, which participant uploaded it, and what IHIW project it is assigned to.

All analysis code, especially those that analyze HLA genotyping data, will use community-derived open-source tools when applicable. Biopython[28] is commonly used to represent nucleotides and amino acids in sequence data. HLA genotyping is represented as either HML[21] files or as GLStrings.[22] In the case of HML files, these Lambda functions access external REpresentational State Transfer (REST) services to perform data quality validations. Collaborations with National Marrow Donor Program (NMDP) bioinformatics has provided a validator that checks HML genotyping documents for the MIRING[23] metadata checklist standards, and a validator that checks against NMDP gateway requirements. pyGlstring is a package provided by NMDP bioinformatics which provides syntactic validation of GLStrings. The Lambda functions access these validators using standard REST protocols, which are universally used in internet communication.

As files are converted or validated, it is often necessary to provide results or feedback which should be communicated back to the user. This feedback informs the submitter that the data appears to be in the correct format and structure, and can identify any issues in data quality. The DMA interface provides a REST application programming interface (API), containing endpoints that allow authorized users and scripts to interact with the workshop database and filesystem. Using specifically designed REST endpoints within the workshop database application, validation feedback can be written back to the DMA interface and stored within the MariaDB database by authenticated system users. The validation feedback automatically appears within the workshop DMA, as shown in Figure 4C.

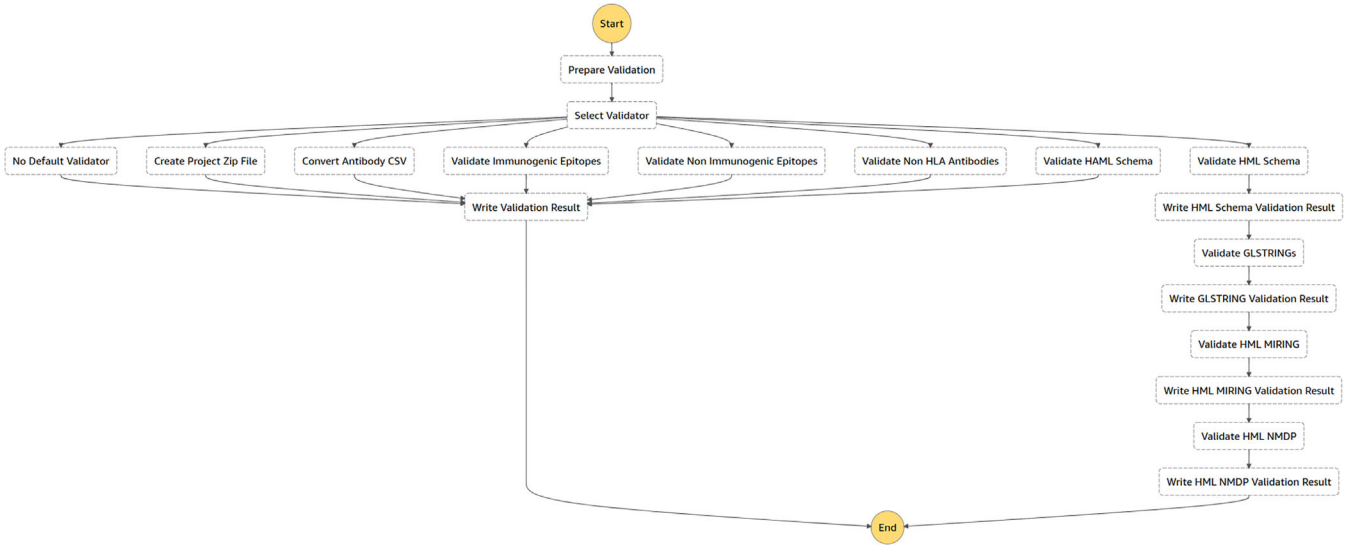

FIGURE 3 Step function architecture for upload conversion and validation. The moment a data file is uploaded to the IHIW database, the AWS cloud architecture triggers analysis using a Step Function design, which is encoded as easily-modifiable JSON. The Step Functions can be thought of as an analysis state machine, where uploaded data is passed between each step in the flowchart. Every uploaded data file is validated or converted differently depending on what type of data it contains, and which project the data is assigned to. Each node in this flowchart represents either a decision, or a Python Lambda function which performs a specific task. When validation or conversion is finished, or if a problem is detected, the Step Function returns feedback back to the IHIW database website to display to the user.

