

UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Predicting the Mode of Action of Bioactive Compounds via High Throughput Screening and Computational Algorithms

Permalink

<https://escholarship.org/uc/item/9976p4ch>

Author

Woehrmann, Marcos H.

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

**PREDICTING THE MODE OF ACTION OF BIOACTIVE
COMPOUNDS VIA HIGH THROUGHPUT SCREENING
AND COMPUTATIONAL ALGORITHMS**

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Marcos H. Woehrmann

March 2015

The Dissertation of Marcos H. Woehrmann
is approved:

Professor Joshua M. Stuart, Chair

Professor Phil Berman

Professor R. Scott Lokey

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Copyright © by

Marcos H. Woehrmann

2015

Table of Contents

List of Figures	iv
List of Tables	vi
Abstract	vii
Dedication	ix
Acknowledgements	x
1 Introduction	1
1.1 Current Prescreening Techniques	9
1.2 Cytological Profiling.....	13
1.3 Screening in Yeast	17
1.3.1 D-Map	24
1.3.2 BioSpace.....	26
2 Cytological Profiling.....	28
3 Screening in Microbes.....	48
3.1 D-Map	50
3.2 BioSpace.....	69
4 Discussion.....	84
Appendices	86
Bibliography	119

List of Figures

1. Synthetic Lethal Interactions.....	20
2. Bliss Independence	22
3. Example knockout/drug interaction	23
4. Cytological profiling features	29
5. Traditional halo assay.....	49
6. Agar plate showing halos.....	51
7. D-Map method overview.....	52
8. Example D-Map plate.....	55
9. Heatmap of known compounds.....	59
10. D-Map correlation histograms	60
11. HALO384 overview	63
12. HALO384 dynamic range and linearity	64
13. HALO384 plate editing.....	66
14. HALO384 Z-factor	67
15. HALO384 plate compare.....	69
16. BioSpace fingerprint method overview.....	71
17. BioSpace direct prediction method overview	72
18. BioSpace screening knockout selection heatmap	75
19. BioSpace correlation vs. MACCS Tanimoto correlation	77

20. Comparison of Cytological Profiling, BioSpace, and MACCS ...	78
21. ECFP_4 correlation heatmap	79
22. Complete D-Map results heatmap	93

List of Tables

1. Significant D-Map classes.....	61
2. BioSpace Target Predictions.....	80
3. D-Map screening compounds.....	84
4. D-Map plate compound classes.....	90
5. D-Map plate compounds.....	91
6. BioSpace knockout screening set.....	118

Abstract

Predicting the Mode of Action of Bioactive Compounds via High Throughput Screening and Computational Algorithms

by

Marcos H. Woehrmann

To develop more effective therapies to treat human diseases, a better method of finding the biological targets and modes of action of new compounds is needed. Target predictions have traditionally been made by comparing a new compound's molecular structure to that of known compounds. In many cases this method does not accurately predict a chemical's function since "small chemical changes in an active molecule can render it either nearly or completely inactive or increase its activity dramatically" (Eckert and Bajorath, 2007). Further, prediction by structural comparison has limited application; it can only be used on chemicals with established structures and only identifies new compounds that are structurally similar to known compounds.

A majority of existing drugs have been discovered by identifying the active ingredient of traditional medicines. More recent techniques of drug discovery screen a library of compounds for effectiveness in treating a single disease. However, this method requires re-screening the library when searching for treatments for other diseases; a critical barrier to expediting and scaling

drug discovery. Screening efficiency is particularly important since advances in robotic chemical synthesis and the search for natural products from the oceans are rapidly increasing the size of drug candidate libraries.

In contrast to current approaches which screen compounds for treatments for single disease; my research focused on creating screening methods that deliver a library of chemical fingerprints which can be used to find potential drug candidates for a multiplicity of diseases.

My work produced three screening methods that generate fingerprints useful for predicting a compound's mode of action: cytological profiling, D-Map, and BioSpace. All of these showed positive results towards solving the screening bottleneck. Finally, combining these approaches to integrate these various fingerprints could increase prediction accuracy of screening methods.

To Jill, Zoë, and Maxwell

Acknowledgements

I thank all of the members of the Stuart Lab, past and present, who motivated me and gave me advice and assistance.

I thank Walter Bray for not only tolerating but actually welcoming me when I rotated in the Chemical Screening Center.

