Published in final edited form as:

Nat Methods. 2009 October ; 6(10): 721–723.

From information to knowledge: new technologies for defining gene function

Sean R Collins¹, Jonathan S Weissman^{2,3,4}, and Nevan J Krogan^{2,4}

Jonathan S Weissman: weissman@cmp.ucsf.edu

¹Chemical and Systems Biology, Bio-X Program, Stanford University, Stanford, California, USA.

²The California Institute for Quantitative Biomedical Research, University of California, San Francisco, San Francisco, California, USA.

³Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, California, USA.

⁴Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, California, USA.

Abstract

A wide range of methodology will be needed to bridge the gap between genome sequence and mechanistic understanding in biology. Recent advances in high-throughput genetic screening address this task.

Breakthroughs in DNA sequencing technology now make the sequencing of essentially any organism's genome fast and inexpensive. Furthermore, new technology and ongoing research promise a deluge of information on the identities and frequencies of splice variants, as well as on the temporal and spatial patterns of gene expression. It remains a major task, however, to define the functional roles corresponding to these cellular and organismal parts lists, and to refine these parts lists into more conceptually tractable subsets likely to be the driving forces behind particular biological behaviors and diseases.

Given the vast array of functions carried out by different proteins, it is unrealistic that a single approach will allow the comprehensive definition of gene functions with the relative ease and simplicity with which DNA sequencing can define an organism's genome. Nonetheless, a series of strategies based on the generalization and systemization of classical genetics are now emerging as powerful tools for overcoming the information bottleneck between sequence and activity (Fig. 1). The common theme of these approaches is that the activity of each gene in the genome (alone or in combination with other genes) is perturbed—for example, by deletion, overexpression, treatment with chemical inhibitors or RNA interference—and the phenotypic consequence is measured. The challenge is to measure phenotypes with enough precision and depth to allow comprehensive definition of gene function in the context of automated and rapid data collection.

Two complementary approaches for determining complex, quantitative phenotypes have been used with great success. In the first approach, a 'high content' screen is conducted in which many different parameters are simultaneously monitored. In the second approach, a

single or limited number of parameters (for example, cell doubling time) are followed, but the effect of perturbing each gene is monitored in combination with a second perturbation, either another mutation or a chemical treatment. The resulting genetic interaction profile then provides a high-resolution view of the function of each gene.

As discussed below, the last few years have brought an impressive and varied array of new screening tools and techniques, including many approaches published in *Nature Methods*, that now make both high-content and genetic-interaction screens feasible in many different single-celled and multicellular models.

New mutant collections

At the foundation of any genetic screen is the systematic perturbation of the function or expression levels of a comprehensive set of genes. In the baker's yeast *Saccharomyces cerevisiae*, the creation of a complete collection of deletion strains for nonessential genes¹ facilitated remarkable progress in defining gene function. However, comparable tools for essential genes had lagged behind, until two recent efforts created collections of constitutive hypomorphic alleles of essential genes in this organism^{2,3}. Most of these strains were found to have modest growth defects in rich media, but certain conditions, such as treatment with known chemical compounds, revealed strong sensitivity, even beyond that of corresponding heterozygous deletion strains. Such collections therefore allow genetic interrogation of essential genes even when a 50% reduction of their dose does not elicit a phenotype.

A natural complement to loss-of-function screening is a systematic method for gene overexpression. Though plasmid libraries for overexpression screens in yeast have existed for some time, these libraries have had significant drawbacks. For example, the proteins from these libraries are usually strongly overexpressed, sometimes 1,000-fold, and often have affinity tags that can interfere with function. For these reasons, the Prelich group constructed a plasmid collection specifically designed for overexpression screening⁴. The construction strategy kept genes untagged and expressed from their native promoters, to avoid perturbing protein function and to achieve a more consistent degree of overexpression (relative to the natural abundance) across the collection. A minimal collection spanning the yeast genome was also rationally selected for use in arrayed screens.