2.4 | HLA antibody data conversion

A notable use of the Lambda and Step Function architecture is the specific conversion of HLA Luminex Single Antigen Bead (SAB) antibody files to the HLA Antibody Markup Language (HAML) format. Data exports from vendor software used to analyze HLA antibody data vary, and while varying formats contain valuable antibody data, such as HLA specificity and mean fluorescence intensity (MFI) values, their structures are not consistent with each other. Efforts to establish a standardized data format for representing antibody result data are ongoing, and an interim IHIW version of the HAML format (https://github.com/IHIW/Converters/tree/master/XmlValidator/schema) was established in order to provide consistency in antibody analysis for IHIW projects. For analysis within current IHIW Projects, the HAML format minimally requires lot/catalog numbers, bead identifiers, HLA specificity, numerical MFI measurements, and also per-sample MFI for positive and negative control beads.

Python Lambda function scripts were written to convert data exported from One Lambda HLA Fusion [Thermo Fisher Scientific Inc., Waltham, MA, USA] and Immucor Match It [Immucor Inc., Norcross, GA, USA] vendor software to HAML, and organized within the Step Function architecture. As seen in Figure 3, when an Antibody CSV file is uploaded, it is immediately converted to a HAML file, and as a final check, is validated against the IHIW HAML .xsd schema file. Feedback from the conversion process, including success messages or any encountered problems, are communicated back to the website and displayed to the submitting user, similar to Figure 4C.

2.5 | Data download

As individual files are uploaded by participating laboratories, they are assigned to a single IHIW project. Participants and project leaders can download individual files, and project leaders can also create and download a .zip file containing summarized project files. This is available to project leaders by visiting the project's [View] page under the DMA Projects entity. There is a button with the text [Create Project Summary Zip], which triggers the creation of a project summary .zip file. This .zip contains the collected uploads, sorted into folders based on the submitting laboratories, as well as a text summary of all project files, specifying the submitting user and file sizes. The .zip becomes available for the project leader to download after a few minutes from the [Uploads] page.

2.6 | Data analysis

For online data analysis, project data is combined and analyzed using project-specific summary scripts, written

