

UC San Diego

UC San Diego Previously Published Works

Title

VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements

Permalink

<https://escholarship.org/uc/item/3kj329p4>

Authors

Christley, Scott
Scarborough, Walter
Salinas, Eddie
et al.

Publication Date

2018

DOI

10.3389/fimmu.2018.00976

Peer reviewed



VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements

Scott Christley¹, Walter Scarborough², Eddie Salinas¹, William H. Rounds¹, Inimary T. Toby¹, John M. Fonner², Mikhail K. Levin³, Min Kim¹, Stephen A. Mock², Christopher Jordan², Jared Ostmeier¹, Adam Buntzman⁴, Florian Rubelt⁵, Marco L. Davila⁶, Nancy L. Monson^{7,8}, Richard H. Scheuermann^{9,10,11} and Lindsay G. Cowell^{1*}

¹Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, United States, ²Texas Advanced Computing Center, University of Texas at Austin, Austin, TX, United States, ³Bank of America Corporate Center, Charlotte, NC, United States, ⁴Bio5 Institute, University of Arizona, Tucson, AZ, United States, ⁵Department of Microbiology and Immunology, Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA, United States, ⁶H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States, ⁷Department of Neurology and Neurotherapeutics, University of Texas Southwestern Medical Center, Dallas, TX, United States, ⁸Department of Immunology, University of Texas Southwestern Medical Center, Dallas, TX, United States, ⁹J. Craig Venter Institute, La Jolla, CA, United States, ¹⁰Department of Pathology, University of California, San Diego, San Diego, CA, United States, ¹¹La Jolla Institute for Allergy & Immunology, La Jolla, CA, United States

OPEN ACCESS

Edited by:

Deborah K. Dunn-Walters,
University of Surrey,
United Kingdom

Reviewed by:

Johannes Trück,
University Children's Hospital
Zurich, Switzerland
Michael Zemlin,
Universitätsklinikum
des Saarlandes, Germany
Nina Luning Prak,
University of Pennsylvania,
United States

*Correspondence:

Lindsay G. Cowell
lindsay.cowell@utsouthwestern.edu

Specialty section:

This article was submitted to
B Cell Biology, a
section of the journal
Frontiers in Immunology

Received: 18 January 2018

Accepted: 19 April 2018

Published: 08 May 2018

Citation:

Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM, Levin MK, Kim M, Mock SA, Jordan C, Ostmeier J, Buntzman A, Rubelt F, Davila ML, Monson NL, Scheuermann RH and Cowell LG (2018) VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements. *Front. Immunol.* 9:976. doi: 10.3389/fimmu.2018.00976

Background: Recent technological advances in immune repertoire sequencing have created tremendous potential for advancing our understanding of adaptive immune response dynamics in various states of health and disease. Immune repertoire sequencing produces large, highly complex data sets, however, which require specialized methods and software tools for their effective analysis and interpretation.

Results: VDJServer is a cloud-based analysis portal for immune repertoire sequence data that provide access to a suite of tools for a complete analysis workflow, including modules for preprocessing and quality control of sequence reads, V(D)J gene segment assignment, repertoire characterization, and repertoire comparison. VDJServer also provides sophisticated visualizations for exploratory analysis. It is accessible through a standard web browser via a graphical user interface designed for use by immunologists, clinicians, and bioinformatics researchers. VDJServer provides a data commons for public sharing of repertoire sequencing data, as well as private sharing of data between users. We describe the main functionality and architecture of VDJServer and demonstrate its capabilities with use cases from cancer immunology and autoimmunity.

Conclusion: VDJServer provides a complete analysis suite for human and mouse T-cell and B-cell receptor repertoire sequencing data. The combination of its user-friendly interface and high-performance computing allows large immune repertoire sequencing projects to be analyzed with no programming or software installation required. VDJServer is a web-accessible cloud platform that provides access through a graphical user interface to a data management infrastructure, a collection of analysis tools covering all steps in an analysis, and an infrastructure for sharing data along with workflows, results, and computational provenance. VDJServer is a free, publicly available, and open-source licensed resource.

Keywords: bioinformatics, cloud computing, Rep-seq, immune repertoire, B-cell receptor, T cell receptor

BACKGROUND

The adaptive immune system is composed of specialized cells, molecules, and processes that evolved to defend the organism against foreign pathogens and tumorous cells. In jawed vertebrates, the primary actors in adaptive immunity are B and T lymphocytes, which express immune receptors on their surface, and, in the case of B lymphocytes, secrete antibodies, a soluble form of the receptor. The genes encoding immune receptors are somatically generated through a DNA recombination process, V(D)J recombination, that assembles variable (V), diversity (D), and joining (J) gene segments into mature, composite genes (1). In some species, including mice and humans, the rearranged genes in B lymphocytes are further diversified through somatic hypermutation (SHM) (2). As a result of these processes, each individual has millions of unique immune receptor genes (3, 4), although some lymphocytes have identical genes due to clonal expansion. The full collection of functional immune receptor gene sequences in an individual at a single point in time is referred to as the adaptive immune receptor repertoire (AIRR). Somatic generation of a tremendously diverse repertoire enables effective immune responses against an essentially infinite array of antigens, such as those derived from pathogens or tumors. Components of this somatically generated repertoire can also recognize self-antigens, however, leading to autoimmune responses.

The composition of immune repertoires shifts in response to immunological events (5). Thus, immune repertoires reflect the history and current state of adaptive immune responses, and analysis of repertoire composition is critical in both basic and translational research, clinical diagnostics, and in pharmaceutical development. Recent technological advances in immune repertoire sequencing have created tremendous potential for it to advance our understanding of adaptive immune response dynamics and lead to the development of repertoire-based diagnostic and prognostic assays. This has resulted in an explosion of research activity, a trend expected to continue (Figure 1). Recent examples of the power of repertoire analysis to have significant impact across various research and clinical areas include: understanding the development of a healthy immune system and immune senescence (6, 7); determining the nature of successful and unsuccessful immune responses for vaccine design (8–13); identifying the targets of autoimmune responses (14–18); diagnosing and monitoring hematologic malignancies (19–23); predicting clinical outcomes in cancer patients (24–30); and monitoring patients for graft-versus-host disease (31) or for organ rejection (32–34) after transplantation.

Immune repertoire sequencing produces large, highly complex data sets that require specialized analysis methods and software tools. Since the first studies demonstrating the technology were published (35–39), a research community has arisen around the development of new methods and tools (40, 41). As part of that effort, we developed VDJServer to address critical barriers in broader adoption of immune repertoire sequencing, namely, the lack of a complete, start-to-finish analysis pipeline, the lack of a data management infrastructure, and limited access for many researchers to high-performance computing (HPC) resources. VDJServer fills these gaps, specifically providing (1) an open suite

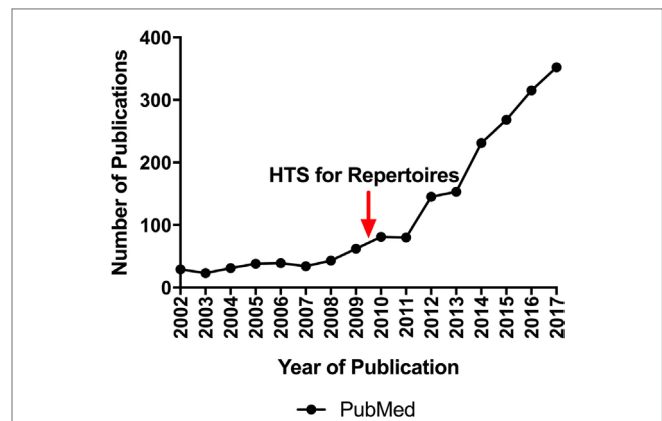


FIGURE 1 | The number of newly published repertoire sequencing papers appearing in PubMed each year over the last 15 years. For 2002, the number was obtained by the query ((repertoire sequencing) AND (2002 [dp])). The number for 2017 is a projection based on the number appearing between January and December 2017. The red arrow indicates the year in which publications demonstrated the feasibility of applying high-throughput sequencing to immune repertoires.

of interoperable repertoire analysis tools that allows users to upload a set of sequences and pass them through a seamless workflow that executes all steps in an analysis, (2) access to sophisticated analysis tools running in an HPC environment, (3) interactive visualization capabilities for exploratory analysis, (4) a data management infrastructure, and (5) a graphical user interface to facilitate use by experimental and clinical research groups that lack bioinformatics expertise. VDJServer provides these capabilities as a free and publicly available resource.

