Opinion

Chemical genomics: what will it take and who gets to play?

Gavin MacBeath

Address: Center for Genomics Research, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.
E-mail: gavin_macbeath@harvard.edu

Published: 6 June 2001

Genome ***Biology*** 2001, **2(6)**:comment2005.1–2005.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/6/comment/2005

Abstract

Chemical genomics requires continued advances in combinatorial chemistry, protein biochemistry, miniaturization, automation, and global profiling technology. Although innovation in each of these areas can come from individual academic labs, it will require large, well-funded centers to integrate these components and freely distribute both data and reagents.

When it comes to genomics, the question that often arises is: who gets to play? Certainly the sequencing of the human genome left no doubt that extensive funding, be it public or private, is required for such efforts, and that factory-scale science can get the job done. But the past decade has also shown us that genomics is a field of innovation, as well as automation. As a result, genomics currently enjoys a place in academic labs, both large and small, as well as in industry. But will the same be true of chemical genomics?

What is chemical genomics?

As far as I can tell, two definitions for chemical genomics are currently in use. The first is: any study directed at gaining a holistic understanding of how small molecules interact with cells. Using this definition, we would include, for example, experiments in which drug treatment of cells has been studied using large-scale expression analysis [1-3], or large-scale protein analysis [4]. We would also include experiments in which many different, related cells (such as yeast cells each carrying a single gene knockout) are tested for changes in their sensitivities to various drugs [5,6]. If we use this definition, we see that chemical genomics is simply a subset of genomics in which the focus is on small molecules. Many of the same tools are used, and hence chemical genomics, by this definition, is just as accessible to the scientific community as are other types of 'genomic' study. In this article, I will therefore focus on the second definition of chemical genomics that is currently in use, and on how I think studies of this type will play out in both academia and industry.

The second definition goes something like this: just as genomics is the extension of genetics to a genome-wide scale, chemical genomics is the corresponding extension of chemical genetics. What, then, is chemical genetics? Broadly defined, it is the study of biological processes using small-molecule intervention, rather than genetic intervention. Just as genetics offers a way to study biology by modulating gene function through mutation, chemical genetics seeks to study biology by modulating protein function with small molecules. As a much-cited historical example, the drug colchicine was used in early studies of mitosis to identify the protein tubulin [7]. Over 30 years later, colchicine remains an invaluable weapon in the arsenal of cell biologists, as do other compounds, such as taxol, which target the same protein but with different biological effects.

The advantages and limitations of chemical genetics

Perhaps the greatest advantage afforded by small molecules is that they can induce their biological effects rapidly and often reversibly (just ask anyone who has induced gene expression with IPTG or arrested cell-cycle progression with hydroxyurea). While genetics provides analogous control through the use of conditional alleles, such as temperature-sensitive mutations, unwanted effects caused by the induction

itself (for example, inducing the heat-shock response) may obfuscate the data. Moreover, induction of conditional alleles is rarely possible in animal models (have you ever tried to heat-shock a mouse?).

In addition to this advantage, small molecules offer the capacity to study essential genes at any stage in development. If a gene knockout has an embryonic-lethal phenotype, the protein that the gene encodes can nevertheless be 'knocked out' in the adult animal by inhibiting its function with a small molecule. Moreover, multiple knockouts can be combined with ease, in contrast to the nightmare facing a geneticist who wants to see the effect of simultaneously deleting four different genes in a mutant mouse.

The downside of chemical genetics is, of course, that it is not as generally applicable as conventional genetics. Whereas any gene can, in principle, be activated or knocked out by genetic intervention, a chemical genetic study is limited by the availability of appropriate reagents. Although the list of useful compounds is quite large, it almost certainly doesn't include a ligand for the protein you would most like to study.

So where do the compounds on the available 'list' come from? Historically, they have been identified as bioactive natural products isolated from a variety of organisms, such as fungi, plants, and bacteria. Initially, discoveries were made in academic labs (consider, for example, the discovery of penicillin). Lately, however, the identification of new bioactive compounds has occurred predominantly within pharmaceutical companies (consider, by contrast, almost every new antibiotic made or identified in the last 10 years). Although academia continues to remain engaged in the identification of bioactive natural products, it is fair to say that discovery in this area is overwhelmingly dominated by big pharmaceutical companies. So, what of chemical genomics?

