BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 18 2010, pages 2354–2356
doi:10.1093/bioinformatics/btq415

Databases and ontologies

Advance Access publication August 2, 2010

ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level

Philippe Rocca-Serra[1,2,†], Marco Brandizi[1,†], Eamonn Maguire[1,2,†], Nataliya Sklyar[1,†], Chris Taylor[1,3], Kimberly Begley[4], Dawn Field[3,5], Stephen Harris[6,‡], Winston Hide[4], Oliver Hofmann[4], Steffen Neumann[7], Peter Sterk[3,5], Weida Tong[6,‡] and Susanna-Assunta Sansone[1,2,*]

[1]The European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD and [2]Oxford e-Research Centre, University of Oxford, 7 Keble Road, Oxford OX1 3QG, [3]Natural Environment Research Council, Environmental Bioinformatics Centre, Wallingford CEH, Benson Lane, Mansfield Road, Oxford OX10 8BB, UK, [4]Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA, [5]Genomic Standards Consortium, Wellcome Trust Sanger Institute, Cambridge CB10 1SD, UK, [6]US Food and Drug Administration, Center for Bioinformatics, National Center for Toxicological Research, 3900 NCTR Road, Jefferson, AR 72079, USA and [7]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Summary: The first open source software suite for experimentalists and curators that (i) assists in the annotation and local management of experimental metadata from high-throughput studies employing one or a combination of omics and other technologies; (ii) empowers users to uptake community-defined checklists and ontologies; and (iii) facilitates submission to international public repositories.

Availability and Implementation: Software, documentation, case studies and implementations at http://www.isa-tools.org

Contact: isatools@googlegroups.com

Received on February 23, 2010; revised on July 7, 2010; accepted on July 8, 2010

1 HIGH-THROUGHPUT OMICS STUDIES

The development of high-throughput genomic and post-genomic (hereafter, 'omics') technologies entails changes in the handling, processing and sharing of data (Schofield *et al.,* 2009). Omics datasets are often complex and rich in context. Studies may run material through several kinds of assay, using both omics and other technologies; for example, studying the effect of a compound on rat liver through transcriptome, proteome and metabolome profiling (using high-throughput sequencing and two kinds of mass spectrometry, respectively) alongside conventional analyses (e.g. histopathology). Such data must be accompanied by enough contextual information (i.e. metadata; sample characteristics, technology and measurement types; instrument parameters and sample-to-data relationships) to make datasets comprehensible and reusable if they are to underpin future investigations.

Many funders and journals require that researchers share data, and encourage the enrichment and standardization of experimental metadata (Field *et al.,* 2009). Consequently, more and richer studies are flowing into public databases. However, two bottlenecks can significantly hamper this process, necessitating urgent solutions. First, international public repositories for 'omics data such as GEO (Barrett *et al.*, 2009), ArrayExpress (Parkinson *et al.*, 2009), PRIDE (Vizcaíno *et al*., 2010), ENA, SRA and DRA (Shumway *et al*., 2010), have their own submission formats, data models and terminologies, created for specific types of assay. This complicates the submission process for researchers producing multi-assay studies (and greatly increases the risk that these datasets become irrevocably fragmented). Secondly, the shortage of curators to check and annotate submissions to public repositories—a situation unlikely to change soon—necessitates better annotation at source (by experimentalists or community-based efforts; Howe *et al.*, 2008). Free software, with automated content validation, is required to facilitate the collection, management and curation of a variety of study inhouse, and to format those data for submission to public repositories. Such software should support community-defined reporting standards, such as the minimum information checklists listed by the MIBBI Portal (Taylor *et al.*, 2007), and ontologies, (Côté *et al*., 2006; Smith *et al.*, 2007; Noy *et al*., 2009).

The Investigation/Study/Assay (ISA) infrastructure described here is the first general-purpose format and freely available desktop software suite designed to regularize local management of experimental metadata by enabling curation at source, supporting community-defined reporting standards and preparing studies for submission to public repositories.

2 THE ISA FORMAT AND SOFTWARE SUITE

The software suite comprises five platform-independent Java-based software components for local use, including a relational database (Fig. 1), built around the ISA-Tab format. The components work

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

‡The views presented in this article do not necessarily reflect those of the US Food and Drug Administration

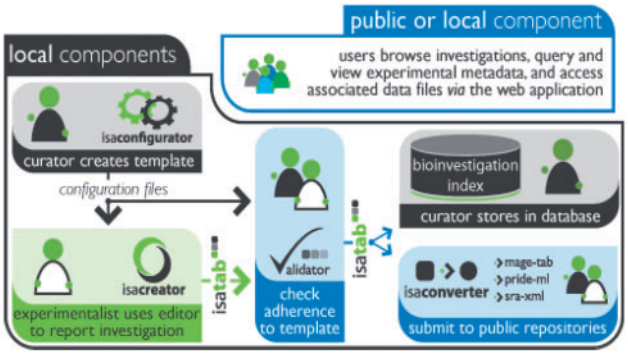

Fig. 1. The role of each ISA software component, showing their interrelations, target users and the flow of information through the system.

both as stand-alone applications and as a unified system to assist in the local management and storage of experimental metadata, and to facilitate data submission to international public repositories. All components run as 'desktop' applications; in addition, the database component features a web-based query interface.