While most immune repertoire analysis tools must be locally installed, there are a few tools that can be accessed over the web. The National Center for Biotechnology Information provides web access to the IgBlast tool for germline gene assignment (42). The Immunogenetics Information System (IMGT) provides access to tools for germline gene assignment (IMGT/V-QUEST) and junction analysis (IMGT/JunctionAnalysis) (43–47). The Vidjil web application (48) provides access to the Vidjil clonotype clustering algorithm, as well as to paired-end read merging *via* PEAR (49), and to germline gene alignment and clone identification using MiXCR (50). The most complete web-based analysis pipeline is provided by IGGalaxy (51) and its successor ARGalaxy (52). These are Galaxy-based (53) pipelines that provide access to demultiplexing and read trimming for 454 data, to downstream analysis tools, such as Change-O (54) and BASELINE (55), and to visualization of the output of those tools. ImmuneDB (56), which must be installed locally, provides a web-based interface to explore results from its analysis pipeline, which includes pre-processing with pRESTO (57), gene and clonal assignment (58), lineage tree construction, and mutation analysis with BASELINE (55). All of these web-based tools are limited in some fashion, however, either by restricting the number of sequences accepted by the web application, providing only a single tool suite, or not providing the tools necessary for all steps in a complete analysis workflow. Furthermore, none of these tools provide an HPC implementation to handle large immune repertoire studies,

they lack metadata capabilities with user-defined sample groups and associated repertoire comparative analysis between groups, and they do not capture the necessary provenance information to allow for reproducibility of the analysis by others (59, 60). Among all currently available tools, VDJServer is the only web-accessible cloud platform that provides access through a graphical user interface to a data management infrastructure, a collection of HPC-enabled analysis tools covering all steps in an analysis, and an infrastructure for sharing data along with workflows, results, and computational provenance.

IMPLEMENTATION

Cloud-Based Architecture Overview

The VDJServer analysis portal is comprised of two main components: a web browser user interface and a web API. VDJServer's architecture is designed upon the Agave Science-As-A-Service cloud platform (61) and augmented with a VDJServer-specific API. Generally, science gateways need to implement a database resource within their architecture for data management. However, the use of Agave allows VDJServer to offload database implementation into the cloud platform. This simplifies VDJServer's architecture and provides the many benefits of cloud computing, such as lower maintenance costs, quick and flexible deployment, and dynamic scaling to accommodate user load. Agave Science APIs are a collection of RESTful web services with user identity management, file management, systems management, application deployment, metadata database, events/notifications, and job execution as some of their main functionality. VDJServer provides an additional RESTful API (Table 1) for project management, Agave event/notification processing, metadata capture for files and jobs, user profile and feedback management, community data publishing, and error logging. The API is implemented as a JavaScript Node.js application using the Express framework, and Nginx is the web server acting as HTTP/HTTPS proxy and serving the user interface code to client browsers.

Web Browser Interface

VDJServer's user interface is a JavaScript single-page application with Backbone.js for the model-view event-driven application framework, Bootstrap for CSS web layout, and Handlebars for HTML templates. Data transport uses JSON with jQuery for some HTTP requests not handled by Backbone.js and Websockets for server-side events, such as file upload and job notifications. Chart and graph visualizations use D3.js, NVD3, and Highcharts. RequireJS is used for file and module loading optimization, and Grunt is used for the build system.

The interface has four primary views: project data management, metadata entry, configuration and execution of analysis jobs, and visualization of results. Each of these primary views is described in more detail in the following sections.

Project Data Management

VDJServer is project-based where a project typically corresponds to a single experimental study. Each project is a logical container for files, jobs, analysis results, and visualizations, and

TABLE 1 | VDJServer release 1.0 API.

Endpoint	Method	Description
/	GET	Status of API service
/feedback	POST	User feedback
/feedback/public	POST	Public feedback
/jobs/queue/pending	GET	Pending jobs for project
/jobs/queue	POST	Submit job
/jobs/archive/:id	POST	Archive job
/jobs/unarchive/:id	POST	Unarchive job
/notifications/files/import	POST	Agave notifications for file import
/notifications/jobs/:id	POST	Agave notifications for job
/permissions/metadata	POST	Update user permissions for metadata
/permissions/username	POST	Add user permissions on project data
/permissions/username	DELETE	Remove user permissions from project data
/projects	POST	Create project
/projects/:id/metadata/export	GET	Export metadata from project into tab-separated values file
/projects/:id/metadata/import	POST	Import metadata into project
/public	GET	Query public data
/telemetry	POST	Error logging
/token	POST	Request an Agave authentication token
/token	PUT	Refresh Agave authentication token
/user	POST	Create user account
/user/change-password	POST	User change password
/user/reset-password	POST	Initiate password reset
/user/reset-password/verify	POST	Verify password reset
/user/:username/verify/email	POST	Send user verification email
/user/verify/:id	POST	Verify user

any number of projects may be created. Figure 2 shows the interface for an example project. Multiple users can be given permissions on a project allowing them to run analysis jobs, access data, and visualize results that are shared with the other users. There is no limit to the number of files, jobs, or users that can be associated with a project. Data files can be uploaded from the user's computer, from the user's Dropbox account, or from a URL (HTTP and FTP supported protocols). VDJServer supports sequencing data in a number of file formats including single-end and paired-end reads in FASTQ format, single-end reads and quality scores in separate FASTA and QUAL files, and sequence data without quality scores in FASTA format. They can be compressed as zip, gzip, or bzip2 for faster upload. Each data file can be tagged with a pre-defined semantic type and with a set of user-defined tags. The semantic type allows VDJServer to automatically infer the appropriate matching between files and tool inputs for analysis jobs and to populate user interface elements with appropriate options. Both help to prevent analysis and job errors. The file search interface allows users to query all files within a project using a user-specified search string that searches against filenames, semantic type, and user-defined tags. Files are the underlying basis for input and output data for tools, and the resulting explosion in the number, type, and size of files can easily overwhelm users. VDJServer encapsulates this complexity and streamlines processing by displaying abbreviated summaries of job output, which is broadly grouped as output data (which may be input to another tool), visualization data, or log/configuration information. Context-specific operations appropriate for each output file type are provided, e.g., figures and charts are shown

VDJ SERVER

+ ADD PROJECT

Molecular diagnostic test for multiple sclerosis

Project Settings

Upload and Browse Project Data

Metadata Entry

Link .fasta/.qual Files

Link Paired Read Files

View Analyses and Results

Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes

Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection

Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue

Antibody repertoire RNA-seq throughout key stages of

COMMUNITY DATA | **DOCUMENTATION** | **FEEDBACK** | | scott_ab

Name: Molecular diagnostic test for multiple sclerosis
 VDJServer UUID: 3011881897146323431-242ac11c-0001-012
 Data: 12 files
 Members: 1