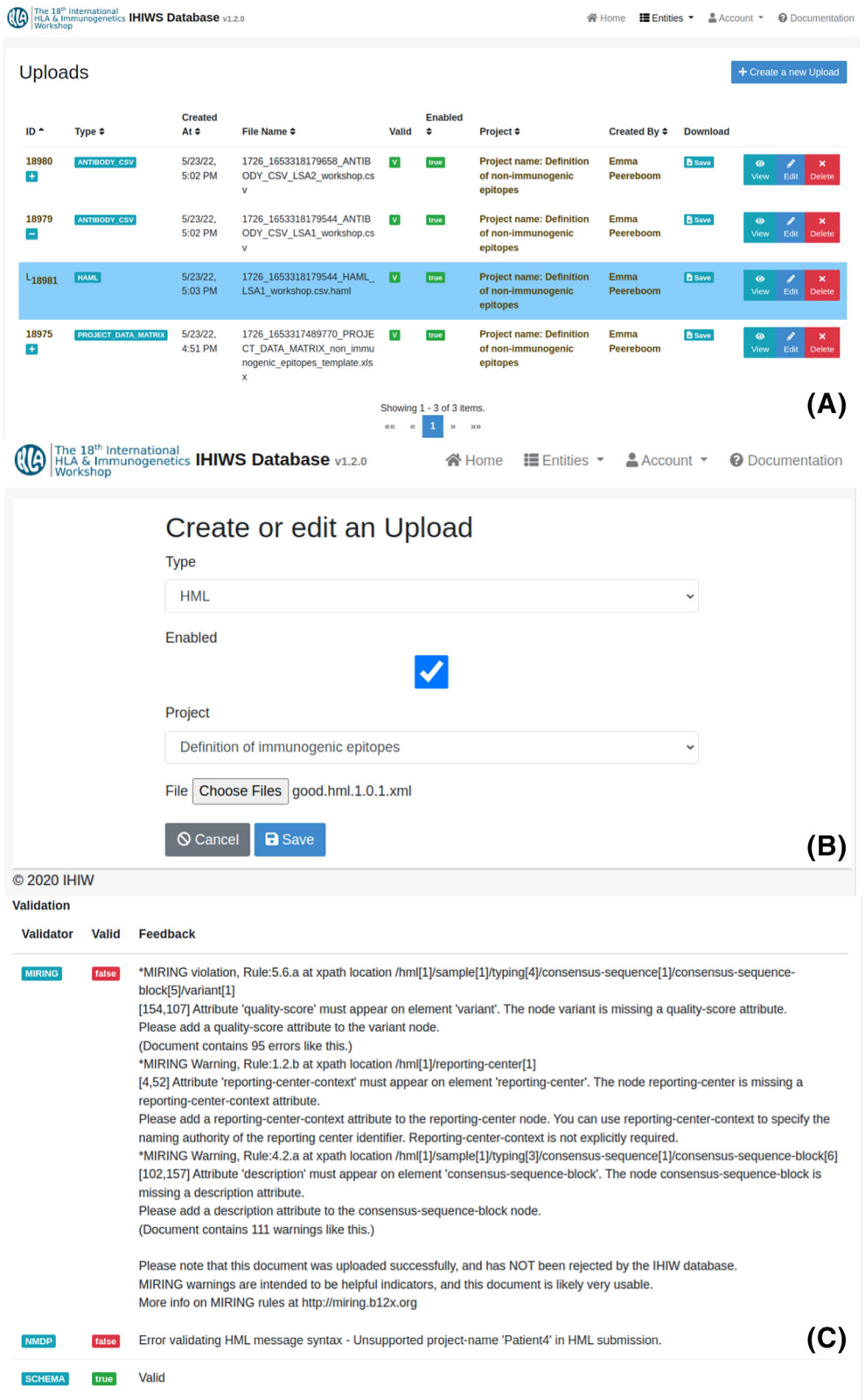

FIGURE 4 Screenshots of the DMA upload interface. Screenshot (A) shows a list of data files which have been uploaded in the database. Visibility of uploads is controlled by the IHIW database user permissions, and standard users have permission to see files created by themselves and members of their own laboratories. Panel (B) shows options that specify which data type and project corresponds to the currently uploaded file(s). Panel (C) shows an example of validation feedback for an HML file containing HLA genotyping results.

in Python. These scripts iterate through the uploaded files that are assigned to a project, to summarize and provide reports to project leaders. A notable example is seen in the case of the Immunogenic Epitopes projects, where participants uploaded data corresponding to transplantations, and provided supporting HLA genotyping and antibody data. In this case, the analysis scripts collect the patient and donor genotyping data, combine it with the antibody results before and after transplantation, and combine the results together in reports. Combined with the uploaded data files, these

reports are used to draw conclusions about the immunogenicity of specific HLA alleles, and work to infer which epitopes may be likely leading to an immune response.

3 | RESULTS

3.1 | Data management application and back-end

The JHipster framework, a widely adopted platform for rapid website development, was applied to generate a slim web application featuring user administration functionality and role-based access control, referred to as the IHIW Data Management Application (DMA). The generated vanilla JHipster code was adjusted to meet the requirements of the IHIW, such as registering participating laboratories, specifying the IHIW components and projects, managing uploads and validations, and adopting access permissions to facilitate access for lab members and project leaders. The DMA stores its information in a MariaDB relational database management system, which is hosted within AWS RDS. The web service runs on a cost-efficient low-performance virtual private cloud compute instance within AWS EC2 (data.ihiws.org). For development purposes, a staging environment was also set up (staging.ihiws.org).

The DMA gives access to users to upload files from their local environment to the IHIW cloud data storage. Using the upload service, project leaders can invite participants to share their project data, which can be combined and analyzed jointly with the data from other participants. Metadata and file pointers to the uploaded files are stored in the relational database, and the data and file contents are stored in a simple file storage service (AWS S3). As users upload a data file, the validation and conversion step functions are triggered via AWS Eventbridge, and they can often see validation feedback and conversion results within a few seconds (Figures 3 and 4).

The JHipster architecture features a built-in API interface for accessing and modifying website data. This API is built upon RESTful architecture, so external scripts that use standard REST methods can access these endpoints. These endpoints are all fully secured using the website's built in access permissions, and require valid user credentials with specific permissions to access. The built-in endpoints allow reading and writing of specific data elements, such as creating an upload or accessing laboratory information. The standard endpoints were extended using specifically designed purposes. For example, one endpoint allows us to assign a validation status for an upload, and another for assigning "child" uploads that use an established parent–child data upload relationship for converted files.

3.2 | User management

User management within a standard JHipster website facilitates common user login features, such as names, email addresses, and encrypted passwords. To adapt to the needs of the IHIW DMA, this functionality was extended to allow storing and linking of IHIW participant-specific information. Specifically, users would normally register information on which laboratory or group they belong to, and what other users belong to this group. Similarly, the DMA stores information about individual projects hosted within the IHIW context, and individual users are assigned as, for example, the head of a laboratory, or the leader of an IHIW project. Defining the "roles" of these users gives some control over which features of the website (such as editing a project or adding lab members). Each user has a password, used to log into the DMA interface, which is encrypted and stored within the relational database. After logging into the website, depending on a user's role, they are able to upload and review their data. Storing user and role information within the JHipster relational database structure clarifies which laboratories are participating in the individual projects, and allows assigning permissions to users to upload data to the projects. Furthermore this structure allows defining new roles and use cases, and enables implementing new data access use cases as IHIW needs may change moving forward.

