

Quantitative human cell encyclopedia

Brenda J. Andrews, A. J. Marian Walhout,* Ravi Iyengar,* The Human Cells Project Working Group[†]

Scientists gathered to discuss the necessity, feasibility, and challenges of generating a quantitative catalog of the components in human cells that is essential for our understanding of human physiology in health and disease and to support future breakthroughs in treating diseases. This report summarizes the discussion that emerged at the Human Quantitative Dynamics Workshop held in Bethesda, MD, USA, in December 2015.

Going from genotype to phenotype requires a full understanding of cell and organ function. Reactions between molecular components of cells drive fundamental processes, such as growth, differentiation, and homeostasis. We need comprehensive quantitative data about the molecular components of cells, in terms of their concentrations and kinetic interactions, to connect genomic and epigenomic characteristics to cellular, tissue, and organ functions. Quantitative molecular data can produce transformative knowledge that revolutionizes our understanding of human physiology. A workshop supported by the National Institute of General Medical Sciences (1) brought together experts in experimental high-throughput technologies, informatics, and modeling to discuss how to obtain a quantitative catalog of human cells.

The grand vision is to catalog the absolute amounts of all cellular constituents and their rates of reactions in all cell types in the human body and then determine how these components and reactions change in healthy and disease states. Such a catalog will enable computable models of various cell types, which serve as a step for building precise predictive models of organ physiology. With these predictive models, we can understand how molecular changes drive physiology from healthy to disease states and the effects of drug therapy. A comprehensive catalog of all human cell types may not yet be achievable; however, we could start with the catalog with detailed information of a few cell types, like beating cardiac myocytes or insulin-secreting pancreatic cells. These cells have already been extensively studied, and there is sufficient previous knowledge to use these as the models for developing plans for a comprehensive cataloging project. Furthermore, the ability to create these cell types in functionally active forms from inducible pluripotent stem cells (iPSCs) (2) enables their large-scale production that is necessary for the success of such a cataloging study. Because iPSCs are obtained from individual human subjects, cells from the same individual can also be used for genomic, epigenomic, transcriptomic, proteomic, and metabolomic characterization, thereby connecting genomic and epigenomic data to the biochemical molecular data in a precise and reproducible manner.

For the catalog, data on proteins, metabolites (such as nucleotides, lipids, and sugars), and ions are needed. Because the proteins within a cell define the cell's characteristic features, the identification and quantification of all proteins in a cell are essential. A starting point is to exhaustively identify and quantify the mRNAs in the selected cell types. These measurements can be integrated with data in other mRNA databases, such as GTEx (Genotype-Tissue Expression) (3) and ENCODE (Encyclopedia of DNA Elements) (4), and serve as a basis for assessing the correlation between mRNA and protein abundance within the cells.

Not only are the protein composition and abundance in a cell critical for the catalog but also the protein-protein interaction data, posttrans-

lational modification information, and protein complexes data. Advances in protein mass spectrometry enable the identification and quantitative measurement of most proteins in a cell type (5), the quantification of post-translational modifications, and the assessment of protein complexes. Protein-protein interaction data and the kinetic parameters associated with all interactions in the cells of interest are needed, and the proteome-scale project on protein-protein interactions for human proteins makes this feasible (6). Protein-protein interactions often involve discrete regions (domains) within proteins; hence, current methods to quantify domain interactions (7) can be leveraged to obtain steady-state affinity measurements and forward and reverse rate constants. Although technology development and adaptation are needed to obtain proteome-wide kinetic interaction data, the basic methodologies are available and can be readily extended. Mass spectrometry and antibody-based methods can also be used to assess posttranslational modifications, as well as the composition and stoichiometry of protein complexes (8). Measuring of enzyme activities in a high-throughput manner in which activities of multiple enzymes are simultaneously measured remains a challenge; however, the use of stable isotope-labeled substrates in metabolomic analyses combined with appropriate informatics enables the quantification of enzyme activities, the estimation of kinetic parameters, and the dynamics of metabolic pathways (9).