I thank Kathryn Siegel and Vikram Sundar, two talented high school students I was privileged to mentor and who are now undergraduates at MIT and Harvard, respectively. In particular Kathryn worked on the BioSpace project and several ideas that I've incorporated in my research were developed during discussions with her.

I thank my committee members Scott Lokey and Phil Berman; Scott not only served on my committee, he supervised much of my work with the UCSC Chemical Screening Center.

And finally, I thank my advisor, Joshua Stuart, who put up with me for more years than I care to remember.

The text of this dissertation includes reprints of the following previously published material:

Woehrmann, M. H., N. C. Gassner, W. M. Bray, J. M. Stuart and S. Lokey (2010). "HALO384: A Halo-Based Potency Prediction Algorithm for High-

Throughput Detection of Antimicrobial Agents.” *Journal of Biomolecular Screening* 15(2): 196-205.

Woehrmann, M. H., W. M. Bray, J. K. Durbin, S. C. Nisam, A. K. Michael, E. Glassey, J. M. Stuart and R. S. Lokey (2013). “Large-scale cytological profiling for functional analysis of bioactive compounds.” *Molecular BioSystems* 9(11): 2604-2617.

Per UCSC policy I have received permission from N. C. Gassner, W. M. Bray, J. K. Durbin, S. C. Nisam, A. K. Michael, and E. Glassey to republish these papers in my dissertation. The senior authors (J. M. Stuart and R. S. Lokey) directed and supervised the research which forms the basis for the dissertation.

The journals that published the papers, *The Journal of Biomolecular Screening* and *Molecular BioSystems*, both allow reproduction of published articles in the author’s dissertation without requesting formal permission.

1 Introduction

The emergence of new diseases and existing bacteria and viruses becoming drug resistant makes the need for new drugs more urgent than ever. Additionally there is a lack effective treatment for many orphaned diseases (diseases for which it is not cost-effective to develop treatment because they afflict very few people or primarily people in developing countries where the cost of development is difficult to recoup). Traditional methods of screening and identifying compounds for drug candidates are slow and expensive; the ability to assess new potential drugs is lagging behind the discovery of new compounds. New ways of conducting and automating compound modes of action discovery or target predictions are needed.

The Pressure of New and Resistant Diseases

The challenge for medical researchers is to keep pace with and react quickly to the emergence of drug resistant pathogens and new viral and bacterial diseases, which now cross geographies rapidly. Contributing to the problem, physicians are prescribing antibiotics for diseases that do not require them. This combined with patients failing to take medications completely as prescribed is causing microbials to develop antibiotic resistance at a more rapid rate. “In the last several years, the frequency and spectrum of antimicrobial-resistant infections have increased in both the hospital and the community. ... The increasing frequency of drug resistance has been attributed to combinations of microbial characteristics, selective pressures of antimicrobial use, and societal and technologic changes that enhance the transmission of drug-resistant organisms.” (Cohen, 1992)

Antibacterial soap has also been identified as exerting pressure on bacteria to evolve antibiotic resistance (Levy, 2001). “The emergence of antibiotic-resistant bacteria as a result of the ubiquity of antimicrobials in the environment is an evolutionary lesson on microbial adaptation.” (Morse, 1995) Compounding this problem is that genes transfer frequently in microbials, allowing newly developed resistant genes to transfer from nonpathogenic to pathogenic microbials.

There is an increase in emergence of new diseases, with scientists discovering one to two new agents of disease per year. “It seems there is something about modern times – these are good times of pathogens to be

invading the human population,” according to Professor Woolhouse, epidemiologist at the University of Edinburgh (Rincon, 2006). Factors contributing to this rise of new diseases include changes in the ways humans interact with other animals in their environment due to deforestation, agricultural changes, and use of exotic livestock (Woolhouse and Gowtage-Sequeria, 2005). Examples of diseases introduced in this way include the H1N1 virus which was acquired from pigs in Europe (Brown, et al., 1997) and caused the 2009 swine flu pandemic; H5N1, the bird flu virus, from wild birds and poultry in Asia (Lewis, 2006); and HIV/AIDS, transferred from chimpanzees to humans in West Africa (Gao, et al., 1999). “HIV/AIDS, malaria, tuberculosis, influenza, SARS, West Nile virus, Marburg virus, and bioterrorism are examples of some of the emerging and reemerging threats.” (Fauci, 2005)

Global travel, global trade, and hospitalization are exacerbating the urgency surrounding finding treatments for new infectious diseases. Infections that were previously localized, are taking advantage of opportunities to spread globally as a result of the greater volume and speed of travel (Morse, 1995).

