DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Knocking down the obstacles to functional genomics data sharing

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation	Simpson, Kaylene J., and Jennifer A. Smith. 2017. "Knocking down the obstacles to functional genomics data sharing." Scientific Data 4 (1): 170019. doi:10.1038/sdata.2017.19. http://dx.doi.org/10.1038/sdata.2017.19.
Published Version	doi:10.1038/sdata.2017.19
Citable link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:32071962
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

SCIENTIFIC DATA

OPEN

Comment: Knocking down the obstacles to functional genomics data sharing

Kaylene J. Simpson[1,2] & Jennifer A. Smith[3]

Received: 23 December 2016
Accepted: 26 January 2017
Published: 1 March 2017

This week, *Scientific Data* published a collection of eight papers that describe datasets from high-throughput functional genomics screens, primarily utilizing RNA interference (RNAi). The publications explore host-pathogen dependencies, innate immune response, disease pathways, and cell morphology and motility at the genome-level. All data, including raw images from the high content screens, are publically available in PubChem BioAssay, figshare, Harvard Dataverse or the Image Data Resource (IDR). Detailed data descriptors enable use of these data for analysis algorithm design, machine learning, data comparisons, as well as generating new scientific hypotheses.

Over the past decade, the term 'functional genomics' has become synonymous with large, often genome-scale, interrogation of gene function in a highly systematic and unbiased manner, principally relying on laboratory automation and often coupled with quantitative high-content phenotypic imaging. Cell-based experiments are miniaturized to microplate format (96-, 384- or 1536-well) and optimized, when possible, using well-defined positive and negative controls. The goal is to develop biologically relevant, robust assays. These efforts are intensive in terms of the cost and time investment required to develop, optimize and conduct them, as well as to perform subsequent data analysis and validation experiments. This, however, has not diminished the interest in, and value of, RNA interference (RNAi) as a discovery tool to explore a broad range of biological questions.

Arrayed RNAi screens generate a substantial amount of data. In its simplest form, a genome-scale viability screen performed in duplicate that relies on a plate reader for assay readout generates more than 40,000 data points. At the other end of the spectrum, a full-genome high-content imaging screen utilizing multiple fluorescent channels and capturing a large number of cellular features will produce millions of data points. Extensive statistical analysis is required to interpret such datasets and, historically, the 'screen' portion of a scientific publication is relegated to a supplemental figure, with both the data and methodology behind it inadequately described, thus preventing data reuse. It is vital that these screen data and their accompanying methodologies be made available to the scientific community in a usable format. Thus far, several obstacles have hindered this. While genome-scale arrayed RNAi screens have been performed and published for over a decade, there have been relatively few public data repositories. GenomeRNAi, initially described in 2007 by Michael Boutros and colleagues at DKFZ, was a pioneering database in this field[1]. Yet, the majority of datasets have been acquired from publications and thus suffered from a lack of adequate author-provided descriptions and, at least for mammalian cell-based screens, frequently contained only a subset of the data. Several years ago the NCBI's PubChem BioAssay[2], an established repository for compound screening data, expanded to host RNAi screening data. Commercial vendors released their siRNA duplex sequence information, specified as substance identifiers, thus enabling scientists to deposit both raw and analyzed numerical data into the BioAssay standardized format portal. Even with this advance, however, reuse was hindered by the lack of a detailed description of how data were generated and analyzed. Data usability was also impeded by the lack of

[1]Victorian Centre for Functional Genomics, Peter MacCallum Cancer Centre, Melbourne 3000, Australia. [2]The Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville 3010, Australia. [3]ICCB-Longwood Screening Facility, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence and requests for materials should be addressed to K.J.S. (email: kaylene.simpson@petermac.org) or to J.A.S. (email: jennifer_smith@hms.harvard.edu).

access to the corresponding raw images, thus preventing re-analysis and extraction of additional parameters.

This week, *Scientific Data* launched a collection of eight papers describing high-throughput functional genomics screens, exploring diverse biological processes at the genome-level, from host-pathogen dependencies to cell morphology and motility (http://www.nature.com/sdata/collections/funcgenom). Each contribution represents a concrete example of progress in overcoming the above-mentioned obstacles. Screen data are openly available via public data repositories, and the associated data records are clearly described in each publication. Notably, three of the screens relied on high-content imaging[3–5] with all raw images made available in either figshare, Harvard Dataverse or the Image Data Resource (IDR; based on the Open Microscopy Environment[6]), demonstrating the feasibility of effectively describing and sharing data even in these challenging cases.

The publications within this collection represent recent functional genomics screening efforts and, for many, this is the first time they have been described. Four of the contributions describe siRNA[7–9] or miRNA mimic and inhibitor[10] screens with relatively straightforward plate reader-based outputs. The data are provided through PubChem BioAssay[2]. Even with one or two data points per well, the complexity of the screen designs and data analysis creates substantial variation that requires in depth description. The primary screening protocols, raw and analyzed data, as well as data analysis methodologies and hit prioritization are well described. These publications delineate secondary and tertiary screens that were conducted to confirm positives, eliminate off-target effects and determine specificity. The use of multiple siRNA reagents in the primary and secondary screens conducted in Iain Fraser's laboratory[7,8] are informative, as scientists now have a variety of strategies to address off-target effects in RNAi screens, and these large-scale siRNA duplex comparisons are essential in designing and testing analysis algorithms to predict off-target effects[11–13]. Validation also extended to additional, complementary genomics approaches. For example, Wu *et al.*[9] present a valuable comparison of results from RNAi-mediated knockdown and CRISPR/Cas9-mediated knockout, and Sun *et al.*[8] include a comparison to transcriptomics data.

