LINCS Dataset Registry (LDR): A Web-Based System to Capture and Manage LINCS Data Releases with Autocomplete Web Forms

Available at http://amp.pharm.mssm.edu/LDR

Michael G. McDermott, BS[1,2], Qiaonan Duan, BS[1,2], Sherry L. Jenkins, MS[1,2], Amar Koleti, MS[2,3], Dusica Vidovic, PhD[2,3] , Stephan Schurer, PhD[2,3], Chris Mader, MS[2,3], Michael J. MacCoss, PhD[4], Avi Ma'ayan, PhD[1,2]

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1215, New York, NY 10029 USA; [2]BD2K-LINCS Data Coordination and Integration Center (DCIC); [3]Center for Computational Science, University of Miami, Miami, FL, USA; [4]Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA

BD2K - LINCS DATA COORDINATION AND INTEGRATION CENTER

Icahn School of Medicine at Mount Sinai

Abstract

One of the challenges of BD2K is to capture metadata about dataset instances and link such metadata to controlled dictionaries and ontologies. This metadata capture is expected to improve dataset search and facilitate data integration. This challenge is central to the LINCS program and the BD2K-LINCS DCIC because different LINCS data generation centers use different but overlapping assays, perturbations (genes/proteins, small-molecules), cell-lines, disease models, readouts and other common entities. Most major data repositories such as GEO or Chorus currently do not have advanced web-based forms to capture metadata about dataset instances from their data submitters. In year 1, the BD2K-LINCS DCIC developed the LINCS Dataset Registry (LDR) system to capture, visualize and manage all LINCS released datasets. LDR is a modern, mobile-friendly web application designed to streamline the process of submitting, approving, and releasing datasets. LDR consists of a client-side application created with the JavaScript library AngularJS and a web server application written in NodeJS. The server's extensive API's communicate to a MongoDB database responsible for storing and querying each center's data. LDR has login authentication functionality that enables the security of unreleased datasets, and its advanced input forms allow for fast, hassle-free data entry. Form entities have autocomplete functionality drawing from ontologies and dictionaries managed by live remote servers. LDR also contains a dataset-specific message board that enables communication between the LINCS data generation centers and the NIH staff to facilitate an approval process. While designed for LINCS, LDR will be generalized to facilitate data capture for other projects. For a BD2K Common Fund supplement in collaboration with the MacCoss Lab at the University of Washington, we will reuse the LDR code as an additional feature for Chorus, a new cloud-based application that provides scientists with the ability to securely store, analyze and share their MS data regardless of the original raw file format. The goal of Chorus is to create a complete catalogue of the world's mass spectrometric data that can be openly accessed by, and freely accessible to, the global scientific community as well as the general public. Hence, the web-forms, developed for LINCS, will be provided as a web-service and will be embedded within Chorus. This effort will serve as a model for the Chorus/proteomics community to better annotate their proteomics data submissions but will also introduce a global solution for other repositories.

Methods

The LINCS Dataset Registry is composed of a web application written in HTML, CSS, and Javascript (AngularJS), a web server written in NodeJS that interacts with the metadata registry hosted at the University of Miami in addition to its own MongoDB database. The interaction of these components is shown in figure 1.

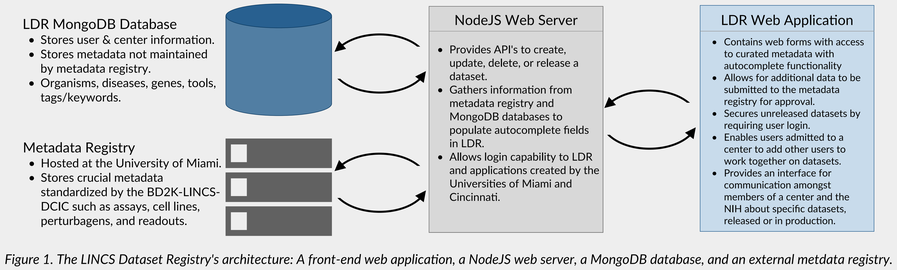

Figure 1. The LINCS Dataset Registry's architecture: A front-end web application, a NodeJS web server, a MongoDB database, and an external metdata registry.