IP blocking and a login cooldown was implemented to add an extra layer of website security. In one case, as is common in modern websites, some questionable traffic was observed in AWS web hosting logs which seemed to indicate that some scripts were repeatedly sending requests to the IHIW web servers. In response we implemented a login attempt cooldown, and subsequent IP address blocking. When repeated unsuccessful attempts to login occur, the originating IP address is temporarily blocked.

3.3 | Data collection

The IHIW DMA allows the upload of data files in a scaled data storage system. As data has been collected for the IHIW projects, the storage requirements grow with every transmission. As is common with CSP data storage, storage within the S3 bucket is billed in a pay-what-you-consume fashion, where costs increase and decrease

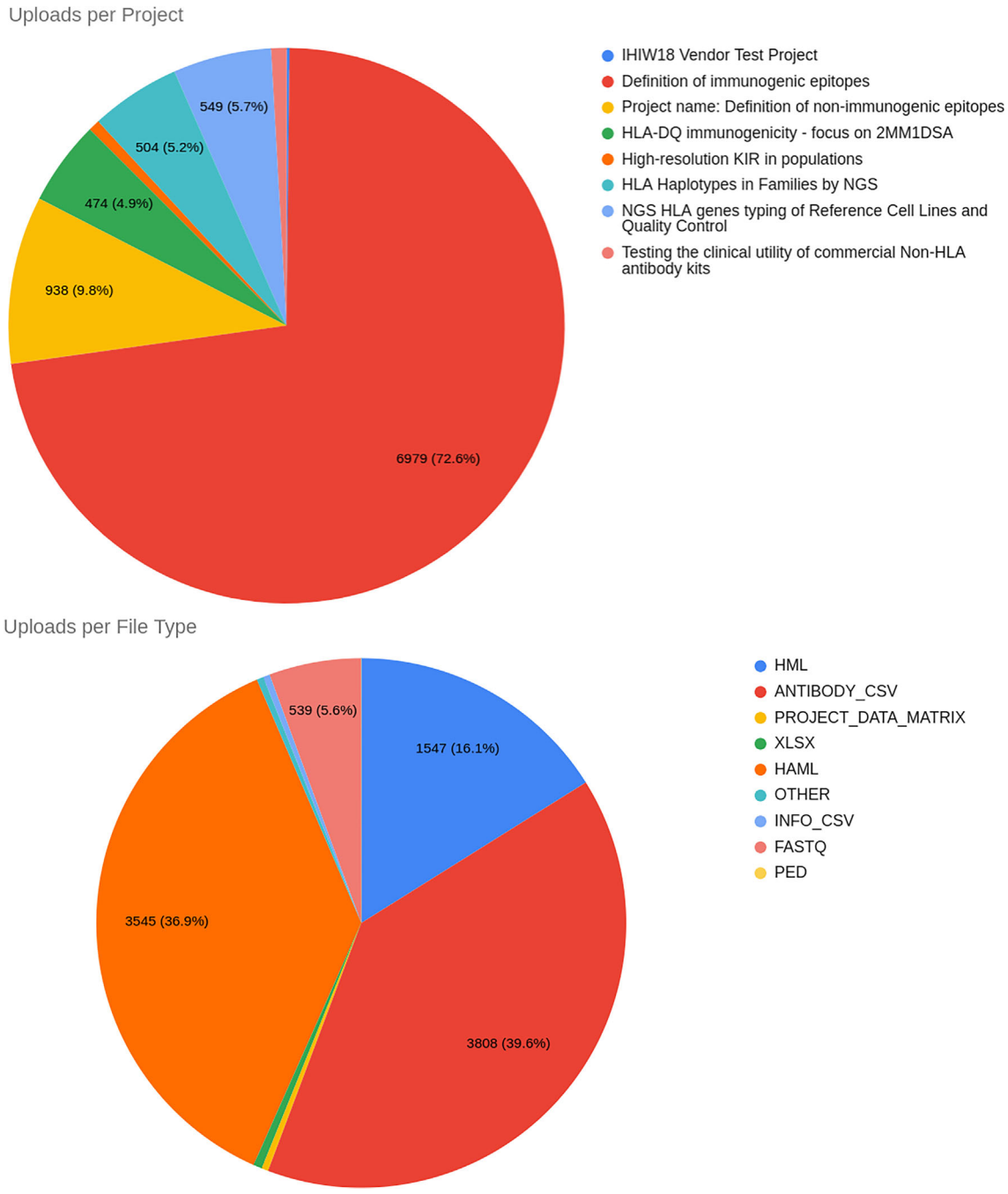

FIGURE 5 Overview of uploaded files in the IHIW database. As files are uploaded using the DMA, the IHIW participant assigns the file to a specific project and specifies the upload type. Panel (A) shows the number and percentages of uploads stratified by project. 72.6% of the uploaded files (by count) were for the Definition of Immunogenic Epitopes project. (Red) Panel (B) shows the number of uploaded files stratified by file type. The majority of files are HML (Blue), and Antibody CSV files (Red) which the database automatically converts to HAML files (Orange).