The Evolution and Explosion of Drug Sources

Historically, natural products have been the source of drugs. Until 2006, 73% of all drugs were biological, natural products, or synthetic modifications of natural products (Newman and Cragg, 2007). For example, aspirin (acetylsalicylic acid), a chemical found in the bark of willow trees, has been documented as early

as 450 BCE to relieve pain and reduce fevers. Penicillin, the ur-antibiotic, produced by the fungus *Penicillium rubens*, was serendipitously found in 1928 by Alexander Fleming to inhibit the growth of bacteria (Chain, et al., 1940; Fleming, 1929). More recently, in 1975, another natural compound, Rapamycin, was discovered in soil from Rapa Nui (Easter Island) (Vezina, et al., 1975). Rapamycin was initially thought to be an anti-fungal agent but has since been utilized as a transplant immunosuppressant (Saunders, et al., 2001), an anticancer drug (Seto, 2012), and has shown promise in treating Alzheimer's (Spilman, et al., 2010), Muscular Dystrophy (Ramos, et al., 2012), and Lupus (Warner, et al., 1994).

Before the twentieth century ethnopharmacologists looking for medicines interviewed indigenous peoples to learn which plants and animals were used medicinally. In 1955, the National Cancer Institute founded the Cancer Chemotherapy National Service Center (CCNSC) with the mission to systematically collect and screen natural products for anticancer activity. Over the next 22 years, the CCNSC tested 114,000 extracts from 35,00 plants along with 16,000 extracts from 550 species of animals (Suffness and Douros, 1982) leading to the discovery of chemotherapy drugs Paclitaxel (trade name Taxol®) (Seto, 2012) and Ellipticine (Miller and McCarthy, 2012), among others.

Having largely exhausted land-based collecting, scientists turned to the oceans. Mostly unexplored and particularly diverse in terms of temperature, pH, and salinity (DeLong, 2007), the oceans are proving to be an especially promising

source of new compounds. Marine sponges, which feed by filtering sea water and have therefore developed efficient defense mechanisms against foreign attackers such as viruses, bacteria, and eukaryotic organisms, are a rich source of novel compounds. Of the 18,000 marine products identified as of 2009, more than 30% came from sponges (Koopmans, et al., 2009). As of 2009, it is estimated that only 0.01% to 0.1% of the ocean's microbes are known (DeLong, 2007). These discoveries are leading to an explosion in natural products awaiting analysis.

Further expanding the pool of potential drugs, innovations in synthetic chemistry have made possible the production of millions of new compounds. "The advent of combinatorial chemistry for the high-throughput synthesis of compounds has driven the advancement of new and emerging technologies for synthetic chemistry laboratories. Automated methods for reaction design, information management, chemical synthesis, compound analysis, and biological testing are necessary to realize the full potential of combinatorial chemistry efforts." (DeWitt and Czarnik, 1995).

The capacity to screen compounds has kept pace with the rate of discovery until recently. As of 2000, only 6% of the approximately 250,000 known plant species have been screened for biological activity (Fabricant and Farnsworth, 2001).

The screening backlog has further worsened because of the growing use of laboratory robotics for chemical synthesis has accelerated the productions of new compounds. As robots become more sophisticated this will to lead to a further

explosion of new molecules: “ ‘I would consider it entirely feasible to build a synthesis machine which could make any one of a billion defined small molecules on demand,’ declares Richard Whitby, a chemist at the University of Southampton, UK.” (Peplow, 2014).

Deficiencies in Drug Discovery Methods

Traditional screening techniques, which typically involve testing hundreds of thousands of compounds for effectiveness against an individual disease, are not efficient nor cost effective. Typically high-throughput screens look for a drug candidate that is effective against a general target class (ion channels, nuclear receptors, etc.), specific biological target (i.e. pathway), or general function (the NCI screens compounds it collects only for anti-cancer activity (Fabricant and Farnsworth, 2001)). Compounds that show activity are subjected to further analysis and optimization before moving on to animal trials. This system is inefficient; when screening for drug candidates for the next disease all the compounds need to be screened again.