2.1 ISA-Tab: an extensible, cross-domain format

'*Investigation*', '*Study*' and '*Assay*' are the three key entities around which the general-purpose ISA-Tab format for structuring and communicating metadata is built (Sansone *et al.,* 2008). *Investigation* contains all the information needed to understand the overall goals and means used in an experiment; *Study* is the central unit, containing information on the subject under study, its characteristics and any treatments applied. Each *Study* has associated *Assay(s)*, producing qualitative or quantitative data, defined by the type of measurement (i.e. gene expression) and the technology employed (i.e. high-throughput sequencing). The hierarchical structure of ISA-Tab enables the representation of studies employing one or a combination of omics and other technologies, overcoming the fragmentation of the existing submission formats built for specific types of assay. To ensure conversion, ISA-Tab has been designed with reference to these existing 'omics formats (Jones *et al.,* 2007), complementing and extending their work where necessary; for example, it shares both syntax and the use of easily-manipulable tab-delimited text files with ArrayExpress' MAGE-Tab (Rayner *et al.*, 2006). Additionally, where omics-based technologies are used in clinical or non-clinical studies, ISA-Tab complements existing biomedical formats such as the Study Data Tabulation Model (http://www.cdisc.org/sdtm), endorsed by the US Food and Drug Administration. ISA-Tab also complements the XML formats used by the PRIDE, ENA, SRA and DRA repositories, and consequently offers a way to render their experimental metadata documents in a more user-friendly format. Note though that ISA-Tab is simply a format; the decision on how to regulate its use (i.e. enforcing the filling of required fields, or the use of ontologies) is left to local administrators' use of ISA software components, or the growing number of other systems and groups implementing the format (e.g. Krestyaninova *et al.*, 2009; SysMO-DB http://www.sysmo-db.org/community; XperimentR, http://www.imperial.ac.uk/bioinfsupport/resources/data_management/; more given on the ISA web site).

2.2 ISAcreator: a user-friendly editor

This desktop application enables users (i.e. experimentalists) to compile experimental metadata sets, and to import and edit existing ISA-Tab formatted files. It breaks down overall descriptions into relatively simple parts, uses graphical abstraction to enable visualization of the information described and facilitates time-efficient description of experimental steps by remembering prior behaviour (through user profiles). ISAcreator's aesthetically pleasing interface makes extensive use of Java Swing and external open source libraries (e.g. Prefuse, http://prefuse.org/). The editor uses a style of form- and spreadsheet-based data entry that is likely to be familiar to researchers, augmenting basic functionality such as 'auto-fill' and 'undo' with advanced features, listed below.

2.2.1 Ontology support A dedicated 'widget' allows ontology terms to be searched for and inserted in real time *via* the BioPortal (Noy *et al.*, 2009) and the Ontology Lookup Service (Côté *et al.*, 2006). Terms from those sources are imported along with core metadata (identifiers, definitions and ontology version); term selection is facilitated by a search history displaying prior choices (through user profiles).

2.2.2 Design wizard An alternative way for users to enter information that leverages common patterns to reduce repetitive tasks by guiding users through a series of questions that elicit information about the design of the *Study* and associated *Assay(s)*.

2.2.3 Spreadsheet import As a second alternative, this widget enables the mapping and import of information from existing spreadsheets; also the reformatting and reannotation of *legacy* data.

2.2.4 Data file chooser This widget appends data files located either local to the operator, or identified by FTP on a remote system, to an experimental metadata sets. Upon completion of a valid investigation report, ISAcreator outputs a compressed 'ISArchive' containing the ISA-Tab-formatted metadata and either the actual data files, or a reference to them, if necessary (e.g. because of their large size), consisting of their address and file name.

2.3 ISAconfigurator: standards-compliant templates

This desktop application allows 'power users' (i.e. community curators) to customize the fields displayed by ISAcreator, and for example, to meet the requirements of one or more MIBBI minimum information checklists by declaring certain fields mandatory, or by specifying allowed values (e.g. drawn from a set of ontology terms, or formatted in a specific manner). Configuration files from ISAconfigurator are read by ISAcreator, which then generates interface components as required.

2.4 ISAvalidator: adherence to templates

This desktop application also reads configuration files and checks both that completed ISA-Tab files meet specified requirements and that associated data files have been linked. Whether ISA-Tab files are created with ISAcreator or another way (e.g. with spreadsheet software), ISAvalidator checks that the document is syntactically correct and internally consistent, and reports on errors (i.e. missing or incorrect values).

