Marc Vidal, Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA and Department of Genetics, Harvard Medical School, Boston, MA 02215, USA. Email: marc_vidal@dfci.harvard.edu

Citation:
M. Vidal, How much of the human protein interactome remains to be mapped? *Sci. Signal.* **9**, eg7 (2016).

10.1126/scisignal.aaf6030

How much of the human protein interactome remains to be mapped?

One of the greatest challenges in biology is to mechanistically explain and predict genotype-phenotype relationships. Substantial progress has been made in the quest to identify the genes and mutations that underlie human disease. Unfortunately, we lack mechanistic information on how most mutations or genomic variants identified so far affect molecular functions.

An implicit assumption of classical 20th-century genetics was that simple, linear paths connect genotype to phenotype. The reality is that most genotype-phenotype relationships arise from much greater underlying complexity. This is certainly true for complex traits and also for Mendelian disorders, which are often complicated by phenomena such as incomplete penetrance.

Nonlinear connections between genotype and phenotype may arise because, rather than functioning in isolation, macromolecules are involved in intricate biophysical, biochemical, and functional interactions. These interactions form highly connected interactome networks that underlie complex cellular dynamic systems. It follows that, just as reference genome sequences revolutionized human genetics at the turn of the century, reference maps of interactome networks may be critical to functionalize and contextualize large numbers of genomic variants.

Between humans and various model organisms, roughly a dozen types of reference maps are at various stages of completion, including protein-protein and DNA-protein biophysical maps, as well as synthetic lethal and co-expression functional maps. What criteria should be used to define a reference interactome map? How do we address completeness and quality? How can we estimate the total number of biophysical interactions in a reference interactome map? Which aspects of the interactome network do we know already? Mapping the human binary protein-protein interactome illustrates how we could tackle these questions.

The network of the human binary protein-protein interactome is defined as all biophysical interactions that can be detected in a biologically relevant range of affinities between the proteins encoded by all 20,000 genes. From this static representation of the entire possible network, dynamic subsets of the interactome can be extracted that occur in specific cell types or tissues, at certain stages of development, or under particular environmental conditions. Ideally, an interactome reference map that attempts to represent this biological reality should (i) approach completeness for at least one splicing isoform of all protein-coding genes and (ii) contain only interactions that have been validated by orthogonal interaction assays.

With at least one splicing isoform as the minimum limit, the possible interactions that need to be assayed for the human proteome amount to ~200 million pairwise combinations. To assay this enormous search space, resources and strategies for systematic analyses—such as ORFeome collections (clones of protein-coding open reading frames) and interaction assays amenable to high-throughput settings [for example, the yeast two-hybrid (Y2H)]—are required.

To address quality, high-throughput results need to be validated using orthogonal methods, and benchmark data sets are also needed. These benchmark data sets are available and consist of (i) well-characterized interaction pairs that function as the positive reference set (PRS) and (ii) randomly picked protein pairs that function as the random reference set (RRS) (*1*). Any high-throughput interaction assay can then be adjusted to maximize the recovery of the PRS while minimizing the recovery of the RRS. The quality of any large-scale data set obtained with any assay can be assessed by comparing the recovery rate of a representative sample of its interactions against that of PRS and RRS pairs, using one or more other orthogonal assay(s).

Using this empirical framework, we can ask how much of the human binary protein-protein interactome network remains to be mapped. Computational assessments based on overlaps between radically different data sets have suggested that the human interactome may contain up to ~600,000 protein interactions. Furthermore, current databases containing literature-curated information report as many as ~200,000 interactions. These estimates are probably too high,

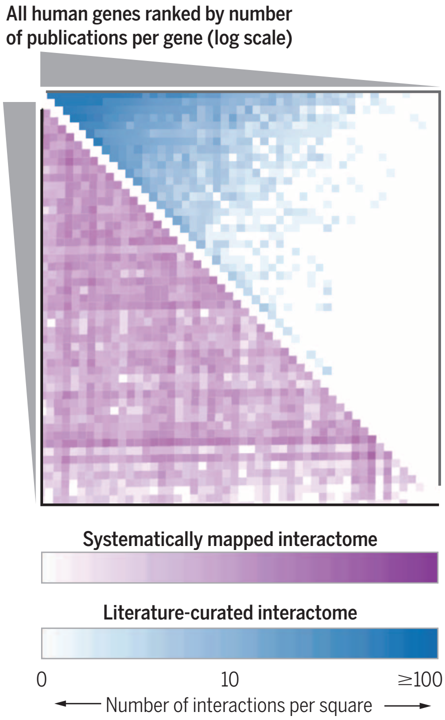

Fig. 1. Sampling the protein-protein interaction space Systematic proteome-wide analyses (purple) sample the space more completely than do targeted efforts in the published literature (blue).

because these numbers drop drastically when indirect co-complex associations are removed. Additionally, many reported binary interactions are supported by only a single piece of experimental evidence. Such "BS" (binary singleton) protein pairs perform very poorly in orthogonal assays (*1*, *2*) and should not be used for any serious attempt at characterizing properties of the human binary interactome.

As of 2014, the collective findings from three decades of published research amounted to ~11,000 high-quality interactions (*2*). However, these interactions from the literature are highly biased toward pairs of very "popular" proteins, representing only a very small fraction, or "dense zone," of the full interactome search space (Fig. 1, blue region; note the F1 use of a logarithmic scale to depict interaction density). Thus, a large fraction of the human interactome search space, or "sparse zone," has not yet been covered by small-scale efforts from individual laboratories. That same year, Rolland *et al*. published the second version of a systematic interactome map, based on screening approximately half of the search space and leading to ~14,000 high-quality interactions, which are publicly available (*2*). Based on the empirical framework, these 14,000 interactions represent ~10% of the full binary interactome (*1*, *2*). In marked contrast to literature-curated information, this systematic, high-quality, proteome-wide strategy covers the search space in a much more homogenous manner (Fig. 1, purple region). Whereas the vast majority of interactions from the systematic approach have not been previously reported (inside the literature-defined sparse zone), one-third of the systematic interactions had been found in the literature when the analysis is limited to the top 1% of the dense zone of the curated information.

These observations demonstrate that systematically generated and empirically controlled large-scale data sets constitute the best solution to mapping interactome networks and will eventually be instrumental for leveraging the genomic revolution. As a result of screening a nearly complete search space and using alternative versions of the Y2H, the systematic map has been expanded by approximately a factor of 3 to ~45,000 interactions. With the development of additional interaction assays and assuming sufficient funding, a reference map of the human binary protein-protein interactome is within reach by the end of this decade.

–Marc Vidal

REFERENCES

1. K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabási, M. Vidal, An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
2. T. Rolland, M. Taşan, B. Charloteaux, 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. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, 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).

CREDIT: P. HUEY/*SCIENCE SIGNALING*

Science Signaling

AAAS

How much of the human protein interactome remains to be mapped?
Marc Vidal (May 10, 2016)
Science Signaling **9** (427), eg7. [doi: 10.1126/scisignal.aaf6030]

Science Signaling (ISSN 1937-9145) is published weekly, except the last December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.