The Stem Cell Commons: an exemplar for data integration in the biomedical domain driven by the ISA framework

The Stem Cell Commons: an exemplar for data integration in the biomedical domain driven by the ISA framework

Shannan Ho Sui[1,2] , Emily Merrill[3], Nils Gehlenborg[4], Psalm Haseley[4], Ilya Sytchev[1], Richard Park[4,5], Philippe Rocca-Serra[7], Stephane Corlosquet[3], Alejandra Gonzalez-Beltran[7], Eamonn Maguire[7], Oliver Hofmann[1,2], Peter Park[4,2], Sudeshna Das[3,6], Susanna-Assunta Sansone[7], Winston Hide[1,2]

[1]Harvard School of Public Health, Biostatistics, Boston, MA, USA, [2]Harvard Stem Cell Institute, Cambridge, MA, USA,[3] Massachusetts General Hospital, MassGeneral Institute for Neurodegenerative Disease, Cambridge, MA, USA, [4]Harvard Medical School, Center for Biomedical Informatics, Boston, MA, USA, [5]Bioinformatics, Boston University, Boston, MA, USA, [6]Harvard Medical School, Neurology, Boston, MA, USA [7]University of Oxford, e-Research Centre, Oxford, UK

Abstract

Comparisons of stem cell experiments at both molecular and semantic levels remain challenging due to inconsistencies in results, data formats, and descriptions among biomedical research discoveries. The Harvard Stem Cell Institute (HSCI) has created the Stem Cell Commons (stemcellcommons.org), an open, community-based approach to data sharing. Experimental information is integrated using the Investigation-Study-Assay tabular format (ISA-Tab) used by over 30 organizations (ISA Commons, isacommons.org). The early adoption of this format permitted the novel integration of three independent systems to facilitate stem cell data storage, exchange and analysis: the Blood Genomics Repository, the Stem Cell Discovery Engine, and the new Refinery platform that links the Galaxy analytical engine to data repositories.

Introduction

Progress in stem cell research is challenged by inconsistent experimental descriptions (e.g. cell types, developmental stages, surface markers, disease states) and difficulties integrating diverse data from gene expression, chromatin modification and transcription factor binding studies. The Stem Cell Commons (stemcellcommons.org) enables researchers to share projects and compare stem cell studies using ISA-Tab, a generic format for describing experiments, supported by an open source software suite enabling configuration, curation, validation, storage and search, with connections to existing analysis tools and formatting for submission to public repositories (isacommons.org). Using this common framework means that methods can be applied consistently, and that experiments can be matched across studies using common experimental descriptions.

Integration of data across repositories and with the Galaxy analytical platform driven by ISA-Tab

The Commons has been derived from two projects that implemented the ISA framework: Blood Genomics (bloodprogram.hsci.harvard.edu), a microarray resource for hematopoietic studies,and the Stem Cell Discovery Engine (SCDE, discovery.hsci.harvard.edu), a tissue and cancer stem cell resource linked to Galaxy. The consistent annotation of key data fields using structured vocabularies, anchored whenever available to the same semantic framework, in both databases was a major advantage that made it possible for 53 SCDE and 87 Blood Genomics studies to be integrated into a single, comprehensive stem cell resource. The resulting Drupal-based Commons repository contains 155 experiments with 2727 assays spanning multiple cell types, organisms, diseases and technologies. The challenge has been to manually curate each experiment to ensure consistent use of a predefined set of ontologies, and to establish a standard operating procedure for future curation activities, which can be performed using ISAcreator, a Java desktop application supporting the ISA-Tab format and ontology-based annotation.

The Commons team is actively developing capabilities for data manipulation, analysis and visualization. We have extended the Galaxy core API to allow us to control and automate analyses in Galaxy, and have developed Refinery, a Django-based visualization interface (refinery.med.harvard.edu). Experimental metadata in ISA-Tab format are exported from the Commons repository to Refinery. The metadata is parsed and stored in Refinery's internal database, and presented as search facets to guide users to choose samples of interest along with the proper workflows, parameters, and data sources to run and interpret analyses. Publicly available data from ArrayExpress, converted to ISA-Tab using the MAGE2ISAconverter, can also be imported. After analyses are complete, the ISA-Tab archive for that experiment will be augmented with the workflow information and parameters used, and saved back to the repository. This ongoing project demonstrates a novel, ISA-Tab driven integration of experimental metadata and raw data across stem cell data repositories, and with the Galaxy framework.