based on the used storage, as opposed to a fixed price of a large file storage.

When a data file is uploaded to the IHIW DMA, the user chooses a few data options (Figure 4B) to specify what type of data the file contains, as well as which project the data is intended to be used for. Categorizing data files in this way helps to facilitate project analysis. The specific collected files are mainly focused on raw biological data, especially HML files for HLA genotypes, and the previously mentioned antibody csv files to specify identified antibody specificities and corresponding MFI values. To summarize project data, we also collect Project Data Matrix files. A data matrix is a project-specific Excel spreadsheet, which is used to tie together raw data into a

context specific for a project. In the Immunogenic Epitopes project, for example, these are used to specify which HLA genotyping results and antibody assays correspond to patients and donors in a given transplantation.

The script "AnalyzeIhiwUploads.py" was written to summarize and count the uploads within the IHIW database. As of this writing, (Oct 03 2022), there are 9610 uploaded files within the IHIW filesystem. This includes HML, HAML, and Project Data Matrix files, as well as some FASTQ read data and a few cases of family pedigree files. Figure 5 summarizes the files which were submitted for different IHIW Projects, and shows what files are submitted of each type.

3.4 | Data analysis

The AWS environment provides API access to access the uploaded data by authenticated users using encryption keys within the Lambda function web servers. Likewise, the JHipster platform provides permission-restricted API access for authenticated users, which has been modified for IHIW-specific analysis. The combination of file access and DMA access allows detailed reporting and summarization of uploaded data from a variety of methods, most notably in project-specific reports.

We have created scripts, which can be run on an ad-hoc basis by administrators, which perform project-specific analyses for several IHIW projects. The immunogenic epitopes projects provide a good example. We created the script "ImmunogenicEpitopesProjectReport.py" which collects and summarizes the data assigned to this project. When this analysis script is run, it will first .zip together all project files, to provide a dataset to project leaders that can be run in an offline setting. Secondly, the script iterates over all the submitted project data to create summaries. The Project data matrices, which aim to represent transplantations by linking together the raw data, provide pointers to the individual genotyping files and GLstrings. These data are combined together with the antibody csv files to correlate HLA genotype data with de novo HLA-specific antibody formation and facilitate analysis. The results of these scripts are summaries, which are designed to give specifically formatted overviews of the submitted data. The .zip files and project summaries are written back to the S3 buckets and made available through the DMA.

3.5 | Code

The source code is completely open-source and IHIW participants are invited to review the website features and code, and collaborate on improvements of the platform. The software encoding the IHIW database management application (DMA) is provided within IHIW github repositories.[29] The "IHIW_Management" repository (https://github.com/IHIW/IHIW_Management) contains the "front-end" JHipster website code, which specifies the website page design, as well as the management database structure. The "Converters" repository (https://github.com/IHIW/Converters) contains the "back-end" code that specifies the Lambda functions for data converters and validators. With the exception of passwords and encryption keys, the code is freely available for analysis and contribution, and will be available for use and adaptation in future IHIWs.

The IHIW_Management github repository also features a wiki (https://github.com/IHIW/IHIW_Management/wiki), which has been used to create and host some website development documentation. This gives developer-focused instructions on how to set up a development environment, and some specific details on how we would deploy new versions of the website to the AWS environment. It also features, in the "Projects" view, a Kanban board (https://github.com/IHIW/IHIW_Management/projects/1), which we used to track and prioritize new features and bug reports. These features are in place to allow planning of development tasks, but also to encourage new feature suggestions and contribution from the IHIW community.

Some documentation on use of the DMA interface is available on the IHIW website (https://www.ihiw18.org/docs/ihiws-database-user-documentation). This provides more detail to participants on what participants can expect to see, regarding user & laboratory registration, data uploads, and validation feedback.