There are numerous classes of metabolites. Cataloging and quantifying all of them may not be feasible initially. Identifying and quantifying metabolites involved in key cellular functions in the cell type of interest is a logical start. LIPID Metabolites and Pathways Strategy and the National Institutes of Health Common Fund Metabolomics Projects have made great strides in quantifying and identifying metabolites by mass spectrometry and nuclear magnetic resonance; these data can be leveraged for the quantitative catalog. Correlating the metabolite data with the measurements of enzyme activities can produce a comprehensive picture of the dynamics of biochemical pathways. Data on the interactions of metabolites with proteins are required. Protein-lipid interaction studies (10) can inform technologies for high-throughput data gathering for interactions of proteins with other small molecules. Reagents for quantifying important ions are available and can be used to gather this information.

The informatics challenges for a quantitative catalog are many. Each data type needs to be cataloged in a consistent manner, and there has to be interoperability across data types. For this, a "smart" database that learns and flags inconsistencies during data input and integration is needed. The data could be organized modularly as pathways and networks, because such modules can be used in graph theory modeling and in statistical and dynamical models to perform predictive simulations of cell functions. A successful catalog will have fully organized and functionally annotated lists of the biochemical and biophysical properties of all cellular proteins and components and their participation in reactions to produce cell-level functions, such as action potentials and contractility in myocytes or glucose-regulated insulin secretion in pancreatic cells. Such reactions can be readily

*Corresponding author. Email: ravi.iyengar@mssm.edu (R.I.); marian.walhout@umassmed.edu (A.J.M.W.)

[†]Members of The Human Cells Project Working Group are listed in Appendix 1.

assembled into pathways and subnetworks and can be used as modules to build larger-scale, well-constrained dynamical models that, when integrated with genomic and epigenomic information, will enable computer-based algorithms for precision medicine by coupling clinical phenotyping to cellular mechanisms. For long-term sustainability, it is paramount that this catalog use existing data repositories when possible.

The consensus at the workshop was that there are sufficient individual technologies to serve as a base for the Quantitative Encyclopedia of Human Cell Components for a selected set of human cell types. Model organisms, such as yeasts or worms, could serve as catalysts for the development of new technologies. The workshop involved a limited number of participants, and because a project of this scale may require a global endeavor, greater outreach and engagement of various research communities will be important.

Appendix 1

Brenda J. Andrews, Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 3E1, Canada.

Rolf Apweiler, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

Kristin Ardlie, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

Evren U. Azeloglu, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

Marc R. Birtwistle, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

Joshua J. Coon, Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706, USA.

Kara Dolinski, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.

Teresa Fan, Department of Toxicology and Cancer Biology, University of Kentucky, Lexington, KY 40536, USA.

Garret A. FitzGerald, Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA 19104, USA.

Anne-Claude Gavin, European Molecular Biology Laboratory, EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.

Anne-Claude Gingras, Lunenfeld-Tanenbaum Research Institute at Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada.

Nancy R. Gough, *Science Signaling*, American Association for the Advancement of Science, Washington, DC 20005, USA.

Alexander Hoffmann, Department of Microbiology, University of California, Los Angeles, Los Angeles, CA 90095, USA.

Ravi Iyengar, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

Michael J. Lee, Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA.

Leslie M. Loew, Department of Cell Biology, University of Connecticut, Farmington, CT 06030, USA.

H. Craig Mak, Cell Systems, Cell Press, Cambridge, MA 02139, USA.

Robert C. Murphy, Department of Pharmacology, University of Colorado Denver, Aurora, CO 80045, USA.

Chad Myers, Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, MN, USA.

Michael P. Snyder, Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.

Peter K. Sorger, Department of Systems Biology, Harvard Medical School, Cambridge, MA 02115, USA.

Gustavo Stolovitzky, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA.

Shankar Subramaniam, Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093, USA.

Mikko Taipale, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada.

Gilles Travé, Centre national de la recherche scientifique (CNRS), 1, rue Laurent Fries, BP 10142, F-67404 Illkirch Cedex, France.

