Building R objects from local ISA-tab files

Alejandra Gonzalez-Beltran, Steffen Neumann, Audrey Kauffmann,
Gabriella Rustici, ISA Team
<isatools@googlegroups.com>

August 1, 2012

1 ISA-tab format

The Investigation / Study / Assay (ISA) Tab-delimited (Tab) format is a general purpose framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from experiments employing a combination of technologies (http://isatab.sourceforge.net). In particular, ISA-Tab has been developed for - but not limited to - experiments using genomics, transcriptomics, proteomics or metabol/nomics techniques (the 'omics').

ISA-Tab uses three types of file to capture the experimental metadata:

- *Investigation file*
- *Study file*
- *Assay file* (with associated data files).

The Investigation file contains an overall description of an experiment while all experimental steps are described in the Study and in the Assay file(s). For each Investigation file there may be one or more Study files; for each Study file there may be one or more Assay files.

1.1 Investigation file

In this file, information is reported on a per-column basis and the fields are organized and divided in sections. The Investigation file is intended to meet three needs: (i) to define key entities, such as factors, protocols, parameters, which may be referenced in the other files; (ii) to relate Assay files to Study files; and optionally, (iii) to relate each Study file to an Investigation (when two or more Study files need to be grouped). The declarative sections cover general information such as contacts, protocols and equipment, and also - where applicable - the description of terminologies (controlled vocabularies or ontologies) and other annotation resources that were used.

1.2 Study file

In this file, information is structured on a per-row basis with the first row being used for column headers. The Study file contains contextualizing information for one or more assays, for example; the subjects studied; their source(s); the sampling methodology; their characteristics; and any treatments or manipulations performed to prepare the specimens.

1.3 Assay file

In this file, as for the Study file, fields are organized on a per-row basis with the first row being used for column headers. The Assay file represents a portion of the experimental graph (i.e., one part of the overall structure of the workflow); each Assay file must contain assays of the same type, defined by the type of measurement (i.e. gene expression) and the technology employed (i.e. DNA microarray). Assay-related information includes protocols, additional information relating to the execution of those protocols and references to data files (whether raw or processed).

For easy transfer, ISA-Tab files and associated data files can be packaged into an ISArchive, using a standalone Java application named ISAcreator (http://isatab.sourceforge.net). In order to facilitate identification of ISA-Tab components in an ISArchive, specific extensions have been created as follows:

- *i_name.txt* for identifying the Investigation file
- *s_name.txt* for identifying Study file (s)
- *a_name.txt* for identifying Assay file (s)

where 'name' is the user-given name.

2 The Risa package

The Risa package is used to build R objects from an ISA archive or dataset. The output is a list of objects containing, for example, the investigation, studies and assays filenames, the contents of their files, the list of samples, among other things.

These objects can then be used by downstream Bioconductor packages for data analysis and visualization (i.e, xcms). The package includes the function processAssayXcmsSet that, for a specific mass spectrometry assay, builds an xcmsSet object.

2.1 Building an R object from local ISArchive

If you have your own ISA archive, you can use the function `isatab2bioc` to convert it into an R object. The arguments for the function `isatab2bioc` are:

- path the name of the directory containing ISAtab files. The default is the working directory.
- verbose a boolean indicating to show messages for the different steps, if TRUE, or not to show them, if FALSE

As an example, we can use the *faahKO* dataset, whose latest version contains an ISA dataset describing the experiment.

```
> library("Risa")
> library("faahKO")
> isaobject = isatab2bioc(find.package("faahKO"))
```

The object `isaobject` contains the following:

- path - the path of the ISA-Tab dataset,
- investigation.filename - the name of the Investigation file
- investigation.file - a data frame with the contents of the Investigation file
- study.identifiers - the list of study identifiers
- study.filenames - the names of the study files
- study.files - a list of data frames wiht the contents of the study files
- assay.filenames - the names of the assay files

- assay.filenames.per.study - the names of the assay files according to the study they belong to
- assay.files - a list of data frames with the contents of the assay files
- assay.files.per.study - a list of data frames with the contents of the assay files divided per study they belong to
- assay.technology.types - a list with the technology types corresponding to each assay
- data.filenames - a list with the names of the data files
- samples - a list with the names of the samples
- samples.per.assay.filename - the samples classified according to the assay filename they belong to
- assay.filenames.per.sample - the names of the assay files classified per sample name
- sample.to.rawdatafile
- sample.to.assayname
- rawdatafile.to.sample
- assayname.to.sample

Augmenting the ISA-Tab dataset after analysis

The Risa package also provides the functionality to augment the original ISA-Tab dataset with more information after analysis.

The function `addAssayMetadata` allows to modify the metadata in a particular assay file. The arguments are:

- isa An isatab object, as retrieved by the isatab2bioc function.
- assay.filename the filename of the assay file to be augmented/modified
- col.name the name of the column of the assay file to be modified
- values the values to be added to the column of the assay file: it could be a single value, and in this case the value is repeated across the column, or it could be a list of values (whose length must match the number of rows of the assay file)

Session Info

```
> toLatex(sessionInfo())
```

- R version 2.15.1 (2012-06-22), `x86_64-apple-darwin9.8.0`
- Locale: `C/en_US.UTF-8/C/C/C/C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Biobase 2.16.0, BiocGenerics 0.2.0, Rcpp 0.9.13, Risa 1.0.7, faahKO 1.2.10, mzR 1.2.2, xcms 1.32.0
- Loaded via a namespace (and not attached): codetools 0.2-8, tools 2.15.1