Pooled screens

A main aim for systematic screening approaches is to maximize throughput, while minimally compromising (or perhaps even enhancing) the accuracy of phenotype detection. In some cases, assay throughput has been a limiting factor for obtaining genome saturation, and in all cases increased throughput translates to faster interrogation of the vast space of interesting phenotypes.

The yeast gene deletion collection was constructed with unique molecular barcodes genetically linked to each disrupted locus¹. This clever strategy allows pooled experiments, in which the strain collection is mixed together and then exposed to an environmental or genetic challenge. Extraction and PCR amplification of the barcodes then allows quantitative deconvolution of the relative enrichment or depletion of each strain within the pool. The pooling approach provides exceptional uniformity in experimental conditions, and it can dramatically simplify the set of manipulations required of the experimenter, thus increasing reproducibility and throughput. To make this approach widely applicable, Yan *et al.* created a library of yeast strains that is capable of introducing unique barcodes into almost any collection of mutant strains by means of an ordered mating and selection strategy³.

The success of pooled approaches in model organisms has inspired the development of analogous tools for mammalian cell culture systems. The Elledge and Hannon laboratories have been among the pioneers of these methods, developing an approach based on microarray deconvolution of molecular barcodes to identify small hairpin RNA constructs affecting the proliferation of human cell lines. As a proof of principle, they showed that analyzing effects of the same shRNA library on multiple cell lines in parallel can identify hairpins with effects specific for cancerous lines⁵.

More recently, Bassik *et al.* took advantage of deep sequencing to obviate the need for a distinct DNA barcode, thus greatly simplifying shRNA library construction⁶. This, in conjunction with advances in oligonucleotide synthesis, makes it possible to rapidly generate highly complex libraries containing 30 distinct shRNAs for each gene, thus maximizing the odds that multiple effective shRNAs will be present for each gene. This in turn helps alleviate both the high rate of false negatives in shRNA screens, which is due to the limited efficacy of most shRNAs, as well as the high rate of false positives due to off-target effects. Elaborations on this approach should serve both as valuable screening tools and as sources of data to improve future hairpin design algorithms.

The strategies described above take advantage of competitive growth or well established technologies (FACS) to separate cells with interesting, albeit simple, phenotypes (for example, increased expression of a GFP reporter or loss of a cell surface antigen) from the mass of uninteresting background. But new technology can extend the power of pooled mutant generation and organism manipulation to more complicated phenotypes, as demonstrated in two recent studies in worms.

Doitsidou *et al.* took advantage of the commercial COPAS Biosort system, which is a fluorescence-activated sorter capable of separating worms on the basis of fluorescence⁷. The authors used a cell type-specific promoter driving one fluorescent protein and a second, more broadly expressed fluorescent protein, to make rapid ratio-metric measurements. The worm sorter, like a standard flow cytometer, measures a fluorescence profile in only a single spatial dimension, yet real-time analysis of these profiles was capable of detecting worms with loss of a single dopaminergic neuron. This study promises an exciting future for the automated analysis of cell fate determination in worms.

Chung *et al.* developed their own micro-fluidic system for sorting worms, based on reporter expression phenotypes extracted from high-resolution images using real-time, customizable selection criteria⁸. Impressively, this system was capable of automatically detecting rare worms showing abnormal relative gene expression between two cells separated by a distance of only ~20 micrometers.

Developments in biomolecule introduction

Although pooled approaches can provide impressive throughput, the need for deconvolution after measurement limits the assays that can be used to those that allow physical separation of cells with a phenotype from the mass of unaffected cells. Automation of more complicated assays, designed to measure with high precision detailed intracellular phenotypes, offer an important avenue for focused screens on specific processes. The difficulty in efficiently and economically introducing nucleic acids or viruses into some cell types (for example, primary cells), and in combining this with high-resolution imaging and automated analysis, is a major challenge.