Upload | Run Job | File Actions

name:exampleSearchFile.fastq tag:exampleSearchTag SEARCH

	Name	Last Modified	Size	File Origin	Type	Tags	Read Direction
<input type="checkbox"/>	Sample00002.fna	24-Oct-2017 11:56 am	3.55 MB	Uploaded File	Single-End Read-Level Data		R
<input type="checkbox"/>	Sample00002.qual	24-Oct-2017 11:16 am	9.21 MB	Uploaded File			
<input type="checkbox"/>	Sample00003.fna	24-Oct-2017 11:56 am	13.74 MB	Uploaded File	Single-End Read-Level Data		R
<input type="checkbox"/>	Sample00003.qual	24-Oct-2017 11:16 am	35.97 MB	Uploaded File			
<input type="checkbox"/>	Sample00005.fna	24-Oct-2017 11:56 am	8.73 MB	Uploaded File	Single-End Read-Level Data		R
<input type="checkbox"/>	Sample00005.qual	24-Oct-2017 11:16 am	22.72 MB	Uploaded File			
<input type="checkbox"/>	Sample00001.fna	24-Oct-2017 11:56 am	17.91 MB	Uploaded File	Single-End Read-Level Data		R
<input type="checkbox"/>	Sample00001.qual	24-Oct-2017 11:16 am	46.59 MB	Uploaded File			

Output Files for Job: VDJPipe pre-processing

<input type="checkbox"/>	Unique Post-Filter Sequences (Sample00001)	24-Oct-2017 12:02 pm	1.53 MB	Job File	Read-Level Data		
<input type="checkbox"/>	Unique Post-Filter Sequences (Sample00002)	24-Oct-2017 12:02 pm	857.05 kB	Job File	Read-Level Data		
<input type="checkbox"/>	Unique Post-Filter Sequences (Sample00005)	24-Oct-2017 12:02 pm	1.52 MB	Job File	Read-Level Data		
<input type="checkbox"/>	Unique Post-Filter Sequences (Sample00003)	24-Oct-2017 12:01 pm	869.45 kB	Job File	Read-Level Data		

for visualization data. Regardless, the underlying files are always available to the user to download if desired.

Study Metadata

Structured metadata are increasingly important to insure adherence to funder and journal data sharing policies for data reusability and study reproducibility. VDJServer provides a comprehensive metadata entry and management interface for an immune repertoire study and its subjects, samples, and biomaterial processing. VDJServer collects metadata according to the recently published Minimal Information about Adaptive Immune Receptor Repertoires (MiAIRR) standards developed by a working group of the AIRR Community¹ (40, 41), as well as any number of user-defined fields. Custom groups of samples and subjects, e.g., Control and Treatment groups, can be defined by the user as part of the metadata. VDJServer uses metadata for performing immune repertoire calculations and comparisons between samples and groups. Metadata can be provided through manual entry, uploading of a spreadsheet table, or a combination

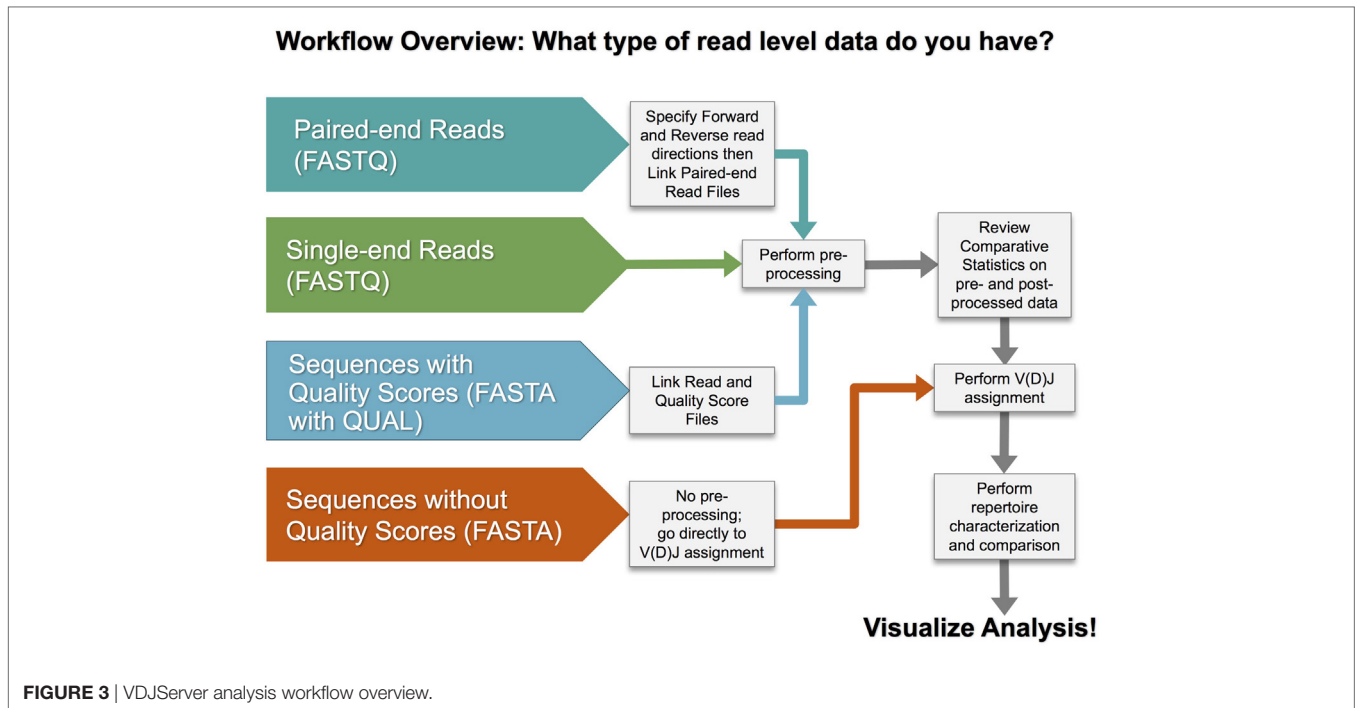
of the two. Metadata can be exported into Tab-Separated Values (TSV) files for easy import into external tools, such as Microsoft Excel. VDJServer automatically captures provenance metadata for all analysis workflows executed by the users, which is described in more detail in Section “Findable, Accessible, Interoperable, and Reusable (FAIR) Principles and Reproducibility.”

Analysis Workflows

There are multiple steps in a typical analysis workflow for repertoire sequencing data, as illustrated in **Figure 3**. The four basic steps include (1) preprocessing of sequence reads, (2) V(D)J gene segment assignment, (3) rearrangement annotation, and (4) repertoire characterization and comparison. VDJServer provides tools for all steps consisting of VDJPipe (62) and pRESTO (57) for step 1, IgBlast (42) for step 2, RepSum and Change-O (54) for step 3, and RepCalc, Change-O (54), Alakazam (54), and SHazaM (54) for step 4. Not all steps are required, and specific workflows will vary based upon the input sequencing data and desired analysis.

Preprocessing of sequence reads is a common step for next-generation sequencing data that remove low quality reads and

¹<http://www.airr-community.org> (Accessed: April 27, 2018).



prepares the reads for further analysis. VDJServer provides preprocessing tools designed specifically for immune repertoire sequencing data, since preprocessing of these data has unique characteristics, such as 5' and 3' PCR primer targeting, complex multilevel barcode demultiplexing, and duplicate sequence read collapsing, when compared with other next-generation sequencing data.^{2,3} Preprocessing tasks may include quality filtering, homopolymer filtering, length and nucleotide filtering, merging of paired-end reads, barcode demultiplexing, forward and reverse primer matching, and duplicate reads collapsing. Users can choose either or both VDJPipe and pRESTO for preprocessing as they are similar in capability but each with some unique features (e.g., pRESTO supports preprocessing of reads containing unique molecular identifiers). VDJServer calculates base composition statistics and read quality statistics before and after preprocessing and provides comparative visualization for user assessment, as described in Section “Visualizations.”

VDJServer combines the V(D)J gene segment assignment and rearrangement annotation steps together in a single job execution, and there is no restriction on the number of sequences that may be submitted. IgBlast is used for V(D)J gene segment assignment. The germline database used by VDJServer contains T cell and B cell gene sequences for human and mouse based upon the IMGT database (63). Multiple annotation outputs are provided by VDJServer including: a VDJML (64) file of the IgBlast output alignments, a TSV file provided by RepSum, a TSV file provided by Change-O, and an AIRR rearrangements file.

²FastQC. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (Accessed: April 27, 2018).

