Title

RImmPort: an R/Bioconductor package that enables ready-for-analysis immunology research data.

Journal

Bioinformatics, 33(7)

ISSN

1367-4803

Authors

Shankar, Ravi D
Bhattacharya, Sanchita
Jujjavarapu, Chethan
et al.

Publication Date

2017-04-01

DOI

10.1093/bioinformatics/btw719

Peer reviewed

Bioinformatics, 33(7), 2017, 1101–1103
doi: 10.1093/bioinformatics/btw719
Advance Access Publication Date: 30 December 2016
Applications Note

OXFORD

Databases and ontologies

RImmPort: an R/Bioconductor package that enables ready-for-analysis immunology research data

Ravi D. Shankar[1,*], Sanchita Bhattacharya[2], Chethan Jujjavarapu[2], Sandra Andorf[1], Jeffery A. Wiser[3] and Atul J. Butte[2]

[1]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA, [2]Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA 94158, USA and [3]Northrop Grumman Information Technology Health Solutions, Rockville, MD 20850, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 16, 2016; revised on November 7, 2016; editorial decision on November 8, 2016; accepted on November 8, 2016

Abstract

Summary: Open access to raw clinical and molecular data related to immunological studies has created a tremendous opportunity for data-driven science. We have developed RImmPort that prepares NIAID-funded research study datasets in ImmPort (immport.org) for analysis in R. RImmPort comprises of three main components: (i) a specification of R classes that encapsulate study data, (ii) foundational methods to load data of a specific study and (iii) generic methods to slice and dice data across different dimensions in one or more studies. Furthermore, RImmPort supports open formalisms, such as CDISC standards on the open source bioinformatics platform Bioconductor, to ensure that ImmPort curated study datasets are seamlessly accessible and ready for analysis, thus enabling innovative bioinformatics research in immunology.

Availability and Implementation: RImmPort is available as part of Bioconductor (bioconductor.org/packages/RImmPort).

Contact: rshankar@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Overview of ImmPort and RImmPort

Publicly available individual-level clinical trial data has created a tremendous opportunity to evaluate new research hypotheses that were not originally formulated in the studies, either by reanalyzing data from a study, by performing cross analysis of multiple studies, or by combining study data with other public research datasets (Warren, 2016). Such analysis of disparate data presupposes (i) uniform representation of clinical trial data using data standards and (ii) easy access to standardized clinical trial data in analytical environments.

ImmPort (Bhattacharya *et al.*, 2014), the Immunology Database and Analysis Portal system (immport.org), is the archival repository for clinical study data generated by scientific researchers primarily funded by the National Institute of Allergy and Infectious Diseases (NIAID)-Division of Allergy, Immunology and Transplantation (DAIT). With over 200 studies now publicly available, ImmPort is an important source of raw data and standard protocols from clinical trials, biological studies and novel immunologic assays for cellular and molecular measurements. The high-dimensional data generated by the assays hold big promise in the area of human medicine, especially recently in cancer immunotherapy (Yuan *et al.*, 2016).

ImmPort studies generate large datasets, comprising of different data types including study design, clinical and mechanistic assays data. The study data in ImmPort is structured using a relational data model and is available for download in Tab and MySQL formats.

The data model is designed for efficient data storage, and distributes datasets across multiple relational tables. However,

accessing specific data for analysis becomes challenging for research scientists. We illustrate the data access issues with ImmPort Study: *SDY1* (immport.org/immport-open/public/study/study/displayStudyDetail/SDY1). This study evaluated the efficacy and safety of allergen immunotherapy when combined with Omalizumab (Casale *et al.*, 2006). The study data included subject demographics, clinical assessments, adverse events, interventions, concomitant medications, results of hematology and total IgE lab test results, raw and/or processed data from flow cytometry and ELISA experiments on 4211 biosamples collected at different time points over 12 weeks from 159 subjects. All this study data is stored across 31 tables in the ImmPort data model. Loading the data into R (https://www.r-project.org/) requires enormous effort in first understanding the ImmPort data model and then writing complex queries to retrieve the data from a large number of tables. The situation is further exasperated when the underlying data model has to change, especially to accommodate data from new assay technologies.

We have built the RImmPort package to streamline the accessibility and inter-operability of ImmPort data for analysis in the R statistical environment. To aid in the secondary reuse of ImmPort data, RImmPort implements a data model that is based on the Clinical Data Interchange Standards Consortium (CDISC: www.cdisc.org) clinical trial data standards, and supports a suite of functions that provide access to different types of ImmPort study data. The package has been released in R/Bioconductor (bioconductor.org/packages/RImmPort).