Guignet *et al.* engineered a 96-well gold-plated electroporation device aimed at attacking these challenges⁹. This device naturally interfaces with standard 96-well microplates that can then be used for automated microscopy. It was designed to use simple extracellular

buffer rather than expensive transfection reagents, and it works efficiently with difficult cell types including primary neurons. This device has already proven useful for focused screens monitoring complex cell behaviors such as endothelial sheet migration¹⁰.

An alternative strategy for pairing systematic biomolecule introduction with automated microscopy is the printing of ordered arrays of target molecules onto glass slides, as is done in the generation of conventional microarrays. In this case, a monolayer of cells is deposited on top of the arrayed molecules, and cells become transfected by taking up material from the solid phase beneath them. This technique, which was initially developed by the Sabatini group¹¹, has recently been extended in two important ways. Bailey *et al.* paired the printing technology with a lentivirus delivery system¹². The printed viruses are highly stable and capable of high-efficiency infection, and their broad tropism opens the cell microarray system to many new cell types (including primary cells). The system is also naturally compatible with RNAi-based loss-of-function techniques, as well as with transgene introduction and gene overexpression. Neumann *et al.* combined cell-based microarray transfection with live cell imaging and automated image analysis¹³. By using time-lapse microscopy after plating, the authors were able to distinguish early phenotypes from later ones and to distinguish primary phenotypes from secondary effects.

Genetic interaction strategies

Biological systems often use compensation mechanisms to maintain homeostasis. These mechanisms can mask the effects of individual mutations, posing a major obstacle for standard screening strategies. Genetic interaction studies, which analyze the combined effects of pairs of mutations, can overcome this issue and have provided a wealth of biological insight in a variety of systems.

Systematic characterization of genetic interactions was originally developed in budding yeast, first focusing on identifying synthetic lethal interactions^{14,15} and more recently on quantitative phenotypes using a variety of different methodologies and phenotypic readouts^{16,17}. Large-scale genetic interaction studies typically yield a genetic interaction profile, or a phenotypic signature, for each mutation. Comparison of these profiles is a powerful way to identify sets of genes that act in the same pathway.

Recent work has extended this approach to other single-celled organisms. Roguev and colleagues developed a high-throughput approach for creating double mutants in the fission yeast *Schizosaccharomyces pombe*¹⁸, in which many biological processes are more similar to those in mammalian cells than are their counterparts in *S. cerevisiae*. Furthermore, fission yeast, unlike *S. cerevisiae*, did not undergo a genome duplication event, and hence there is less overall functional redundancy within the *S. pombe* genome. Individual mutants in this organism, therefore, should provide stronger genetic profiles. This system in *S. pombe* was recently applied on a large scale, allowing an investigation of the conservation of genetic interactions and the ‘rewiring’ of functional modules between *S. pombe* and *S. cerevisiae*¹⁹.

Systems for high-throughput genetic interaction mapping of a prokaryotic organism were also recently described. Two groups independently devised approaches in the gram-negative bacterium *Escherichia coli*, based on a bacterial form of sexual reproduction, conjugation, that allows transfer of genetic material from a donor to a recipient^{20,21}. Because double mutants can be generated in *E. coli* much more quickly than in yeast, it should be possible to create massive genetic interaction maps rapidly. It will be of great interest to compare the pathways derived from unbiased genetic interaction mapping to the detailed bacterial biochemical pathways characterized over the past 60 years using more traditional methodologies. Extension of this approach to other organisms, especially gram-positive

bacteria such as *Bacillus subtilis*, will expand our knowledge of the genetic architecture of prokaryotic organisms and hopefully aid in the discovery of new types of drug targets.