³ea-utils. Available from: <https://github.com/ExpressionAnalysis/ea-utils> (Accessed: April 27, 2018).

The TSV files contain annotations, such as gene calls and CDR3 sequences. The AIRR rearrangements file is an evolving community standard to facilitate interoperability between immune repertoire analysis tools,⁴ and it specifies mandatory and optional annotations. Mandatory annotations are all essential data fields from V(D)J gene segment assignment, such as the V, D, and J gene calls, functional versus non-functional rearrangements, CIGAR strings for sequence alignment, and sequences and their identifiers. Optional annotations are those which can be computed from the mandatory annotations or are specific to some V(D)J gene segment assignment tools.

After the initial processing steps to assemble individual VDJ rearrangements, repertoire characterization and comparison becomes the primary focus of the researcher to extract insights and test hypotheses with their data. VDJServer provides a wide range of functionality applicable to both T-cell and B-cell data. Repertoire characterization encompasses various factors for each repertoire sample, including (1) enumeration of V, D, and J gene segment usage, and V–J and V–D–J combinations; (2) identification of clonally related sequences; (3) estimation of repertoire diversity utilizing common measures of diversity; (4) characterization of CDR3 patterns, such as length, amino acid utilization, and physicochemical properties; and (5) enumeration of unique CDR3 sequences and unique V gene, J gene, and CDR3 sequence combinations. VDJServer also provides B-cell-specific functionality for analysis of SHM: (1) characterization of somatic mutation patterns, including identification of mutations, determination of replacement (R) or silent (S) mutations, calculation of V gene, CDR, and framework mutation frequencies and R:S ratios; (2) characterization of patterns

⁴AIRR Rearrangements File Format. Available from: <http://docs.airr-community.org> (Accessed: April 27, 2018).

of selection across framework and CDR regions; and (3) inference of B cell lineage trees. VDJServer provides comparison of many of these characteristics between repertoire samples and repertoire groups. As described earlier in Section “Study Metadata,” users can define their own groupings of repertoire samples, which will be pairwise compared and provides increased flexibility for *ad hoc* analysis. How group comparisons are performed depends upon the nature of the characteristic. For numerical values, such as gene segment usage, mean and variance are calculated for the set of repertoires that comprise the group. Not all characteristics, such as a diversity curve, have a well-established aggregation metric and thus do not have a meaningful group comparison. While other characteristics enable additional analyses, such as shared CDR3 sequences, with intragroup comparison quantifying sharing between repertoires within the same group and intergroup comparison quantifying sharing between two groups. Results from repertoire characterization and comparison can be visually examined through a set of charts and figures, as described in the next section. Furthermore, all of the results are stored in TSV files that can be downloaded for import into external tools.

Visualizations

VDJServer provides two primary sets of visualizations. One set of charts for assessing quality and composition statistics before and after preprocessing of sequence reads, and another set of charts for examining repertoire characteristics and comparative repertoire analysis. All charts provide interactive settings that allow the user to manipulate the chart and enable/disable which results are displayed, while tooltips (a small visual popup box) provide additional detail for specific data points when the mouse pointer hovers over them. In addition, the currently displayed chart can be quickly downloaded as a high-resolution image with the click of a button.

The preprocessing visualizations include histograms of read length, average quality score, and GC content; a box-and-whiskers display of the quality score distribution at each read position; and base composition at each read position for all four bases and ambiguous base calls. The quality score distribution has the ability to zoom in on portions of sequence read. Statistics are calculated on the pre- and postprocessed sequence reads, and both are provided on a single figure for easy comparison. Examples of these charts are shown in **Figure 4**.

For repertoire characterization and comparison, VDJServer provides visualizations for CDR3 length, gene segment usage, clonal composition, diversity, mutational distribution by position, and quantification of selection for framework and CDR regions. For gene segment usage, users can interactively drill down through the gene hierarchy and display counts at the locus (e.g., IGH or TRB), gene family, gene segment, and allele levels. Moreover, gene segment usage can display absolute counts, which is useful for magnitude analysis within a repertoire, and relative counts which is useful for comparison across repertoires. Clonal composition is visualized as a ranking of individual clones sorted by their relative abundance, or alternatively as a cumulative abundance curve of those sorted clones. Analysis results are displayed on each chart by selecting specific input files, or preferably by selecting repertoire samples and groups based upon the

study metadata. Individual repertoire samples can be displayed together with repertoire groups on the same chart, with group results shown as an average with error bars or as a set of identically colored data elements. Examples of these charts are shown in **Figure 5**.

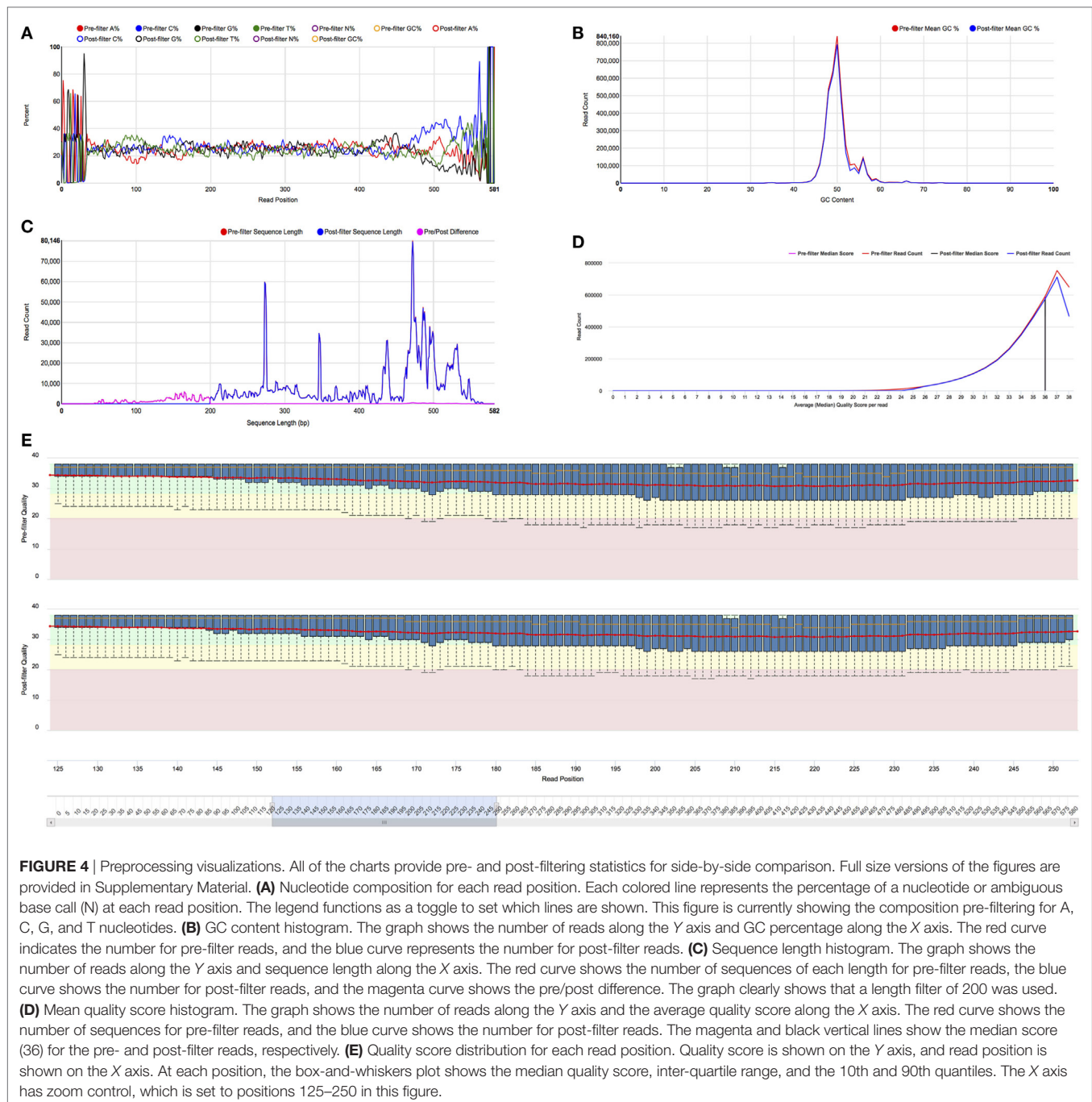
Findable, Accessible, Interoperable, and Reusable (FAIR) Principles and Reproducibility