A pre-screening technique that could limit the number of compounds needed to be screened would improve the efficiency of this system. An ideal pre-screening system would predict the potential use of a library of compounds as a treatment for all diseases. While this is not yet feasible, as you will see, it is possible to compare new compounds to existing drugs to determine which have similar modes of action. Those compounds which have similarities to an existing

drug would then be tested to see if they are an improvement, with reduced side effects, easier production, or higher efficacy for drug resistant disease. Those compounds which show a biological effect during the screen, but do not match any existing drug, could be examined for novel modes of action. Compounds which exhibited no activity, would be omitted from future screens to reduce inefficient and unnecessary testing. For example, one screen of 114,045 compounds found only 4.3% to be active (Suffness and Douros, 1982). “The high attrition rate of drug candidates during clinical trials for poor pharmacokinetic and metabolic properties has created a need to do these studies as early as it is possible during the drug discovery process.” (Darvas, et al., 2002)

As an example of the magnitude of the screening problem, the National Cancer Institute, through its Developmental Therapeutics Program, has over 500,000 synthetic compounds and 170,000 extracts from plants and marine organisms in its repository. The NCI is now testing these for anti-cancer properties at the rate of 20,000 per year; but from 1990 to 2002 only 75,000 had been screened (Monga and Sausville, 2002). A method that reduces the number of the compounds to be screened would reduce the time and cost of drug development by allowing more targeted research efforts and focus on compounds with a greater likelihood of efficacy.

One of the uses of prescreening is to classify compounds by their mode of action, the physical or functional induced in a cell. Modes of action can be broad, i.e. cell cycle arrest, or specific, i.e. cell cycle arrest in M-Phase. Complicating

mode of action classification is that many compounds have multiple modes of action or a mode of action that is concentration dependent. My goal in prescreening is to determine the dominant mode of action that occurs near the middle of the dilution range (the midpoint between having no discernible effect and being 100% toxic).

The efficiency of drug screening could be further improved if there were public repositories that listed prescreening results (the PubChem BioAssay database (Wang, et al., 2013) does have screening data available, but only for specific screens and only for those molecules found in PubChem). Currently most screening is done by pharmaceutical companies, who are not willing to release or share data. However, university screening centers are becoming more common and are willing to deposit screening results into a public database. An additional advantage of a screening data repository is that it would be a place where negative results could be published, reducing duplicated efforts in re-examining compounds that have already been screened.

1.1 Current Prescreening Techniques

Computationally comparing the structure of compounds has traditionally been used to identify potential drugs to be screened. For diseases where existing pharmaceuticals are losing efficacy due to a build up of resistance, or have serious side effects, scientists often start screening structurally similar chemicals with the objective of discovering a minor difference that could make a significant improvement. (Johnson and Maggiora, 1990). “One approach is to design sets of compounds ‘similar’ to known active compounds in the hope that alternative molecular structures are found that maintain the properties required while enhancing e.g. patentability, medicinal chemistry opportunities or even in achieving optimised pharmacokinetic profiles” (Bender and Glen, 2005). However, as we will see, the concept of similar structures having similar functions is not reliable, and cannot be used at all if the structure of a compound is not known, as is the case with newly found natural products.

A widely used measure of structural similarity is the 166 bit long MACCS key (Molecular ACCess System) developed by Molecular Design Limited (MDL) (Durant, et al., 2002). Each bit in the MACCS key is a binary indication of the presence of a specific feature in the molecule being identified. For example, bit

146 is set if there are more than 2 oxygen atoms present in the molecule and bit 14 is set to indicate the presence of a disulfide bond. Two molecules can then be compared for similarity by a variety of distance measures, a common one being the Tanimoto coefficient (Rogers and Tanimoto, 1960; Tanimoto, 1957), also known as the Jacquard index.

The Tanimoto coefficient is calculated by as follows:

$$Tanimoto = \frac{N_{both}}{N_a + N_b - N_{both}} \quad (1)$$

where N_a and N_b are the number of bits set in molecule a and b, respectively, and N_{both} is the number of bits set in common. The Tanimoto coefficient varies between 1, in the case of two molecules have the same MACCS fingerprint, to 0, indicating that the two molecules have no features in common.

Other distance measurements, such as the Cosine Coefficient (Holliday, et al., 2002):

$$Cosine = \frac{N_{both}}{\sqrt{N_a N_b}} \quad (2)$$

can perform somewhat better for similarity searching than the Tanimoto coefficient (Bender, 2005) but they are little used.

While MACCS keys are easy to calculate and to compare, they often do not capture differences and similarities in compounds that have similar biological activity.

