On target: A public repository for large-scale RNAi experiments

Many RNA interference (RNAi) screens are now performed to functionally annotate the mammalian genome. Implementing a public repository based on common data standards is essential to realize the full value of large-scale RNAi data sets.

Caroline E. Shamu, Stefan Wiemann and Michael Boutros

Genome-scale RNAi screening has been possible for more than a decade in model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*, and more recently in mammalian cells. Most publications describing RNAi screen data focus on results from a few RNAi reagents per screen, typically those that are well-characterized experimentally in support of the main biological message. Such articles rarely even provide even full 'hit lists' of reagents that scored positive in the primary screen, partly because no database of record has been established and widely accepted as a repository for large-scale mammalian RNAi data.

We envisage an open-access public repository that includes raw and annotated mammalian RNAi results, well-documented experimental and data analysis protocols, and the sequences of RNAi reagents. Data formats should be suitable for both small and large data sets. This would allow inclusion of information that validates specific RNAi reagents; critical data on the impact of RNAi on gene and protein expression are usually collected only on a smaller scale.

Such a repository would be a valuable community resource. It could be mined for information on phenotypes that result when particular genes are knocked down (which is helpful annotation for gene function) and for information about how specific RNAi reagents score across different assays. Information about reagent performance matched with sequence information is very useful for studying RNAi on-target and off-target mechanisms, and thus for improving design of RNAi reagents. Moreover, bioinformatics analysis of aggregated RNAi data sets may contribute important information to the overall functional annotation of genomes.

To develop this mammalian RNAi data repository, three important challenges must be addressed: first, there must be consensus on where the repository (or repositories) will be hosted and who will participate in its development; second, experimentalists and database experts should collaborate to create standard formats for reporting and annotating RNAi data sets and experimental protocols, similarly to standards that have been implemented in the gene expression community; third, funding agencies and journals must provide incentives for researchers to deposit data sets. This is necessary not only to begin populating databases, but also to promote community interest in developing and implementing robust data standards for RNAi experiments.

The organization hosting the repository must meet three key criteria: first, it should be perceived to be fair, neutral and not favour particular geographical regions or types of science; second, it must have bioinformatic expertise in maintaining databases over an extended period of time; third, it must have sufficient domain-specific knowledge that it can contribute to the development of metadata and data standards, and

implement those that are agreed. The National Centre for Biotechnology Information (NCBI) and the European Bioinformatics Institute are two organizations that fit this profile, but both would require a community mandate and stable funding for this mission. Alternatively, the database could be maintained by an extramural consortium such as the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Databank.

Several pilot projects are underway to provide a solid foundation of experience on which to develop a public mammalian RNAi repository. Repositories for RNAi data from model organisms — WormBase (*C. elegans*), DRSC, FLIGHT and GenomeRNAi (*Drosophila*) — are instructive case studies. Some mammalian RNAi data, but few large screen data sets, are stored in the NCBI PubChem Bioassay database, with RNAi reagents compiled in the NCBI Probe database. GenomeRNAi has been extended recently to include mammalian screens.

The creation of a common repository depends on standardized ontologies and data formats for simple upload, download and mining of data sets. For large-scale RNAi screens, it will be essential to capture information on the biological assay, reagents (including sequence information), data (as well as data formats) and standardized phenotype ontology. An effort that incorporates broad community input to develop data standards for RNAi experiments is essential. This will ensure that data deposition guidelines are sufficiently complete to allow the repository to be mined productively, but also 'minimal' enough so that formatting data for deposition is not unnecessarily burdensome.

Data standards and formats for describing large data sets are being developed simultaneously by many groups and are starting to converge. For example, the ISA (Investigation/Study/Assay) framework for reporting experimental metadata (Sansone *et al. Nat. Genet.*; 2012) is being developed with input from many different areas of biology. The MIARE (Minimum Information About an RNAi Experiment) and MIACA (Minimum Information about a Cellular Assay) guidelines, both of which are relevant for describing RNAi data, are recognized as part of the larger minimum information standards effort. Their further development, however, should be closely tied to that of repositories to be most effective.

In conclusion, the wealth of data accumulating from RNAi screens should not be hidden in supplementary information, but rather made accessible in public data repositories to contribute meaningfully to our understanding of biology. This will require the commitment of the scientific community, funding agencies and journals working together.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

Caroline E. Shamu is at the ICCB-Longwood Screening Facility, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115-5731, USA. Stefan Wiemann is at the Division of Molecular Genome Analysis and Genomics & Proteomics Core Facility, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany. Michael Boutros is at the Division of Signaling and Functional Genomics, German Cancer Research Center (DKFZ) and Department of Cell and Molecular Biology, Heidelberg University, Medical Faculty, Mannheim, D-69120 Heidelberg, Germany. e-mail: caroline_shamu@hms.harvard.edu, s.wiemann@dkfz.de or m.boutros@dkfz.de