Computational analysis is becoming increasingly complex with multiple tools and steps involved in a typical workflow, and the exploratory nature of research often entails running multiple workflows to test alternative hypotheses and perform comparative analysis. The FAIR principles places emphasis on enhancing the ability of machines to find and use data, as well as the processes, tools, and workflows that led to that data (65). VDJServer supports these principles through its implementation with digital object identifiers (DOIs), an open and free API, standard formats, such as TSV and JSON, and detailed provenance. A DOI is a code to permanently and stably identify digital objects along with a standard mechanism to retrieve metadata about the object as well as access the digital object itself. Such digital objects in VDJServer include projects, files, study metadata, and jobs, all of which have a uuid (universally unique identifier) assigned to them. The Agave API, in combination with VDJServer’s API, allows digital objects to be accessed directly with their uuid or to be queried based upon the object’s metadata. The API URL with the object’s uuid provides a permanent and stable identification. Access to private data requires authentication while public data does not.

Replication of scientific results is vitally important, yet the burden of specifying the exact details of a computational analysis can easily lead to mistakes or missing information. VDJServer eliminates this burden by automatically capturing provenance for all jobs in a machine-readable JSON description of the inputs, parameters, and outputs of all computational processes. This provenance provides a trail for data as it is produced and passed from one computational tool to the next in an analysis workflow. Like other digital objects, computational provenance is given a uuid for access and can be queried. Furthermore, VDJServer can utilize the computational provenance to produce a detailed description of the complete analysis workflow for submission to a journal or database. Users are no longer burdened with remembering or keeping careful notes about exactly which tool was run to generate a particular set of results as VDJServer keeps track for them.

High-Performance Computing

VDJServer provides free access to HPC at the Texas Advanced Computing Center (TACC). VDJServer automatically parallelizes tool execution based on the size of the input data. Computational analysis that takes days on a user’s desktop might take only a few hours on VDJServer. Small jobs are run on dedicated VDJServer machines and begin execution immediately, while larger jobs are queued to run on one of the TACC HPC clusters. Job submission to the HPC clusters is scheduled with optimized run time so that



smaller jobs execute quickly, even if the queue is busy. At the time of this writing, VDJServer has processed billions of sequences through its analysis pipelines using TACC's HPC resources.

RESULTS AND DISCUSSION

Immune repertoire analysis is having a significant impact across various research and clinical areas. VDJServer's combination of an easy-to-use web interface, full suite of tools, interactive visualization, and HPC integration, facilitates a rapid, exploratory analysis cycle for scientific discovery. In the following sections,

we describe two scientific use cases of VDJServer and the discoveries it has enabled.

Use Case: Multiple Sclerosis

We used VDJServer to analyze the IGH repertoires of patients with relapsing–remitting multiple sclerosis (RRMS) in search of diagnostic immune biomarkers. The first study examined SHM patterns in the V gene segment (14), while the second study developed a statistical classifier based upon the CDR3 sequence (66). For both studies, B cells were obtained from the cerebrospinal fluid (CSF) of patients diagnosed with either RRMS or other

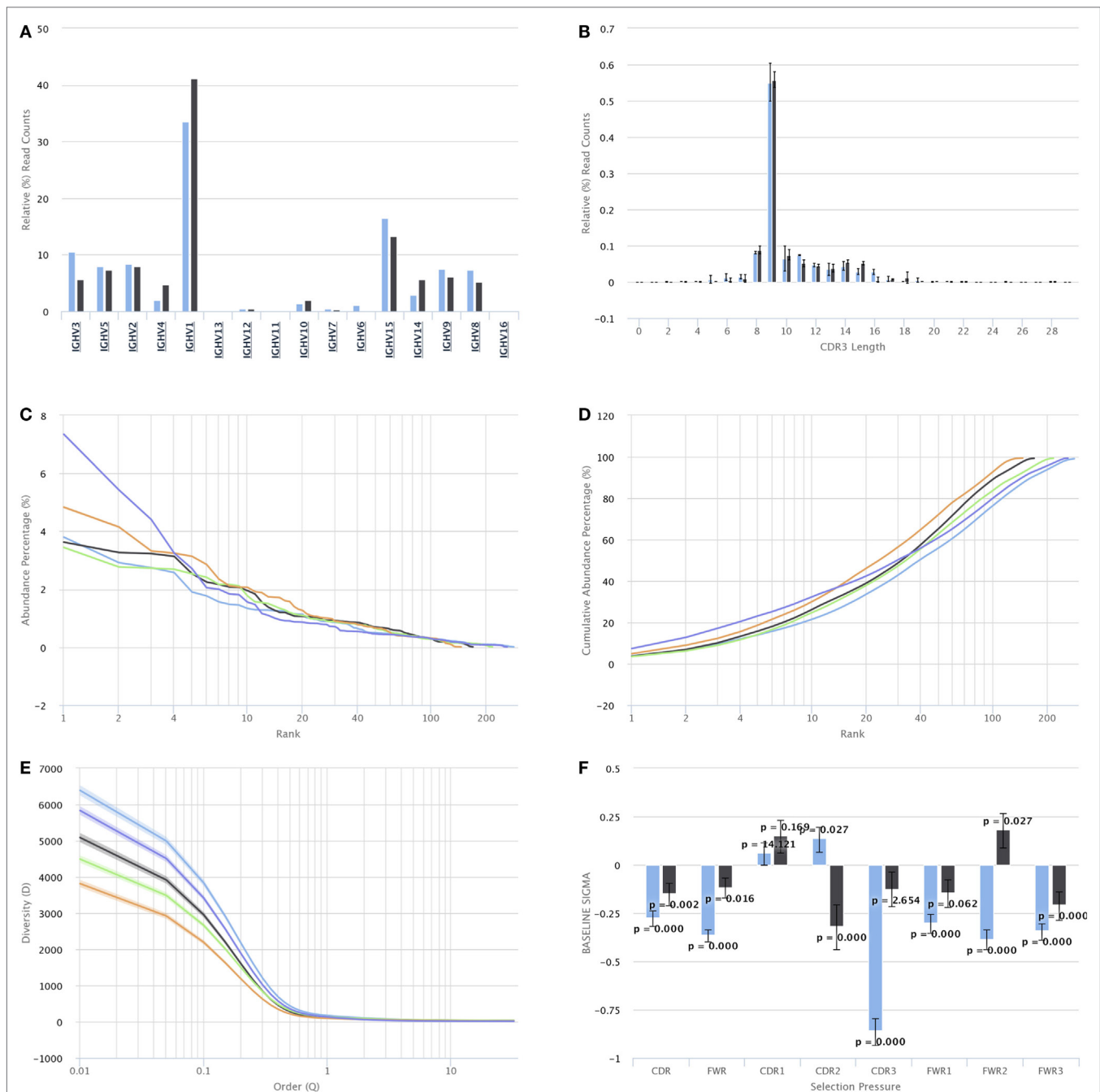


FIGURE 5 | Repertoire characterization and comparison visualizations. For simplicity, the interface buttons for selecting which samples and sample groups to display are not shown. Full size versions of the figures are provided in Supplementary Material. **(A)** Gene segment usage histogram. The graph shows IGH V gene segments along the X axis and the percentage of reads assigned to each V gene segment along the Y axis. The percentage is currently being displayed for two samples, as indicated by the blue and black bars. **(B)** CDR3 length histogram. CDR3 length is shown on the X axis, and the percentage of reads with that CDR3 length is shown on the Y axis. The percentage is currently shown for two sample groups as indicated by the blue and black bars. When the data for sample groups, rather than samples, are displayed, the bar height represents the average percentage across all samples in the group, and the error bars indicate the SD. **(C)** Ranked clonal abundance percentage. Each colored line represents a sample with the clones ranked from highest abundance (rank 1) to lowest abundance along the X axis and the corresponding percentage of reads for each clone along the Y axis. **(D)** Cumulative clonal abundance. Each colored line represents a sample with clones ranked from highest abundance (rank 1) to lowest abundance along the X axis and the cumulative percentage along the Y axis. **(E)** Diversity profile. Each colored line represents a sample where clonal diversity along the Y axis is calculated across a sweep of the ordering parameter (Q) along the X axis. A point of $Q = 1$ corresponds to the Shannon entropy. **(F)** Quantification of selection pressure for CDR and framework regions. Region is shown on the X axis, and the value of the selection parameter is shown on the Y axis. Negative bar values indicate negative selection, and positive bar values indicate positive selection. The error bars show the 95% confidence interval of the selection parameter along with a p -value for significance. Two B cell samples are currently displayed as indicated by the blue and black bars.