Daylight fingerprints (Daylight Chemical Informations Systems, 2011) were developed to overcome the limitations of MACCS keys, but researchers found that even highly similar compounds, those with a Tanimoto coefficient ≥ 0.85 , have only a 30% chance of being active. “Although this enrichment is greater than that found with random screening and docking to three-dimensional structures, this low fraction of actives within similar compounds occurs not only because of deficiencies in the Daylight fingerprints and Tanimoto similarity calculations but also because similar compounds do not necessarily interact with the target macromolecule in similar ways.” (Martin, et al., 2002).

Attempts have been made to develop measures of structural similarity that more accurately model what makes two compounds functionally similar, such as Extended Connectivity FingerPrints (ECFP) which are “explicitly designed to capture molecular features relevant to molecular activity” (Rogers and Hahn, 2010).

ECFP calculates an atom identifier, a hashed value based on an atom and its bonds, and then iteratively steps out from that atom to the neighboring atoms, calculating hash values based on that substructure. In this way, “Each atom identifier contains topological information on the atom that includes the number of immediate heavy atoms, the atom’s mass, its charge, the number of hydrogens attached, the valance minus the number of hydrogens and whether it is part of a ring.” (Sussex Drug Discovery, 2013). Calculating the atom identifier is repeated starting at every atom in the molecule, with duplicate hash values being

discarded; the final set of hash values represents the fingerprint. Typically, the number of iterations is limited to 4 or 6, resulting in ECFP_4 and ECFP_6 fingerprints, respectively. The ECFP fingerprints are then compared using one of several similarity measures, the Tanimoto coefficient being the most common (Hu, et al., 2009).

ECFP fingerprints generally have a better correlation with biological function than MACCS keys, but two dissimilar compounds can have a similar biological effect, and conversely two similar compounds can act very different in vitro. “The great challenge for in silico methods is generation of models that correlate more closely with in vivo systems.” (Darvas, et al., 2002) Due to open source implementations of ECFP fingerprints have only recently become available they infrequently used in published experiments.

In this thesis, I will present my research in and contribution to the field of screening to determine molecular function similarity and predict drugs targets. I developed several different approaches, in both human cells (HeLA) and microorganisms: Cytological Profiling, BioSpace, and D-Map.

1.2 Cytological profiling

First presented in 2004 as a method for “discovering the mechanism of and predicting the toxicity of new drugs” (Perlman, et al., 2004), cytological profiling has the potential to allow screening a large library of compounds to determine functional similarity *in vivo*.

Cytological profiling exposes mammalian cells (typically HeLa) to a library of compounds and then uses staining and automated microscopy to examine the cells for perturbations from control cells. Features examined can include the size of the nucleus, the number of cells, etc. Depending on the number of stains and the analysis software used, up to 700 different features can be gathered forming a cytological fingerprint, profiling a compounds effect on a cell. The degree of similarity between two cytological fingerprints indicate the compounds’ mode of action similarity (Perlman, et al., 2004). It is also possible find clusters of compounds exhibiting common modes of action using a library of cytological fingerprints. Additionally, by comparing cytological profiles of new compounds to those of known drugs, predictions can be made of a compounds mode of action. New compounds that show no similarity to existing drugs may have a novel mode

of action that would make them particularly interesting to be screened for treatment of diseases for which no treatment is available.

Perlman tested 100 compounds (of which 61 showed enough response to be analyzed) in 9 mode of action categories. Of the 100 compounds, 90 were known, 6 were blinded copies of one of the known compounds but at a different concentration, 1 was known to have multiple targets, and 3 had unknown modes of action. All compounds were tested at 13 concentrations and analyzed with 11 stains producing 93 cytological features. The cytological features were clustered and compared using a variety of distance measures including Tanimoto and Euclidian (Deza and Deza, 2009).

The Perlman results showed a very high correlation for compounds in the Microtubule and Protein synthesis categories and lower but still statistically significant correlation in 4 other categories (Actin, DNA replication, Histone deacetylase, and Topoisomerase). There was a high inverse correlation between the number of compounds in a category and that categories' P-value; the three categories with the lowest P-values (Microtubule, Protein Synthesis, and Histone deacetylase) contained the first, second, and third largest number of compounds in each category, respectively. Perlman correctly identify the mode of action of 5 of the 6 blinded compounds, 4 of which were in the Microtubule and Protein synthesis categories.