Arrayed, high-throughput screening is particularly amenable to quantitative high-content imaging, but, as described above, data-sharing is more challenging as a consequence of the vast amount of data, size of raw images, variables in image acquisition and analysis, as well as the multiparametric nature of the data. This collection includes three examples of such screens, each focusing on distinct cellular phenotypes. The authors have addressed what has historically been lacking with publications of these kinds of screens—inclusion of all raw image files and a clear explanation of the entire screening protocol. Vargas and colleagues knocked down the kinome in triple negative breast cancer cell lines, then deep-phenotyped the siRNA-transfected cells, quantifying 127 different features as well as YAP/TAZ localization[4]. A binning strategy was then used to sub-classify the features into 5 distinct shapes. All images are available via IDR. Vascular dynamics was the focus of the genome-wide siRNA screen performed by Williams *et al.*[3]. They utilized a scratch-wound healing assay to quantify cell motility in lymphatic vascular endothelial cells, with secondary screens expanding to blood endothelial cells; this high content screen enabled determination of core, conserved migration genes and cell line dependent targets. All images are available in figshare and the analyzed data are available in PubChem BioAssay. In contrast to the above two transfection-based screens, Ketteler and team relied on electroporation to deliver siRNAs targeting the human kinome into umbilical vein endothelial cells[5]. Image analysis quantified Weibel-Palade Body size, nuclei, the trans-Golgi network and plasma membrane staining. They developed a detailed analytical pipeline to integrate all cellular phenotypes. The raw images and data records are publically available at the Harvard Dataverse repository.

In contrast to the RNAi screens described above, Pettitt and colleagues conducted a forward-genetic screen[14]. They performed a genome-wide barcoded transposon screen to determine what mutations caused sensitivity to common cancer drugs in haploid murine embryonic stem cells. All scripts have been shared in GitHub and all FASTQ files are deposited in figshare, enabling re-use for a variety of applications.

Functional genomics datasets are a valuable resource to optimize new genomics tools, refine informatics approaches and improve machine learning. Re-analysis of screen data and raw images is an effective approach for hypothesis generation and identification of novel targets. While publications will no doubt continue to focus on a pathway or several targets identified in a large-scale screen, this requires extensive secondary and tertiary analysis, generally with different cell lines, biological assays, and validation with complementary approaches. The time investment is enormous. To share this abundance of data prior to or in conjunction with a more detailed publication focused on a specific aspect of the screen provides the research community with an opportunity to advance our knowledge, decrease duplicated efforts and conserve basic science resources.

References

1. Horn, T., Arziman, Z., Berger, J. & Boutros, M. GenomeRNAi: a database for cell-based RNAi phenotypes. *Nucleic Acids Res* **35,** D492–D497 (2007).
2. Wang, Y. *et al.* PubChem BioAssay: 2017 update. *Nucleic Acids Res* **45,** D955–D963 (2017).
3. Williams, S. P. *et al.* Systematic high-content genome-wide RNAi screens of endothelial cell migration and morphology. *Sci. Data* **4,** 170009 (2017).

4. Pascual-Vargas, P. RNAi screens for Rho GTPase regulators of cell shape and YAP/TAZ localisation in triple negative breast cancer. *Sci. Data* **4,** 170018 (2017).
5. Ketteler, R. Image-based siRNA screen to identify kinases regulating Weibel-Palade body size control using electroporation. *Sci. Data* **4,** 170022 (2017).
6. Goldberg, I. G. *et al.* The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol.* **6,** R47 (2005).
7. Li, N. *et al.* Genome-wide siRNA screen of genes regulating the LPS-induced NF-κB and TNF-α responses in mouse macrophages. *Sci. Data* **4,** 170008 (2017).
8. Sun, J., Katz, S., Dutta, B., Wang, Z. & Fraser, I. D. C. Genome-wide siRNA screen of genes regulating the LPS-induced TNF-α response in human macrophages. *Sci. Data* **4,** 170007 (2017).
9. Wu, W., Orr-Burks, N. L. & Tripp, R. A. Development of improved vaccine cell lines against rotavirus. *Sci. Data* **4,** 170021 (2017).
10. Orr-Burks, N. L. MicroRNA screening identifies miR-134 as a regulator of poliovirus and enterovirus 71 infection. *Sci. Data* **4,** 170023 (2017).
11. Yilmazel, B. *et al.* Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis. *BMC Bioinformatics* **15,** 192 (2014).
12. Buehler, E., Chen, Y. C. & Martin, S. C911: a bench-level control for sequence specific siRNA off-target effects. *PLoS ONE* **7,** e51942 (2012).
13. Buehler, E. *et al.* siRNA off-target effects in genome-wide screens identify signaling pathway members. *Sci. Rep* **2,** 428 (2012).
14. Pettitt, S. J. Genome-wide barcoded transposon screen for cancer drug sensitivity in haploid mouse embryonic stem cells. *Sci. Data* **4,** 170020 (2017).

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Simpson, K. J. & Smith, J. A. Knocking down the obstacles to functional genomics data sharing. *Sci. Data* 4:170019 doi: 10.1038/sdata.2017.19 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0

© The Author(s) 2017