neurological disease (OND). The first study also included CD19+ CD27– naïve B cells from the peripheral blood of one healthy donor. 454 sequencing was performed upon 40 samples for the first study (13 RRMS, 26 OND, 10 replicates from the single healthy donor), and an additional 102 samples for the second study (60 RRMS, 42 OND). All CSF samples were collected by lumbar puncture in accordance with IRB-approved protocols at UT Southwestern Medical Center, the University of Massachusetts Memorial Medical Center (UMass), Johns Hopkins University, or purchased from a commercial biorepository (PrecisionMed, Solana Beach, CA, USA).

The analysis workflow included VDJPipe for preprocessing, IgBlast for V(D)J assignment, and RepSum for rearrangement annotation. VDJPipe preprocessing trimmed the reads of primers and sample barcode sequences as well as filtered reads with average quality <35 or length shorter than 200 bp. Duplicate sequences were collapsed by VDJPipe with the resultant set of unique sequences input to IgBlast. RepSum annotation provided a secondary filtering step by eliminating sequences with frame-shifting insertions or deletions, out of frame junctions, stop codons present, truncated read lengths, less than 85% homology to germline sequences, missing CDR3s, or missing read coverage between Chothia-numbered codons 31 and 92. Also, unique sequence reads with fewer than two copies in the raw sequence data were discarded, as containing possible sequencing errors. Because sufficient B cells are not always obtained from CSF, samples were required to have at least 10 unique reads, otherwise they were excluded from further analysis. The filtering was performed with basic logical operations on the annotation fields produced by RepSum, thus providing a high-quality set of sequences for mutational analysis. VDJServer reports silent and replacement mutations at the nucleotide and codon level as part of the rearrangement annotation with replacement mutations indicating an amino acid substitution.

A second study applied machine learning to the CDR3 sequences within each sample repertoire to identify a biochemical motif present in the CDR3 sequences of RRMS repertoires but not OND repertoires (66). The analysis workflow was as described earlier, except pRESTO was used for preprocessing, and length filtering removed reads shorter than 300 bp. An additional step was performed to remove all reads shared between two or more samples. The remaining CDR3 sequences were used as input to the statistical classifier algorithm described in Ref. (66).

MS is an autoimmune disease that is notoriously difficult to diagnose, but early detection is needed because prompt intervention can significantly slow the progression of the disease. The VDJServer analysis was used for the development and refinement of a more accurate diagnostic test (MSPrecise) for RRMS based on the replacement mutation frequency for five V gene segment codons (14), and for the discovery of a biochemical motif present in the CDR3 sequences of RRMS repertoires but not OND repertoires (66).

Use Case: Chimeric Antigen Receptor (CAR) T-Cells

Acute myeloid leukemia (AML) is the most common acute leukemia in adults, presenting greater than 20,000 new cases per

year in the US and representing 80% of acute leukemias. The field of anti-cancer T-cell therapy has had major advances in recent years with the development of the CAR. A recent study examining AML has shown that a TIM3 monoclonal antibody blocks AML engraftment and eliminates leukemic stem cells (67). Those results, as well as others, led Davila and colleagues to hypothesize that anti-TIM3 CAR T-cells could target and kill AML. To test this hypothesis, mice were immunized with either Chinese hamster ovary (CHO) cells or CHO cells expressing TIM3. Immune repertoire sequencing was performed on splenocytes from both groups of mice, and repertoire comparison between the two groups was used to identify immunoglobulin (Ig) heavy and light chain genes that together form a receptor with TIM3 specificity. These genes were then used to develop novel, anti-TIM3 CARs. Animal studies were performed in accordance with the principles of the Basel Declaration and following protocols reviewed and approved by the Institutional Animal Care and Use Committee at the University of South Florida.

The Illumina sequencing platform was utilized to provide a total of over 18 million raw, paired-end sequence reads for Ig heavy and light chains from two control mice and three TIM3-treated mice. The analysis workflow included VDJPipe for preprocessing, IgBlast for V(D)J gene segment assignment, RepSum for rearrangement annotation, and RepCalc for repertoire comparison. preprocessing with VDJPipe merged the paired-end reads, filtered reads with average quality <25 or length shorter than 200 bp, and collapsed duplicate sequences. The resulting 12 million unique sequences for the five samples were given V(D)J gene segment assignments by IgBlast and annotated with RepSum. The analysis design entailed identifying rearranged Ig genes that were highly abundant in the TIM3-treated mice yet not present or very low in abundance in the control mice. We defined TIM3 and control sample groups using VDJServer's study metadata entries, and RepCalc performed the repertoire comparison between the two sample groups. RepCalc performed both intragroup and intergroup comparison for V-J combinations, and V-J-CDR3 combinations, where the CDR3 was compared at both the amino acid and nucleotide sequence levels. RepCalc identifies shared and unique combinations and calculates abundance amounts for each sample and sample group. Sharing levels are provided that indicate how many samples within a sample group share a specific combination along with their abundance, and those sharing levels are also compared between sample groups. For example, a particular V-J combination might be present in one of three (intragroup unique), two of three (intragroup shared), or all three (intragroup shared) TIM3-treated mice, and one of two (intragroup unique) or all two (intragroup shared) control mice. While a combination that is present in at least one sample within a sample group but not present in any samples of the other sample group is intergroup unique, a combination that is present in at least one sample in both sample groups is intergroup shared. By analyzing the intragroup and intergroup comparisons along with their associated abundances, the Davila group could confidently identify rearranged Ig genes that were likely to respond to TIM3.

The use of immune repertoire sequencing with VDJServer analysis, in particular the detailed sharing levels produced by

RepCalc, provides a rapid, economical system for the development of novel CARs that eliminated the need for hybridoma production and screening.

CONCLUSION

VDJServer is a web-accessible cloud platform that provides access through a graphical user interface to a data management infrastructure, a collection of analysis tools covering all steps in an analysis, and an infrastructure for sharing data along with workflows, results, and computational provenance. VDJServer has been successfully used across a broad range of immune repertoire analysis projects, and is a free, publicly available, and open-source licensed resource.

VDJServer has been designed for long-term stability regardless of financial constraints that can often hamper research-oriented web resources. Specifically, VDJServer's primary dependency is the cloud infrastructure provided by Agave (61), which is a core technology of the Texas Advanced Computer Center with a dedicated development team that is being continually enhanced and supports numerous other web resources besides VDJServer. The VDJServer-specific components are relatively small and easily fit on a small virtual machine, and the components have been containerized with Docker, which means the system can run on a user's desktop if desired or run on a commercial cloud platform, such as Amazon Web Services or Microsoft Azure. Notably, Agave does not require use of TACC computing or storage resources. Because Agave abstracts the notion of execution systems (where jobs are run) and storage systems (where data are stored), users can define execution and storage systems that point to their own compute resources (or a commercial cloud platform) without loss of functionality, though the various analysis tools would need to be installed on the execution system. Nevertheless, in this scenario, the financial costs of the compute resources are shifted to the user.