4 | DISCUSSION

With the aim of effectively managing the heterogeneous data required for a global consortium of multifaceted projects with detailed data requirements, we have presented a resource-efficient data repository and website for collecting and storing data for the 18th and future IHIWs. It was designed to accommodate the data requirements from a variety of IHIW projects, and to provide a platform that can be adapted for future IHIW needs. The use of a cloud-based architecture was critical to circumvent the challenge of analyzing data from participants who may live across the world, where on-site analysis of large datasets from multiple participants by a single laboratory is often not an effective strategy.

4.1 | Architecture

Since the IHIW is an umbrella for various research projects with evolving requirements, there is no single data

structure that universally matches the needs of all project leaders. Furthermore, the use of high-performance computers or computational clusters does not apply in the context of on-demand data uploads and analysis. Therefore, the IHIW architecture is mainly a slim web service, which facilitates flexible data collection by the IHIW projects. The architecture is designed to be powerful, yet cost-efficient with on-demand data processing when it is needed. This does not result in a comprehensive overarching repository including all HLA genotyping results with all supporting metadata for every sample, but by reducing the specific needs and complex requirements of the data collection, the analytic pipeline is flexible and adaptable to individual projects.

We could have mapped all data, especially HML files, to an internal relational database. This option has been discussed, and would provide a valuable and queryable database, that can be repurposed for many studies. This would also facilitate, for example, checking the sample ids of the submitted data to inform the user if duplicate data has already been submitted for a given sample or patient. It would however necessitate even more strict requirements on data, which may discourage participants from uploading their data. To encourage more data upload, the strategy was focused on the validation environment, where there are fewer restrictions on data upload. In this way, we can provide constructive feedback to the users on data quality, but in the case of imperfect or incomplete data, we can still accept the data and perform some analyses.

AWS was selected as a cloud provider, as it is a widely-used and well-supported platform and we had expertise and experience within the IHIW Database team. AWS is, however, one of the many competing cloud providers. Although the software is designed for the AWS environment, most of the applied concepts can be implemented with competing CSP, such as Google Cloud Platform, Microsoft Azure or other competitors. This is consistent with the general IHIW database design, where specific providers and platforms may be substituted or replaced in the future to avoid being permanently linked to specific software or vendors. However, a major portion of the code is specifically written for the AWS environment. Changes in major parts of the architecture, especially to a different CSP, would require non-trivial redesign and rewriting. Although this issue is nearly unavoidable, it is a notable drawback of the IHIW cloud architecture.

4.2 | Data availability and analysis

The strategy for developing tools for analysis and the strategy for making data available has been discussed in depth. There must be a balance between making data available for all participants and restricting access to specific summarized data in specific contexts. Participating laboratories may not be willing to submit any data, no matter how anonymized, if the access and analysis is not strictly defined before submission. This can be a limiting factor in the goals of collecting a re-usable dataset which can be analyzed in customizable ways. This challenge can be addressed moving forward by creating and providing well-defined documentation for how data will be anonymized, summarized, and provided back to participants.

The current DMA website is designed such that individual files are available for download to the submitter and those participants from the same laboratory. The project leader can also download all files for the whole project, and can create and download a combined .zip file to facilitate some offline analysis. However, the main analysis scripts are not fully available within the DMA user interface. The goal of providing flexible analysis tools that can be applied in multiple contexts continues to be a major goal but has not been fully realized in the 18th IHIW. From a practical perspective, creating web-based user interface tools around the data-focused analysis algorithms adds additional layers of complexity, especially if the analysis algorithms must be reliable and usable for several projects. This requires significant work, and the design and usability of the analysis tools can be improved moving forward by involving project leaders and project leaders, and inviting them to contribute time and talent to develop the analysis algorithms. In forthcoming IHIWs, the IHIW database team should be expanded to decentralize the efforts and increase shared value as much as possible.

4.3 | Future IHIWs

A major focus of the IHIW database design and infrastructure is transferability. For such a collaborative and global effort, the task of collecting and organizing data may move for each IHIW. Workshop organizers and administrators play an important role in guiding the data collection, and in order to keep pace with the changing technologies and analysis requirements, the data collection strategy may evolve. The DMA and database architecture were designed keeping in mind that the use cases and projects will change, and future administrators will be able to adapt the existing data structures to, for example, define new participant roles to expand or restrict access to uploaded files. It is critical that the data repository, and collected project data, is not owned by an

individual or single research group. The use of open-source code, and development and using established and novel data standards and rulesets plays an important role in transferability and reusability.

The code for this project is completely open-source. Although the actual submitted data is encrypted within the AWS S3 buckets, the data management and analyses are provided as open-source, to facilitate review and contributions from the community. Workshop participants, especially the organizers for future IHIW events, are encouraged to participate in the use and development of the IHIW database. This could be by submitting improvement requests or bug reports, but direct contribution of analysis techniques would be even better. The described architecture using Lambda Functions for analysis can be seen as a template which can be extended and improved for other purpose-built big-data analysis. Ideas, suggestions, and contributions are always welcome to improve the functionality moving forward, to ensure high-quality future IHIWs.