2 RImmPort software components

RImmPort comprises of (i) a foundational data model that encapsulates ImmPort study data. The model leverages and incorporates terms and semantics supported by CDISC standards. (ii) A suite of data access methods to query and load different attributes of a specific study, methods to integrate assay-specific datasets from multiple studies and utility methods to serialize study data in RImmPort data model.

To access ImmPort data using RImmPort, a user (i) downloads the ZIP file that contains the Tab-formatted study data of interest from ImmPort website into a local directory; (ii) in R, loads the RImmPort library, (iii) using RImmPort provided commands, builds a SQLite database from the ZIP file that contains the study data, and sets the new SQLite database as the data source; and (iv) calls RImmPort methods to load ImmPort data of interest into the R environment.

Using RImmPort, an entire study can be loaded into R with a single command. For example, a researcher can run RImmPort method *getStudy('SDY1')* to load the large dataset associated with ImmPort Study: *SDY1* into R.

3 RImmPort data model

The RImmPort data model is based on the CDISC clinical trial data standards. CDISC has been developing a suite of data standards that support data management at different stages of the clinical research process: study design, study conduct, data analysis and reporting. CDISC's reporting standard is the Study Data Tabulation Model (SDTM) (Kubick *et al.*, 2007) that can be used to structure study data when submitting to regulatory agencies such as the United States Food and Drug Administration (FDA). In SDTM, the data is structured into domains and each domain defines a standard set of variables that are appropriate for that domain. For example, the *Demographics* domain comprises of variables such as age, sex, ethnicity and race. RImmPort adapts SDTM data standard to encapsulate the different types of ImmPort study data as domains including *Demographics, Laboratory Test Results, Genetics Findings, and Cellular Quantification.*

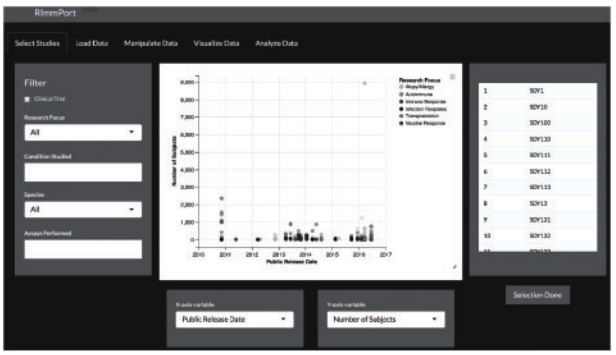

Fig. 1. Using the visual interfaces, research scientists can easily select studies of interest and load specific domain data of these studies

4 RImmPort data access functions

RImmPort supports a set of functions to load different types of ImmPort study data. Functions include *getStudy* that loads entire data of a study and *getDomainDataOfStudies* that loads specific domain data of one or more studies. For example, the R code snippet

> sdy1_dm <- getDomainDataOfStudies('Demographics', 'SDY1')

loads the Demographics data of ImmPort Study: *SDY1*.

Internally, the functions query for specific study data from the ImmPort data source, organize the data into the RImmPort model, and then load the data into the R environment. We are also building interactive visual components in RImmPort that allows users to easily access and manipulate study data of interest (Fig. 1).

The vignettes that are bundled with the RImmPort package give additional details on the RImmPort data model and usage of RImmPort functions.

5 Conclusion

We have based the RImmPort data model on the CDISC open clinical study data standards. The RImmPort package has been released on Bioconductor, thus enabling access to ImmPort data in a rich environment for various data analyses. Besides providing utility in enabling data scientists familiar with R to access ImmPort clinical studies, we believe this package is the first instantiation of FDA-relevant standards like CDISC to be implemented in R. We continue to enhance RImmPort with new data access and data manipulation functions, and assay-specific analysis pipelines. We are building an interactive RImmPort workbench that research scientists can use to easily access and analyze ImmPort data. RImmPort is a standards-based open platform that enables innovative bioinformatics research in immunology.

Funding

This work was supported in part by the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN272201200028C).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Bhattacharya,S. *et al*. (2014) ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res*., **5858**, 234–239.

Casale,T.B. *et al*. (2006) Omalizumab pretreatment decreases acute reactions after rush immunotherapy for ragweed-induced seasonal allergic rhinitis. *J. Allergy Clin. Immunol*., **117117**, 134–140.

Kubick,W.R. *et al*. (2007) Toward a comprehensive CDISC submission data standard. *Therap. Innov. Regul. Sci*., **4141**, 373–382.

Warren,E. (2016) Strengthening research through data sharing. *N. Engl. J. Med*., **375375**, 401–403.

Yuan,J. *et al*. (2016) Novel technologies and emerging biomarkers for personalized cancer immunotherapy. *J. Immunother. Cancer*, **44**, 3.