From chemical genetics to chemical genomics

To move the somewhat *ad hoc* field of chemical genetics to a genome-wide scale requires approaches that are general. (The ability to study something on a large scale and systematically, rather than one-by-one and on a case-by-case basis, is what gives you the right to append '-omics' to your field of study.) Given that small molecules substitute for mutations in chemical genetics, we need either general methods to design compounds that will modulate the function of any gene or protein in a cell, or methods to prepare and screen large collections of diverse compounds for biological activity.

The design approach

To scale up small-molecule design to a genome-wide scale requires methods that are general. Although structure-based design of ligands has made great strides in recent years [8,9], it is not, at this point, extensible to a genome-wide endeavor. This does not, however, preclude all so-called 'rational' methods of ligand design. A few noteworthy approaches have been described that can be generalized to some extent. For example, Dervan and coworkers [10] have described ways to design and synthesize cell-permeable polyamides that bind with high affinity and selectivity to any given sequence of double-stranded DNA. Such molecules have now been shown to affect the expression of genes controlled by transcription factors that bind at or near the target sequence [11]. These compounds should prove extremely valuable in studying gene expression in an inducible and reversible manner.

In an approach that focuses on the proteins themselves, Shokat and coworkers [12] have found a way very specifically to inhibit almost any protein kinase. A kinase of interest is mutated to introduce a cavity in its active site, and existing, broad-specificity chemical inhibitors are then modified to fit the new, carved-out enzyme. Such molecules no longer bind promiscuously to many different kinases; instead, they exhibit high specificity for their genetically altered target. Since the site where such a mutation must be introduced can be predicted from primary sequence alignments, the method can be generalized to almost any kinase.

The diversity-based approach

Although the examples described above demonstrate how rational approaches can be exploited to expand the use of small molecules in studying biology, the ultimate goal of chemical genomics is to produce one or more specific ligands for every single protein in a cell, tissue, or organism. For this formidable task, we can take our cue, once again, from the field of genetics. Just as classical genetic studies proceed by generating large collections of random mutants and then screening or selecting for a given phenotype, so chemical genetics can exploit the same strategy. And just as genomics has taken the techniques of genetics and applied them in a systematic and large-scale manner, so chemical genomics must do likewise. But what will it take to do this? At a minimum, it will require four essential components: small-molecule compounds; proteins; fast, efficient scalable methods; and profiling technology.

Compounds, compounds, compounds

In chemical genetics, small molecules are the equivalent of mutations. Just as a classical genetic study begins with the generation of a large set of mutants, a directed (as opposed to serendipitous) chemical genetic study must start with a large collection of compounds. And this is where the two fields differ. A single graduate student in a small academic lab can easily generate millions of mutations in a genetically tractable model organism, such as yeast. In contrast, it has taken hundreds of people several decades, spending millions of dollars in large pharmaceutical companies, to assemble on

the order of a million natural products, many of which are available in only limited amounts and are not yet completely characterized or even purified. So, it is little wonder that chemical genetics has not yet been generalized or that most discovery to date has occurred in industrial labs.

But that is now set to change. With the advent of combinatorial chemistry has come the ability to prepare large collections of diverse compounds synthetically, using relatively few chemical steps. The field began with an emphasis on synthesizing peptides and other relatively simple compounds, but it has expanded to include the preparation of more complex molecules. To illustrate, in 1998 Schreiber's lab [13] reported the stereoselective synthesis of over two million 'natural-product-like' compounds using split-pool combinatorial synthesis. While the chemistry took several years to develop, the synthesis itself was carried out by two graduate students over the course of two weeks, with most of their time devoted to encoding the beads for subsequent identification of the attached molecules. In two weeks, two chemists prepared more compounds than there are natural products on the shelves at any pharmaceutical company. To be fair, natural product collections almost certainly contain a much higher proportion of compounds that are likely to display bioactive properties, since the molecules were isolated from organisms that make these compounds for a reason (often as a defense against other organisms). Nevertheless, the fact remains that combinatorial chemistry provides a way for academics to re-enter the small-molecule game.