AUTHOR CONTRIBUTIONS

Benedict M. Matern, Matthias Niemann, Ioannis Nemparis, Andreas Schimanski, and Eric Spierings contributed to development of the IHIW database platform and software. Benedict M. Matern, Matthias Niemann, Emma T. M. Peereboom, Cynthia S. M. Kramer, Sebastiaan Heidt and Eric Spierings developed the data analysis and interpretation strategy. Sebastiaan Heidt and Eric Spierings organized and hosted the 18th IHIW conference and established the IHIW database team. Benedict M. Matern and Matthias Niemann drafted the manuscript, and all authors contributed to manuscript revision.

ACKNOWLEDGMENTS

Thanks to all participating laboratories who have submitted the excellent data for analysis, and to all participants of the 18th IHIW. Thanks to participants of the Data Standards Hackathon (DASH) meetings, especially Martin Maiers, Bob Milius, Loren Gragert, Steve Mack, and Kazutoyo Osoegawa who have contributed ideas, effort, and data standards which are essential in IHIW efforts. Thanks also to Anat Tambur and Lloyd D'Orsogna for insightful discussions on epitopes analysis.

FUNDING INFORMATION

This work was supported by a grant of the International HLA & Immunogenetics Workshop Foundation and by an internal grant from the UMC Utrecht.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

The data repository and analysis software are available at https://github.com/IHIW, and workshop data is available to workshop participants at https://data.ihiws.org/ or by reasonable request to IHIW organizers.

ORCID

Benedict M. Matern https://orcid.org/0000-0002-9548-732X
Matthias Niemann https://orcid.org/0000-0002-3514-9812
Emma T. M. Peereboom https://orcid.org/0000-0002-9656-8076
Cynthia S. M. Kramer https://orcid.org/0000-0003-1350-2336
Sebastiaan Heidt https://orcid.org/0000-0002-6700-188X
Eric Spierings https://orcid.org/0000-0001-9441-1019

REFERENCES

1. *Histocompatibility Testing*. The National Academies Press; 1965. 10.17226/21294
2. Holdsworth R, Hurley CK, Marsh SG, et al. The HLA dictionary 2008: a summary of HLA-A, -B, -C, -DRB1/3/4/5, and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Tissue Antigens*. 2009; 73(2):95-170. doi:10.1111/j.1399-0039.2008.01183.x
3. Osoegawa K, Marsh SGE, Holdsworth R, et al. A new strategy for systematically classifying HLA alleles into serological specificities. *HLA*. 2022;100(3):193-231. doi:10.1111/tan.14662
4. Kramer CSM, Koster J, Haasnoot GW, Roelen DL, Claas FHJ, Heidt S. HLA-EMMA: a user-friendly tool to analyse HLA class I and class II compatibility on the amino acid level. *HLA*. 2020; 96(1):43-51. doi:10.1111/tan.13883
5. Niemann M, Matern BM, Spierings E. Snowflake: a deep learning-based human leukocyte antigen matching algorithm considering allele-specific surface accessibility. *Front Immunol*. 2022;13:937587. doi:10.3389/fimmu.2022.937587
6. Geneugelijk K, Spierings E. PIRCHE-II: an algorithm to predict indirectly recognizable HLA epitopes in solid organ transplantation. *Immunogenetics*. 2020;72(1-2):119-129. doi:10.1007/s00251-019-01140-x
7. Matern BM, Olieslagers TI, Voorter CEM, Groeneweg M, Tilanus MGJ. Insights into the polymorphism in HLA-DRA and its evolutionary relationship with HLA haplotypes. *HLA*. 2020;95(2):117-127. doi:10.1111/tan.13730
8. Askar M, Madbouly A, Zhrebker L, et al. HLA haplotypes In 250 families: the Baylor Laboratory results and a perspective on a core NGS testing model for the 17th international HLA and immunogenetics workshop. *Hum Immunol*. 2019;80(11): 897-905. doi:10.1016/j.humimm.2019.07.298
9. Osoegawa K, Mallempati KC, Gangavarapu S, et al. HLA alleles and haplotypes observed in 263 US families. *Hum Immunol*. 2019;80(9):644-660. doi:10.1016/j.humimm.2019.05.018
10. Hurley CK, Kempenich J, Wadsworth K, et al. Common, intermediate and well-documented HLA alleles in world