2.5 BioInvestigation Index: local storage

An ISArchive provides a simple way to store and share information in a structured manner, but those tasks are better performed by uploading such a file to an instance of our 'BioInvestigation Index' (BII), or another system that implements ISA-Tab import. The BII includes a management tool and relational database (tested with Oracle, MySQL and PostgreSQL). The former enables validation and loading of an ISArchive and provides simple permissions functionality to link users (or groups of users) to studies. The latter manages the storage of experimental metadata, which can be collectively searched and browsed *via* a query interface or web services; the destination for associated data files, and their protocol for transfer, is custom defined by the local administrator on installation. As an example, a publicly accessible instance of the BII, maintained by the European Bioinformatics Institute (http://www.ebi.ac.uk/bioinvindex), has proven useful as a curation and storage system for multi-assay studies, and as a mechanism for submitting data files to ArrayExpress, PRIDE, ENA and SRA. Installation of the BII system requires some knowledge of database management. However, it is portable enough to be easily installed in individual labs, to maximize the efficiency with which high-throughput studies can be managed and shared among users that have been granted access to them.

2.6 ISAconverter: submission to public repositories

ISAconverter recodes the relevant parts of ISArchives as MAGE-Tab, PRIDE XML or SRA-XML (used by ArrayExpress, PRIDE and ENA, SRA and DRA, respectively), enabling combined submission to public omics repositories. It is readily extensible to support export of other formats, e.g. SOFT required by GEO (Barrett *et al.*, 2009). Mappings for format elements are available in the ISA-Tab specification and documentation on the ISA web site.

3 COLLABORATIONS AND CASE STUDIES

Developed for the European multi-site 'CarcinoGENOMICS' project (Vinken *et al.*, 2008), the ISA software suite version one was released in early 2009. The core ISA developers are engaged with an ever-growing number of collaborators: case studies from early implementers already provide evidence of the diverse life science scenarios in which the suite's various components have been successfully tested and are being used with large datasets (details on the ISA web site). The main limitations recorded to date are simply the person hours required to specify the standards and ontologies to be used and to actually curate studies.

Demonstrable acceptance and community engagement has also brought a new funding stream for this project, allowing us to continue the collaborative development of this exemplar system that supports data sharing policies, promotes the uptake of community-defined reporting standards and ontologies and enables curation at source (Field *et al.,* 2009). The ISA components, in particular the BII, have been designed to provide core functionalities. Inevitably, each collaborator has additional in-house requirements that are too specific to be included as core functionality. This may be due to the nature of their studies or their need for one or more ISA software components to be interoperable with existing systems. To support further collaborative development, the core ISA developers are setting up an environment for distributed development, and are augmenting the ISA code base with Application Programming Interfaces (APIs). Ongoing collaborative activities include: a module to enable the analysis of ISA-Tab formatted metadata and any associated data, using R; integration with other data management and analysis systems (e.g. Fang *et al.*, 2009; MetWare, http://metware.org); and giving assistance to the growing number of projects exploring the tools and underlying format (e.g. Sage http://sagecongress.org/WP/workstreams/Standards; Kawaji *et al.*, 2009). Other collaborative activities include an enhanced user authentication system, support for additional formats such as RDF, OWL and SOFT, converters to/from lab equipment-related file formats (e.g. sampling robots and mass spectrometers) and improved packaging and distribution mechanisms to offer a single download bundle to facilitate installation.

ACKNOWLEDGEMENTS

The ISA developers owe debts of gratitude to many collaborators, as listed at: http://isatab.sf.net/people_funding.html.

Funding: CarcinoGENOMICS, NuGO, BBSRC (BB/I000917/1, BB/G000638/1, BB/E025080/1), NERC-NEBC and EMBL.

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, 885–890.

Côté,R.G. *et al.* (2006) The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.

Fang,H. *et al.* (2009) ArrayTrack: an FDA and public genomic tool. *Methods Mol. Biol.*, **563**, 379–398.

Field,D. *et al.* (2009) 'Omics Data Sharing. *Science*, **9**, 234–236.

Howe,D. *et al.* (2008) Big data: the future of biocuration. *Nature*, **4**, 47–50.

Jones,A.R. *et al.* (2007) The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat. Biotechnol.*, **25**, 1127–1133.

Kawaji,H. *et al.* (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol.*, **10**, R40.

Krestyaninova,M. *et al.* (2009) A System for Information Management in BioMedical Studies—SIMBioMS. *Bioinformatics*, **25**, 2768–2769.

Noy,N.F. *et al.* (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W173.

Parkinson,H. *et al.* (2009) ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, 868–872.

Rayner,T.F. *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.

Sansone,S.A. *et al.* (2008) The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?" *OMICS*, **12**, 143–149.

Schofield,P.N. *et al.* (2009) Post-publication sharing of data and tools. *Nature*, **10**, 171–173.

Shumway,M. *et al.* (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, 870–871.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255

Taylor,C.F. *et al.* (2007) MIBBI: a minimum information checklist resource. *Nat. Biotechnol.*, **26**, 889–896.

Vinken,M. *et al.* (2008) The carcinoGENOMICS project: critical selection of model compounds for the development of omics-based in vitro carcinogenicity screening assays. *Mutat. Res.*, **659**, 202–210.

Vizcaíno,J.A. *et al.* (2010) The Proteomics Identifications database: 2010 update. *Nucleic Acids Res.*, **38**, 736–742.