But what about labs that are not dedicated to synthetic chemistry? Can they enter this game, in which organic chemistry plays such a central part? Again, there is hope. Recent years have seen the formation of several commercial companies that sell libraries of compounds, derived either through their own combinatorial synthetic efforts (see, for example, [14]) or by collecting small molecules from a variety of sources (see, for example, [15]). Unfortunately, these libraries are typically very expensive, requiring a change in the way this type of science is organized and funded within the academic setting.

Proteins, proteins, proteins

Compounds are just one side of the story, however. They enable the generalization of chemical genetics, but to expand to a genome-wide scale requires more. It is not sufficient to screen many compounds against one or a few targets of interest; it must be done against a large collection of targets - and, at the limit, against the whole proteome. But where will all these proteins come from?

As most biochemists know from experience, expression and purification of one or a few proteins can sometimes prove to be an exasperating endeavor. To attempt this on a genome-wide scale sounds impossible. If we hope to pull this off, we will need to use every trick in the book. Some proteins will express well in *Escherichia coli*; others will require yeast, baculovirus, or even mammalian expression systems. Some proteins will tolerate amino-terminal tags (to allow them to be affinity purified); others will require carboxy-terminal tags. Some proteins will function normally with their affinity tags intact; others will require tag removal. Some proteins will emerge cleanly from a single affinity purification step; others will require multiple steps. For this reason, indexed collections of cDNA clones, both full-length and subdivided by the domain structure of the encoded protein, must be prepared, not in conventional cloning vectors, but in flexible, recombination-based systems that will facilitate future cloning and expression. In this way, sequence-verified clones can be transferred *en masse* into multiple expression systems using an automation-friendly subcloning step. Efforts of this sort are now underway in a number of labs (see, for example, [16,17]) and will ultimately provide an enormously valuable resource, not only for chemical genomics but also for a variety of applications aimed at the high-throughput study of protein function.

Smaller, faster, cheaper

The third essential component of chemical genomics is one that applies to all '-omic' studies: the need to miniaturize and automate. When you are dealing with only a few elements in an experiment - be they small molecules, proteins, nucleic acids, or others - scale and manpower are relatively unimportant issues. When thousands of elements are being considered, however, the mantra becomes 'smaller, faster, cheaper'. Conservation of resources and conservation of effort are essential.

I had the pleasure of observing and, to a small extent, participating in the evolution of screening at the Harvard Institute of Chemistry and Cell Biology (ICCB) [18]. This is an institute that is now about four years old and is directed at generalizing chemical genetics. When I arrived at ICCB as a postdoc in 1998, efforts were in their early stages and small-molecule libraries were being prepared on relatively small synthesis beads (90 μm diameter beads producing about 0.1 nmol of compound per bead). Cell-based assays were being designed in which the beads bearing the compounds were physically present throughout the screening process [19,20]. This meant that only a single assay could be run with a given bead. Although this strategy is perfectly acceptable for one or a few assays, it cannot realistically be scaled to a larger, and ultimately genome-wide, effort. The decision was therefore made to use larger beads that hold a thousand times more compound and to release the molecules from the beads into separate wells of microtiter plates, so as to generate stock solutions. Very small volumes of these stock solutions are then transferred robotically into separate assays, enabling hundreds or thousands of screens to be performed with the compound from each bead. Because minute amounts of each compound are used, the assays themselves

must also be miniaturized. Following the decision to move to this flexible assay format, many different microtiter-based screens have been designed and implemented (see Figure 1), including screens with purified proteins [21], whole cells [22], and even multicellular organisms [23].

In thinking about new technologies that can be applied to the advancement of chemical genomics, it is instructive to look at the contribution DNA microarrays have made to genomics. Whereas the northern blot first enabled the straightforward quantitation of transcript abundance, microarray technology extended this to a genome-wide scale because the assays were extremely miniaturized, making it possible to use very small sample volumes and reducing costs; assays were also multiplexed, enabling the rapid and simultaneous analysis of thousands of different transcripts; they were automated, enabling thousands of samples to be handled with minimal human intervention; and they were replicated, enabling hundreds of different analyses to be performed on the same set of transcripts. These same principles, in essentially the same format, can be exploited directly for small-molecule discovery. With compounds present as stock solutions in microtiter plate format, standard arraying robots can be used to deposit the compounds at extremely high spatial density onto chemically derivatized glass slides [24]. Because the compounds are generated by combinatorial synthesis, every compound in the library can be engineered to have a common reactive functional group. This group can then be used to direct the covalent attachment of each compound to appropriately