populations: CIWD version 3.0.0. *HLA*. 2020;95(6):516-531. doi:10.1111/tan.13811
11. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol*. 2013;74(10):1313-1320. doi:10.1016/j.humimm.2013.06.025
12. Gonzalez-Galarza FF, McCabe A, Santos EJMD, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res*. 2020;48(D1):D783-D788. doi:10.1093/nar/gkz1029
13. Balgansuren G, Regen L, Sprague M, Shelton N, Petersdorf E, Hansen JA. Identification of the rs9277534 HLA-DP expression marker by next generation sequencing for the selection of unrelated donors for hematopoietic cell transplantation. *Hum Immunol*. 2019;80(10):828-833. doi:10.1016/j.humimm.2019.05.010
14. Schone B, Bergmann S, Lang K, et al. Predicting an HLA-DPB1 expression marker based on standard DPB1 genotyping: linkage analysis of over 32,000 samples. *Hum Immunol*. 2018;79(1):20-27. doi:10.1016/j.humimm.2017.11.001
15. Turner TR, Hayward DR, Gymer AW, et al. Widespread non-coding polymorphism in HLA class II genes of international HLA and immunogenetics workshop cell lines. *HLA*. 2022;99(4):328-356. doi:10.1111/tan.14571
16. Truong L, Matern BM, Groeneweg M, et al. Polymorphism clustering of the 21.5 kb DPA-promoter-DPB region reveals novel extended full-length haplotypes. *HLA*. 2020;96(3):299-311. doi:10.1111/tan.13975
17. Arrieta-Bolaños E, Crivello P, Shaw BE, et al. In silico prediction of nonpermissive HLA-DPB1 mismatches in unrelated HCT by functional distance. *Blood Adv*. 2018;2(14):1773-1783. doi:10.1182/bloodadvances.2018019620
18. Little AM. HLA antibodies in haematopoietic stem cell transplantation. *HLA*. 2019;94(Suppl 2):21-24. doi:10.1111/tan.13741
19. Everly MJ, Rebellato LM, Haisch CE, et al. Incidence and impact of *de novo* donor-specific alloantibody in primary renal allografts. *Transplantation*. 2013;95(3):410-417. doi:10.1097/TP.0b013e31827d62e3
20. Heidt S, Claas FHJ. Not all HLA epitope mismatches are equal. *Kidney Int*. 2020;97(4):653-655. doi:10.1016/j.kint.2019.12.017
21. Milius RP, Heuer M, Valiga D, et al. Histoimmunogenetics markup language 1.0: reporting next generation sequencing-based HLA and KIR genotyping. *Hum Immunol*. 2015;76(12):963-974. doi:10.1016/j.humimm.2015.08.001
22. Milius RP, Mack SJ, Hollenbach JA, et al. Genotype list string: a grammar for describing HLA and KIR genotyping results in a text string. *Tissue Antigens*. 2013;82(2):106-112. doi:10.1111/tan.12150
23. Mack SJ, Milius RP, Gifford BD, et al. Minimum information for reporting next generation sequence genotyping (MIRING): guidelines for reporting HLA and KIR genotyping via next generation sequencing. *Hum Immunol*. 2015;76(12):954-962. doi:10.1016/j.humimm.2015.09.011
24. Kramer CSM, Israeli M, Mulder A, et al. The long and winding road towards epitope matching in clinical transplantation. *Transpl Int Off J Eur Soc Organ Transplant*. 2019;32(1):16-24. doi:10.1111/tri.13362
25. Voorter CEM, Matern B, Tran TH, et al. Full-length extension of HLA allele sequences by HLA allele-specific hemizygous sanger sequencing (SSBT). *Hum Immunol*. 2018;79(11):763-772. doi:10.1016/j.humimm.2018.08.004
26. Augusto DG, Hollenbach JA. HLA variation and antigen presentation in COVID-19 and SARS-CoV-2 infection. *Curr Opin Immunol*. 2022;76:102178. doi:10.1016/j.coi.2022.102178
27. Chang CJ, Osoegawa K, Milius RP, et al. Collection and storage of HLA NGS genotyping data for the 17th international HLA and immunogenetics workshop. *Hum Immunol*. 2018;79(2):77-86. doi:10.1016/j.humimm.2017.12.004
28. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163
29. International HLA & Immunogenetics Workshop. GitHub. Accessed September 30, 2022. https://github.com/IHIW

How to cite this article: Matern BM, Niemann M, Nemparis I, et al. Using cloud infrastructure to facilitate data collection and conversion of HLA diagnostic data for the 18th International HLA and Immunogenetics Workshop. *HLA*. 2023;101(5):484-495. doi:10.1111/tan.14989