Results

The LINCS Dataset Registry is accessed at http://amp.pharm.mssm.edu/LDR/#/. From the homepage (figure 2), a user can both view and search released datasets, view database statistics, or login to navigate to the rest of the application. Once logged in, the user is given the ability to not only view or manipulate their center's datasets, but edit their center's icon, admit new users, and view statistics specific to their center.

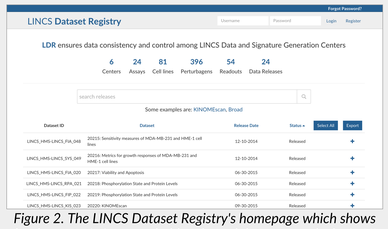

Figure 2. The LINCS Dataset Registry's homepage which shows database statistics, searchable released datasets, and a login UI.

Figure 3. Options in the navigation bar available to a logged in user.

The heart of LDR is its dataset submission process shown in figure 4. This is started by clicking the "Dataset Statistics" drop-down and clicking "Add New" (Shown in figure 3). From here, the user can add all relevant metadata specific to their dataset release instance. Fields that require standardized entries, such as cell lines and perturbagens, have autocomplete functionality. Terms populating the autocomplete options are pulled from the BD2K LINCS-DCIC metadata registry, hosted by the University of Miami, which contains controlled dictionaries and ontologies specific to their relevant field. While it is required to select one of these options, users have the ability to add terms of their own. Figure 4 showcases this process. These new terms are submitted to the registry to be curated and added to the controlled dictionaries if needed. This process ensures that entity names, descriptions, and metadata are consistent across all LINCS Data Signature Generation Centers.

While the LDR application is specific to the LINCS program, its forms and autocomplete functionality work separately from the servers responsible for populating them. This modular nature should make re-implementation in Chorus straight-forward to improve the process of capturing metadata associated to mass spectrometry data.

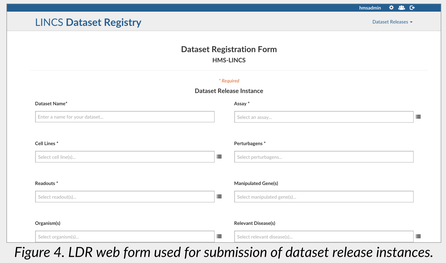

Figure 4. LDR web form used for submission of dataset release instances.

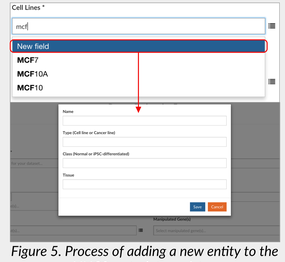

Figure 5. Process of adding a new entity to the LINCS metadata registry.

Conclusions

While the data and signature generation pipelines have been fine-tuned amongst members of the LINCS consortium, their data release processes and associated metadata have not. Improving the process of capturing metadata will not only facilitate an increase in search capabilities and data integration across LINCS applications, but it will entice researchers to increase the quality and quantity of metadata attributed to a given data release.

By generalizing LDR to function outside of LINCS, this increase in productivity can be transferred to other fields of biomedical research. LDR's re-implementation in Chorus will improve the user's ability to catalogue, search, and share their mass spectrometry data. Built with the researcher in mind, the LINCS Dataset Registry allows the user to spend less time submitting datasets and more time generating them.

References

- Vempati,U.D. *et al.* (2014) Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the library of integrated network-based cellular signatures (LINCS). *J. Biomol. Screen.*, 19, 803–816.
- AngularJS: HTML enhanced for web apps! - Google ©2010-2015. Details available at https://angularjs.org/
- Node.js: An open-source, cross-platform runtime environment for developing server-side web applications - ©2015 Node.js foundation, ©2015 Joyent - https://nodejs.org
- MongoDB - ©2015 MongoDB, Inc. - https://mongodb.org

Acknowledgements

This work is supported by NIH grants: R01GM098316, U54HG008230 and U54CA189201 to AM.