VDJServer is continually advancing to meet the needs of the user community in the quickly developing field of immune repertoire analysis. Future goals include integrating new analysis tools and algorithms, providing additional interactive visualizations, enabling user queries across study metadata and rearrangement annotations, and conducting training and community outreach. With publication of the AIRR Minimal Information Standards (41), VDJServer currently ensures study metadata conforms to those standards. In the future, VDJServer will provide users the capability to submit their complete study, including project data, metadata, analysis, and results to a database in the International Nucleotide Sequence Database Collaboration [e.g., National Center for Biotechnology Information's (NCBI) Sequence Read Archive and genetic sequence database (GenBank)] per the AIRR Minimal Standards recommendations. VDJServer will handle all of the technical details regarding data formatting requirements and standards compliance for submission, with accession numbers provided for journal publication. VDJServer will continue to evolve to remain consistent with AIRR standards as they are developed and released.

AVAILABILITY AND REQUIREMENTS

Project name: VDJServer
 Project home page: <https://vdjserver.org/>
 Source code repository: <https://bitbucket.org/vdjserver>
 Docker images: <https://hub.docker.com/r/vdjserver>
 Operating system(s): Platform independent
 Programming language: JavaScript, Python
 License: MIT, GNU GPL
 Any restrictions to use by non-academics: no restrictions
 No datasets were generated for this study.

ETHICS STATEMENT

While the development of VDJServer did not involve the use of human or animal subjects, the two use cases presented to demonstrate VDJServer functionality did. The MS use case utilized human subjects. All samples were collected in accordance with IRB-approved protocols at UT Southwestern Medical Center, the University of Massachusetts Memorial Medical Center (UMass), Johns Hopkins University (JHU), or purchased from a commercial biorepository (PrecisionMed, Solana Beach, CA, USA). The protocols included an informed consent process. The CAR T-cells use case utilized animal subjects. All of the animal studies were performed in accordance with the principles of the Basel Declaration and following protocols reviewed and approved by the Institutional Animal Care and Use Committee at the University of South Florida.

AUTHOR CONTRIBUTIONS

SC, WS, ES, WR, IT, JF, ML, MK, SM, CJ, and JO designed, implemented, and tested the software code. AB and FR participated in testing through the web interface. MD and NM contributed driving use cases. NM, RS, and LC conceived the project, determined required user functionality, and provided feedback on the design. SC and LC wrote the manuscript. All the authors read and approved the final manuscript.

FUNDING

This work was supported by an NIAID-funded R01 (AI097403) to LGC. FR was supported in part through an NIAID-funded U19 (AI057229). The CAR T cell use case was supported by funding to MD from the H. Lee Moffitt Cancer Center and Research Institute. The MS use case was supported by funding to NLM from DioGenix, Inc. and the National Multiple Sclerosis Society.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at <https://www.frontiersin.org/articles/10.3389/fimmu.2018.00976/full#supplementary-material>.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302 (5909):575–81. doi:10.1038/302575a0
2. Berek C, Milstein C. The dynamic nature of the antibody repertoire. *Immunol Rev* (1988) 105:5–26. doi:10.1111/j.1600-065X.1988.tb00763.x
3. Lythe G, Callard RE, Hoare RL, Molina-Paris C. How many TCR clonotypes does a body maintain? *J Theor Biol* (2016) 389:214–24. doi:10.1016/j.jtbi.2015.10.016
4. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) 21(5):790–7. doi:10.1101/gr.115428.110
5. Burnet FM. The clonal selection theory of acquired immunity. *The Abraham Flexner lectures, 1958*. Nashville: Vanderbilt University Press (1959). xiii, 209 p.
6. Nikolich-Zugich J, Rudd BD. Immune memory and aging: an infinite or finite resource? *Curr Opin Immunol* (2010) 22(4):535–40. doi:10.1016/j.coi.2010.06.011
7. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J Immunol* (2014) 192(6):2689–98. doi:10.4049/jimmunol.1302064
8. Sobolev O, Binda E, O'Farrell S, Lorenc A, Pradines J, Huang Y, et al. Adjuvanted influenza-H1N1 vaccination reveals lymphoid signatures of age-dependent early responses and of clinical adverse events. *Nat Immunol* (2016) 17(2):204–13. doi:10.1038/ni.3328
9. Zhu J, Wu X, Zhang B, McKee K, O'Dell S, Soto C, et al. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc Natl Acad Sci U S A* (2013) 110(43):E4088–97. doi:10.1073/pnas.1306262110
10. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* (2016) 6:20842. doi:10.1038/srep20842
11. Wang B, Kluwe CA, Lungu OI, DeKosky BJ, Kerr SA, Johnson EL, et al. Facile discovery of a diverse panel of anti-Ebola virus antibodies by immune repertoire mining. *Sci Rep* (2015) 5:13926. doi:10.1038/srep13926
12. Strauli NB, Hernandez RD. Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Med* (2016) 8(1):60. doi:10.1186/s13073-016-0314-z
13. Chang YH, Kuan HC, Hsieh TC, Ma KH, Yang CH, Hsu WB, et al. Network signatures of IgG immune repertoires in hepatitis B associated chronic infection and vaccination responses. *Sci Rep* (2016) 6:26556. doi:10.1038/srep26556
14. Rounds WH, Salinas EA, Wilks TB II, Levin MK, Ligocki AJ, Ionete C, et al. MSPrecise: a molecular diagnostic test for multiple sclerosis using next generation sequencing. *Gene* (2015) 572(2):191–7. doi:10.1016/j.gene.2015.07.011
15. Johansen JN, Vartdal F, Desmarais C, Tuttunen AE, de Souza GA, Lossius A, et al. Intrathecal BCR transcriptome in multiple sclerosis versus other neuroinflammation: equally diverse and compartmentalized, but more mutated, biased and overlapping with the proteome. *Clin Immunol* (2015) 160(2):211–25. doi:10.1016/j.clim.2015.06.001
16. Lossius A, Johansen JN, Vartdal F, Robins H, Jüratė Šaltytė B, Holmøy T, et al. High-throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV-reactive CD8+ T cells. *Eur J Immunol* (2014) 44(11):3439–52. doi:10.1002/eji.201444662
17. Ria F, Penitente R, De Santis M, Nicolò C, Di Sante G, Orsini M, et al. Collagen-specific T-cell repertoire in blood and synovial fluid varies with disease activity in early rheumatoid arthritis. *Arthritis Res Ther* (2008) 10(6):R135. doi:10.1186/ar2553
18. Thapa DR, Tonikian R, Sun C, Liu M, Dearth A, Petri M, et al. Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing. *Arthritis Res Ther* (2015) 17:132. doi:10.1186/s13075-015-0655-9
19. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, et al. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc Natl Acad Sci U S A* (2011) 108(52):21194–9. doi:10.1073/pnas.1118357109
20. Carlotti E, Wrench D, Rosignoli G, Marzec J, Sangaralingam A, Hazanov L, et al. High throughput sequencing analysis of the immunoglobulin heavy chain gene from flow-sorted B cell sub-populations define the dynamics of follicular lymphoma clonal evolution. *PLoS One* (2015) 10(9):e0134833. doi:10.1371/journal.pone.0134833
21. Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott LL, Elco CP, et al. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Sci Transl Med* (2015) 7(308):308ra158. doi:10.1126/scitranslmed.aaa9122
22. Martinez-Lopez J, Lahuerta JJ, Pepin F, González M, Barrio S, Ayala R, et al. Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma. *Blood* (2014) 123(20):3073–9. doi:10.1182/blood-2014-01-550020
23. Kurtz DM, Green MR, Bratman SV, Scherer F, Liu CL, Kunder CA, et al. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood* (2015) 125(24):3679–87. doi:10.1182/blood-2015-03-635169
24. Robins HS, Ericson NG, Guenthoer J, O'Brian KC, Tewari M, Drescher CW, et al. Digital genomic quantification of tumor-infiltrating lymphocytes. *Sci Transl Med* (2013) 5(214):214ra169. doi:10.1126/scitranslmed.3007247
25. Postow MA, Manuel M, Wong P, Yuan J, Dong Z, Liu C, et al. Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. *J Immunother Cancer* (2015) 3:23. doi:10.1186/s40425-015-0070-4
26. Jia Q, Zhou J, Chen G, Shi Y, Yu H, Guan P, et al. Diversity index of mucosal resident T lymphocyte repertoire predicts clinical prognosis in gastric cancer. *Oncoimmunology* (2015) 4(4):e1001230. doi:10.1080/2162402X.2014.1001230
27. Iglesia MD, Vincent BG, Parker JS, Hoadley KA, Carey LA, Perou CM, et al. Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. *Clin Cancer Res* (2014) 20(14):3818–29. doi:10.1158/1078-0432.CCR-13-3368
28. Han Y, Liu X, Wang Y, Wu X, Guan Y, Li H, et al. Identification of characteristic TRB V usage in HBV-associated HCC by using differential expression profiling analysis. *Oncoimmunology* (2015) 4(8):e1021537. doi:10.1080/2162402X.2015.1021537
29. Robert L, Tsoi J, Wang X, Emerson R, Homet B, Chodon T, et al. CTLA4 blockade broadens the peripheral T-cell receptor repertoire. *Clin Cancer Res* (2014) 20(9):2424–32. doi:10.1158/1078-0432.CCR-13-2648
30. Cooper ZA, Frederick DT, Juneja VR, Sullivan RJ, Lawrence DP, Piris A, et al. BRAF inhibition is associated with increased clonality in tumor-infiltrating lymphocytes. *Oncoimmunology* (2013) 2(10):e26615. doi:10.4161/onci.26615
31. Meyer EH, Hsu AR, Liliental J, Löhr A, Florek M, Zehnder JL, et al. A distinct evolution of the T-cell repertoire categorizes treatment refractory gastrointestinal acute graft-versus-host disease. *Blood* (2013) 121(24):4955–62. doi:10.1182/blood-2013-03-489757
32. Morris H, DeWolf S, Robins H, Sprangers B, LoCascio SA, Shonts BA, et al. Tracking donor-reactive T cells: evidence for clonal deletion in tolerant kidney transplant patients. *Sci Transl Med* (2015) 7(272):272ra10. doi:10.1126/scitranslmed.3010760
33. Vollmers C, De Vlaminck I, Valantine HA, Penland L, Luikart H, Strehl C, et al. Monitoring pharmacologically induced immunosuppression by immune repertoire sequencing to detect acute allograft rejection in heart transplant patients: a proof-of-concept diagnostic accuracy study. *PLoS Med* (2015) 12(10):e1001890. doi:10.1371/journal.pmed.1001890
34. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* (2015) 7(1):49. doi:10.1186/s13073-015-0169-8
35. Robins HS, Campregher PV, Srivastava SK, Wachter A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) 114(19):4099–107. doi:10.1182/blood-2009-04-217604
36. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight