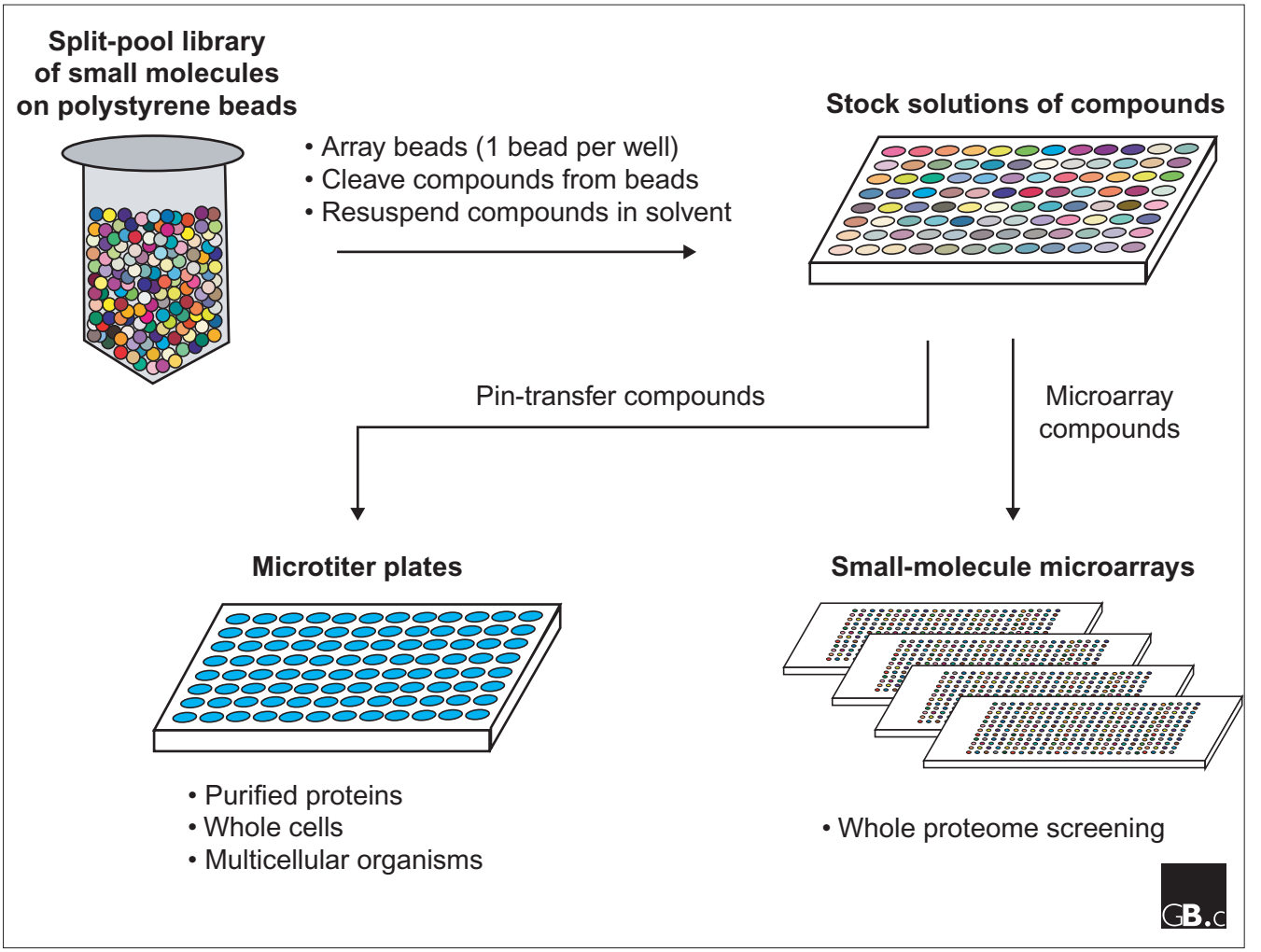

Figure 1
An overview of small-molecule screening, as carried out at the Harvard Institute of Chemistry and Cell Biology. Small-molecule libraries are prepared by split-pool synthesis on large polystyrene synthesis beads. The beads are then arrayed, one per well, into 384-well plates. Following cleavage from the beads, the compounds are resuspended in solvent and transferred into empty 384-well plates to yield stock solutions. At this point, minute amounts of the compounds can either be introduced robotically into microtiter plate assays or printed robotically onto chemically derivatized glass slides to produce small-molecule microarrays.

modified slides. In a microarray format, each slide can display over 10,000 different compounds, which can then be probed with fluorescently labeled proteins to identify new ligands.