Screening complex phenotypes

Automation of screening techniques for complex phenotypes such as behavior holds great potential for gene discovery—and for brightening the eyes of weary graduate students. However, systematic monitoring and quantification of behavior is a major challenge. Two recent studies in *Drosophila melanogaster* attacked this problem using sophisticated real-time imaging systems to track the behavior of flies^{22,23}. Branson *et al.* developed a video platform that can visually maintain the identities of individual flies for hours, in the context of a group²². Using only planar behavior and a simple fly body model, their analyses were able to predict gender and genotype. Dankert *et al.* used machine vision to automatically detect several actions that are part of fly courtship and aggression²³ and showed that the system could automatically discern previously identified effects of genetic and environmental conditions on behavior. For example, disruption of the activity of octopaminergic neurons (octopamine is closely related to noradrenaline) or mutation in males of the sex-specific transcription factor fruitless (*fru*) both result in reduced aggressive behavior. Neither of these approaches have as yet been used in medium- or high-throughput screens, and the results of such application should prove interesting.

Perspective

S. cerevisiae has served as the testing ground for the development and refinement of new approaches for functional genomics. Thanks to these efforts, the number of completely uncharacterized genes in yeast is dwindling rapidly and new functions for well characterized players are emerging. The tools described above are now enabling similar strategies to be applied in many different single-cell systems, including mammalian cells, as well as to intact model organisms. These efforts should make it possible to exploit the vast array of data that have emerged from large-scale genomic and proteomic projects to gain a deeper knowledge of the function and organizational principles of biological systems.

References

1. Winzler EA, et al. Science. 1999; 285:901–906. [PubMed: 10436161]
2. Breslow DK, et al. Nat. Methods. 2008; 5:711–718. [PubMed: 18622397]
3. Yan Z, et al. Nat. Methods. 2008; 5:719–725. [PubMed: 18622398]
4. Jones GM, et al. Nat. Methods. 2008; 5:239–241. [PubMed: 18246075]
5. Schlabach MR, et al. Science. 2008; 319:620–624. [PubMed: 18239126]
6. Bassik MC, et al. Nat. Methods. 2009; 6:443–445. [PubMed: 19448642]
7. Doitsidou M, et al. Nat. Methods. 2008; 5:869–872. [PubMed: 18758453]
8. Chung K, Crane MM, Lu H. Nat. Methods. 2008; 5:637–643. [PubMed: 18568029]
9. Guignet EG, Meyer T. Nat. Methods. 2008; 5:393–395. [PubMed: 18408727]
10. Vitorino P, Meyer T. Genes Dev. 2008; 22:3268–3281. [PubMed: 19056882]
11. Ziauddin J, Sabatini DM. Nature. 2001; 411:107–110. [PubMed: 11333987]
12. Bailey SN, et al. Nat. Methods. 2006; 3:117–122. [PubMed: 16432521]
13. Neumann B, et al. Nat. Methods. 2006; 3:385–390. [PubMed: 16628209]
14. Pan X, et al. Mol. Cell. 2004; 16:487–496. [PubMed: 15525520]
15. Tong AH, et al. Science. 2001; 294:2364–2368. [PubMed: 11743205]
16. Decourty L, et al. Proc. Natl. Acad. Sci. USA. 2008; 105:5821–5826. [PubMed: 18408161]
17. Schuldiner M, et al. Cell. 2005; 123:507–519. [PubMed: 16269340]
18. Roguev A, et al. Nat. Methods. 2007; 4:861–866. [PubMed: 17893680]

19. Roguev A, et al. *Science*. 2008; 322:405–410. [PubMed: 18818364]
20. Butland G, et al. *Nat. Methods*. 2008; 5:789–795. [PubMed: 18677321]
21. Typas A, et al. *Nat. Methods*. 2008; 5:781–787. [PubMed: 19160513]
22. Branson K, et al. *Nat. Methods*. 2009; 6:451–457. [PubMed: 19412169]
23. Dankert H, et al. *Nat. Methods*. 2009; 6:297–303. [PubMed: 19270697]

Figure 1.
Elucidating how genome sequence drives phenotype will require a range of methodologies in multiple organisms.