- into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106(48):20216–21. doi:10.1073/pnas.0909775106
37. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) 324(5928):807–10. doi:10.1126/science.1170020
 38. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* (2009) 19(10):1817–24. doi:10.1101/gr.092924.109
 39. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* (2009) 1(12):12ra23. doi:10.1126/scitranslmed.3000540
 40. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front Immunol* (2017) 8:1418. doi:10.3389/fimmu.2017.01418
 41. Rubelt F, Busse CE, Bukhari SAC, Bürckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* (2017) 18(12):1274–8. doi:10.1038/ni.3873
 42. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41(Web Server issue):W34–40. doi:10.1093/nar/gkt382
 43. Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) 4:2333. doi:10.1038/ncomms3333
 44. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P. IMGT/HighV-QUEST: the IMGT web portal for immunoglobulin (Ig) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immun Res* (2012) 8:26. doi:10.4172/1745-7580.1000056
 45. Giudicelli V, Brochet X, Lefranc MP. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* (2011) 2011(6):695–715. doi:10.1101/pdb.prot5633
 46. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. IMGT/Junction Analysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONS. *Bioinformatics* (2004) 20(Suppl 1):i379–85. doi:10.1093/bioinformatics/bth945
 47. Giudicelli V, Lefranc MP. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* (2011) 2011(6):716–25. doi:10.1101/pdb.prot5634
 48. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F, Vidjil: a web platform for analysis of high-throughput repertoire sequencing. *PLoS One* (2016) 11(11):e0166126. doi:10.1371/journal.pone.0166126
 49. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* (2014) 30(5):614–20. doi:10.1093/bioinformatics/btt593
 50. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* (2015) 12(5):380–1. doi:10.1038/nmeth.3364
 51. Moorhouse MJ, van Zessen D, IJspeert H, Hiltmann S, Horsman S, van der Spek PJ, et al. Immunoglobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunol* (2014) 15:59. doi:10.1186/s12865-014-0059-7
 52. IJspeert H, van Schouwenburg PA, van Zessen D, Pico-Knijnenburg I, Stubbs AP, van der Burg M. Antigen receptor galaxy: a user-friendly, web-based tool for analysis and visualization of T and B cell receptor repertoire data. *J Immunol* (2017) 198(10):4156–65. doi:10.4049/jimmunol.1601921
 53. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* (2005) 15(10):1451–5. doi:10.1101/gr.4086505
 54. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) 31(20):3356–8. doi:10.1093/bioinformatics/btv359
 55. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) 40(17):e134. doi:10.1093/nar/gks457
 56. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. ImmuneDB: a system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* (2017) 33(2):292–3. doi:10.1093/bioinformatics/btw593
 57. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30(13):1930–2. doi:10.1093/bioinformatics/btu138
 58. Zhang B, Meng W, Prak ET, Hershberg U. Discrimination of germline V genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment. *J Immunol Methods* (2015) 427:105–16. doi:10.1016/j.jim.2015.10.009
 59. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. *Gigascience* (2016) 5(1):30. doi:10.1186/s13742-016-0135-4
 60. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* (2012) 13(9):667–72. doi:10.1038/nrg3305
 61. Dooley R, et al. Software-as-a-service: the iPlant Foundation API. *5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers*. Salt Lake City, Utah (2012).
 62. Christley S, Levin MK, Toby IT, Fonner JM, Monson NL, Rounds WH, et al. VDJPipe: a pipelined tool for pre-processing immune repertoire sequencing data. *BMC Bioinformatics* (2017) 18(1):448. doi:10.1186/s12859-017-1853-z
 63. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* (2015) 43(Database issue):D413–22. doi:10.1093/nar/gku1056
 64. Toby IT, Levin MK, Salinas EA, Christley S, Bhattacharya S, Breden F, et al. VDJML: a file format with tools for capturing the results of inferring immune receptor rearrangements. *BMC Bioinformatics* (2016) 17(Suppl 13):333. doi:10.1186/s12859-016-1214-3
 65. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* (2016) 3:160018. doi:10.1038/sdata.2016.18
 66. Ostmeier J, Christley S, Rounds WH, Toby I, Greenberg BM, Monson NL, et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics* (2017) 18(1):401. doi:10.1186/s12859-017-1814-6
 67. Jan M, Chao MP, Cha AC, Alizadeh AA, Gentles AJ, Weissman IL, et al. Prospective separation of normal and leukemic stem cells based on differential expression of TIM3, a human acute myeloid leukemia stem cell marker. *Proc Natl Acad Sci U S A* (2011) 108(12):5009–14. doi:10.1073/pnas.1100551108
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- The reviewer NP declared a past co-authorship with the author to the handling Editor.

Copyright © 2018 Christley, Scarborough, Salinas, Rounds, Toby, Fonner, Levin, Kim, Mock, Jordan, Ostmeier, Buntzman, Rubelt, Davila, Monson, Scheuermann and Cowell. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.