The microarray format for this type of simple binding assay offers two important features. First, so little compound is used for each assay (about 1 nl of the stock solution) that 10,000 assays can be performed with the compound released from a single bead. Second, because the library of small molecules can be replicated onto thousands of different slides, large collections of proteins can be screened, one by one, against each of these compounds. This feature enables the extension of small-molecule screening to a proteome-wide endeavor. If we hope to make chemical genomics a reality, we need technologies of this sort, to enable libraries of compounds to be screened against libraries of proteins.

Profiling

Small-molecule ligands are not much use without information about their specificity. Does a compound bind only to its intended target or does it bind promiscuously to a broad range of related proteins? Although highly specific compounds are ultimately more useful than less specific ones, both can serve as valuable tools provided their binding properties are well understood. This highlights the importance of using global profiling technologies to study the compounds identified from large-scale screening efforts. At the moment, expression profiling is by far the best way to do this - although it is far from ideal. Of particular note are studies that compare the profiles of small-molecule-treated cells with otherwise isogenic cells in which the gene encoding the compound's target has been deleted [25]. This provides an indication of how well the small molecule 'phenocopies' the deletion of its target. In addition, comparative profiling of targetless cells, either treated or not treated with the corresponding small molecule, can be used to investigate that compound's 'off-target' effects.

In an ideal world, it is of greater value to study the cellular effects of small-molecule treatment at the protein level. Unfortunately, effective methods for protein profiling lag far behind those for expression profiling. Although comparative two-dimensional gel electrophoresis coupled with mass spectrometry currently provides the most comprehensive way to profile proteins, it falls far short of ideal. Much effort is now being directed towards affinity capture-based methods for protein profiling [26], and it will be very interesting to see how this field evolves over the next few years. In a related vein, we and others have described methods for preparing microarrays of functionally active proteins on solid supports [27-30], which may also prove useful for evaluating small-molecule specificity - in the short term with relatively small arrays of related proteins and in the longer term with comprehensive 'proteome arrays'.

So, who gets to play?

Genomics has been driven by technological innovation and the same will be true of chemical genomics. At some level, therefore, everyone gets to play. Isolated labs can contribute to advances in combinatorial chemistry, target identification, high-throughput cloning, protein production, assay design, automation, miniaturization, detection technology, profiling, and informatics. But to bring all these components together will require large, multidisciplinary, and well-funded centers that employ scientists from a broad range of disciplines. Pharmaceutical companies fit this description, but they are typically constrained in their research programs by market issues. Moreover, if the results of their efforts are not published, the data effectively don't exist as far as the greater scientific community is concerned. This leaves either privately funded centers that pursue and publish basic science, such as the Genomics Institute of the Novartis Research Foundation [31], or publicly funded academic institutes with the appropriate vision and infrastructure, such as the Harvard Institute of Chemistry and Cell Biology [18].

The past decade has shown us that the genomics revolution is not just a shift in the way we study biology: it is shift in the way we organize and fund science. Genome centers were required to sequence the human genome; analogous centers will be required for chemical genomics. And just as the public human genome project required organization, cooperation, and a commitment to open access, the same will be true for chemical genomics.

Acknowledgements

Work in my lab is supported by the Harvard Center for Genomics Research [32].

References

1. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
2. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
3. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, *et al.*: **A gene expression database for the molecular pharmacology of cancer.** *Nat Genet* 2000, **24:**236-244.
4. Moller A, Soldan M, Volker U, Maser E: **Two-dimensional gel electrophoresis: a powerful method to elucidate cellular responses to toxic compounds.** *Toxicology* 2001, **160:**129-138.
5. Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW: **Genomic profiling of drug sensitivities via induced haploinsufficiency.** *Nat Genet* 1999, **21:**278-283.
6. Chan TF, Carvalho J, Riles L, Zheng XF: **A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR).** *Proc Natl Acad Sci USA* 2000, **97:**13227-13232.
7. Shelanski ML, Taylor EW: **Isolation of a protein subunit from microtubules.** *J Cell Biol* 1967, **34:**549-554.
8. Klebe G: **Recent developments in structure-based drug design.** *J Mol Med* 2000, **78:**269-281.
9. Gane PJ, Dean PM: **Recent advances in structure-based rational drug design.** *Curr Opin Struct Biol* 2000, **10:**401-404.