Olga G. Troyanskaya, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 07412, USA.

Mathias Uhlen, Science for Life Laboratory, KTH-Royal Institute of Technology, Stockholm 106 91, Sweden.

Marc Vidal, Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.

A. J. Marian Walhout, Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA.

REFERENCES AND NOTES

- Human Quantitative Dynamics Workshop, www.sbcny.org/2015_workshop/index.htm.
- S. Yamanaka, Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell* **10**, 678–684 (2012).
- M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segré, S. Djebali, A. Niarchou; GTEx Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigó, Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
- ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- T. W. Rhoads, A. Prasad, N. W. Kwiecien, A. E. Merrill, K. Zawack, M. S. Westphall, F. C. Schroeder, J. Kimble, J. J. Coon, NeuCode labeling in nematodes: Proteomic and phosphoproteomic impact of ascaroside treatment in *Caenorhabditis elegans*. *Mol. Cell. Proteomics* **14**, 2922–2935 (2015).
- T. Rolland, M. Taşan, B. Charlotiaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruysinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejada, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, M. Vidal, A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
- R. Vincentelli, K. Luck, J. Poirson, J. Polanowska, J. Abdat, M. Blémont, J. Turchetto, F. Iv, K. Ricquier, M.-L. Straub, A. Forster, P. Cassonnet, J.-P. Borg, Y. Jacob, M. Masson, Y. Nominé, J. Reboul, N. Wolff, S. Charbonnier, G. Travé, Quantifying domain-ligand affinities and specificities by high-throughput holdup assay. *Nat. Methods* **12**, 787–793 (2015).
- A.-C. Gingras, M. Gstaiger, B. Raught, R. Aebersold, Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8**, 645–654 (2007).
- T. W.-M. Fan, P. K. Lorkiewicz, K. Sellers, H. N. B. Moseley, R. M. Higashi, A. N. Lane, Stable isotope-resolved metabolomics and applications for drug development. *Pharmacol. Ther.* **133**, 366–391 (2012).
- A.-E. Saliba, I. Vonkova, A.-C. Gavin, The systematic analysis of protein–lipid interactions comes of age. *Nat. Rev. Mol. Cell Biol.* **16**, 753–761 (2015).

Funding: This report is based on a workshop that was supported in part by a National Institutes of Health conference grant (R13116481) and a Systems Biology Center grant (P50-071558) from the National Institute of General Medical Sciences and by funds from the Office of the Dean, Icahn School of Medicine at Mount Sinai.

Submitted 27 June 2016

Accepted 19 August 2016

Final Publication 30 August 2016

10.1126/scisignal.aah4406

Citation: B. J. Andrews, A. J. M. Walhout, R. Iyengar, The Human Cells Project Working Group, Quantitative human cell encyclopedia. *Sci. Signal.* **9**, mr1 (2016).

Quantitative human cell encyclopedia

Brenda J. Andrews, A. J. Marian Walhout, Ravi Iyengar and The Human Cells Project Working Group

Sci. Signal. **9** (443), mr1.
DOI: 10.1126/scisignal.aah4406

ARTICLE TOOLS	http://stke.sciencemag.org/content/9/443/mr1
RELATED CONTENT	http://stke.sciencemag.org/content/sigtrans/8/371/fs8.full http://stke.sciencemag.org/content/sigtrans/8/408/eg6.full http://stke.sciencemag.org/content/sigtrans/2/64/eg3.full http://science.sciencemag.org/content/sci/356/6340/806.full http://science.sciencemag.org/content/sci/356/6340/eaal3321.full http://stke.sciencemag.org/content/sigtrans/10/481/eaan8309.full http://stke.sciencemag.org/content/sigtrans/10/481/eaan8308.full
REFERENCES	This article cites 9 articles, 2 of which you can access for free http://stke.sciencemag.org/content/9/443/mr1#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the [Terms of Service](#)

Science Signaling (ISSN 1937-9145) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Signaling* is a registered trademark of AAAS.

Copyright © 2016, American Association for the Advancement of Science