10. Dervan PB, Burli RW: **Sequence-specific DNA recognition by polyamides.** *Curr Opin Chem Biol* 1999, **3:**688-693.
11. Chiang SY, Burli RW, Benz CC, Gawron L, Scott GK, Dervan PB, Beerman TA: **Targeting the ets binding site of the HER2/neu promoter with pyrrole-imidazole polyamides.** *J Biol Chem* 2000, **275:**24246-24254.
12. Bishop AC, Ubersax JA, Petsch DT, Matheos DP, Gray NS, Blethrow J, Shimizu E, Tsien JZ, Schultz PG, Rose MD, *et al.*: **A chemical switch for inhibitor-sensitive alleles of any protein kinase.** *Nature* 2000, **407:**395-401.
13. Tan DS, Foley MA, Shair MD, Schreiber SL: **Stereoselective synthesis of over two million compounds having structural features both reminiscent of natural products and compatible with miniaturized cell-based assays.** *J Am Chem Soc* 1998, **120:**8565-8566.
14. **BioSpace: Combinatorial Chemistry Companies** [http://www.biospace.com/articles/062199_combi_players.cfm]
15. **ChemBridge Corporation** [http://www.chembridge.com/]
16. **Harvard Institute of Proteomics** [http://www.hip.harvard.edu/]
17. **Vidal Laborartory** [http://vidal.dfci.harvard.edu/]
18. **Harvard Institute of Chemistry and Cell Biology** [http://sbweb.med.harvard.edu/~iccb/]
19. Borchardt A, Liberles SD, Biggar SR, Crabtree GR, Schreiber SL: **Small molecule-dependent genetic selection in stochastic nanodroplets as a means of detecting protein-ligand interactions on a large scale.** *Chem Biol* 1997, **4:**961-968.
20. You AJ, Jackman RJ, Whitesides GM, Schreiber SL: **A miniaturized arrayed assay format for detecting small molecule-protein interactions in cells.** *Chem Biol* 1997, **4:**969-975.
21. Degterev A, Lugovskoy A, Cardone M, Mulley B, Wagner G, Mitchison T, Yuan J: **Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-xL.** *Nat Cell Biol* 2001, **3:**173-182.
22. Mayer TU, Kapoor TM, Haggarty SJ, King RW, Schreiber SL, Mitchison TJ: **Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen.** *Science* 1999, **286:**971-974.
23. Peterson RT, Link BA, Dowling JE, Schreiber SL: **Small molecule developmental screens reveal the logic and timing of vertebrate development.** *Proc Natl Acad Sci USA* 2000, **97:**12965-12969.
24. MacBeath G, Koehler AN, Schreiber SL: **Printing small molecules as microarrays and detecting protein-ligand interactions en masse.** *J Am Chem Soc* 1999, **121:**7967-7968.
25. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, *et al.*: **Drug target validation and identification of secondary drug target effects using DNA microarrays.** *Nat Med* 1998, **4:**1293-1301.
26. Haab BB, Dunham MJ, Brown PO: **Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions.** *Genome Biol* 2001, **2**:research0004.1-0004.13.
27. Lueking A, Horn M, Eickhoff H, Bussow K, Lehrach H, Walter G: **Protein microarrays for gene expression and antibody screening.** *Anal Biochem* 1999, **270:**103-111.
28. Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V, Mirzabekov A: **Protein microchips: use for immunoassay and enzymatic reactions.** *Anal Biochem* 2000, **278:**123-131.
29. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination.** *Science* 2000, **289:**1760-1763.
30. Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M: **Analysis of yeast protein kinases using protein chips.** *Nat Genet* 2000, **26:**283-289.
31. **Genomics Institute of the Novartis Research Foundation** [http://www.gnf.org/]
32. **Harvard Center for Genomics Research** [http://www.cgr.harvard.edu/]