Following on the work of Perlman, a study (Young, et al., 2008) was performed that used factor analysis to reduce the number of features. Young

screened 6,547 compounds for 30 cytological features, retaining the compounds in the top 5% phenotypic responses in the two replicate screens, resulting in 211 compounds. The 30 cytological features were reduced to 6 phenotypic attributes by using factor analysis. The similarity of compound pairs using the cytological profiling fingerprints were compared to fingerprints based on structural similarity, as calculated by ECFP_4. Young found that 96% of the time compounds with at least a ECFP_4 Tanimoto coefficient of 0.3 had a high phenotypic attribute Euclidean distance score, which they defined to be less than 1 (Euclidean distances scale from 0 to infinity, with 0 being perfect correlation). Note that the converse was not true; many compound pairs had high phenotypic attribute similarity but a low structural similarity score. They also examined how well the clustering of the phenotypic attributes ordered the structural fingerprints and found only a 0.0746 correlation. They hypothesized this was due to many of the compound pairs showing a high degree of phenotypic similarity showed little or no molecular similarity. However, there were enough highly correlated pairs to produce a P-value of 0.001.

In 2009, Feng wrote a review of cytological screening and proposed that combining it with other phenotypic profiling, such as DNA microarrays and antibody-based protein detection, would produce better results. He argued that “an earlier integration of phenotypic profiling technologies, combined with effective experimental and in silico target identification approaches, can improve

success rates of lead selection and optimization in the drug discovery process.” (Feng, et al., 2009)

Even though cytological profiling has shown to be effective at determining at least some classes of drug targets, there are disadvantages when using mammalian cells in a screen. Because they have a 23 hour doubling time, HeLa cells are time consuming as well as difficult and expensive to grow. (Norcliffe, et al., 2014) Further, they are subject to contamination and can also contaminate other cell lines (Lucey, et al., 2009); this contamination can be difficult to detect and has even caused papers to be retracted (Cai, et al., 2011).

The scanning and analyzing of the HeLa cells is time consuming and computationally complex. The microscopic scanning takes up to 60 minutes per plate and the image analysis takes a further 6 to 12 hours. While the analysis can be done in parallel using a cluster of computers, the licensing costs for the image analysis software is not insignificant.

1.3 Screening in Yeast

As discussed, the growing cells of human origin in automated screening systems is difficult; cheaper and more efficient screening can be performed using yeast. Though humans and yeast differ in appearance, many of the basic biological processes are preserved. “In yeast cells, the function of human proteins can often be reconstituted and aspects of some human physiological processes can be recapitulated because of the high degree of conservation of basic molecular and cellular mechanisms between yeast and human cells.” (Barberis, et al., 2005)

In addition, yeast screens are faster: “As an *experimental* organism, yeast is the model of simplicity. Its 90-minute population doubling time allows for relatively short experiments examining exposure to drugs over many population doublings. In addition, yeast requires very simple and inexpensive media, can grow either in liquid culture or on solid media, and can be readily manipulated to express foreign genes.” (Simon and Bedalov, 2004)

A fundamental difference between mammalian cells and yeast assays is that the former is performed as a High-Content Screen (HCS) and the latter a High-Throughput Screen (HTS). A HCS examines many features, including the

number of cells, size of nucleus, etc., whereas a HTS is typically measures only one or two characteristics, for example, growth inhibition.

Yeast screens can be conducted in a small wells using a liquid medium or on a plate using a semi-solid medium. In either case, the yeast is grown in YPD (Yeast extract Peptone Dextrose); the difference being that the solid medium additionally contains agar (Bergman, 2001). After the yeast has been placed in the medium it is exposed to the compounds being tested and grown for ten or more doubling times. If growth inhibition is being measured, the optical density of the yeast is then read by using a plate reader. By comparing this optical density to yeast grown without exposure, the amount of growth inhibition can be calculated (Gassner, et al., 2007).

The high throughput screening techniques in yeast I have developed makes use of synthetic lethality and Bliss Independence to predict modes of action of unknown compounds.

Synthetic Lethality

Two genes are synthetically lethal if silencing both genes results in a nonviable organism, but removing either gene allows the organism to live. Traditionally synthetic lethality has been used to describe pairs of genes, but it can be extended to a gene-drug pair and even a drug-drug pair (figure 1). I make use of this to build fingerprints and to directly predict drug targets in both D-Map and BioSpace.

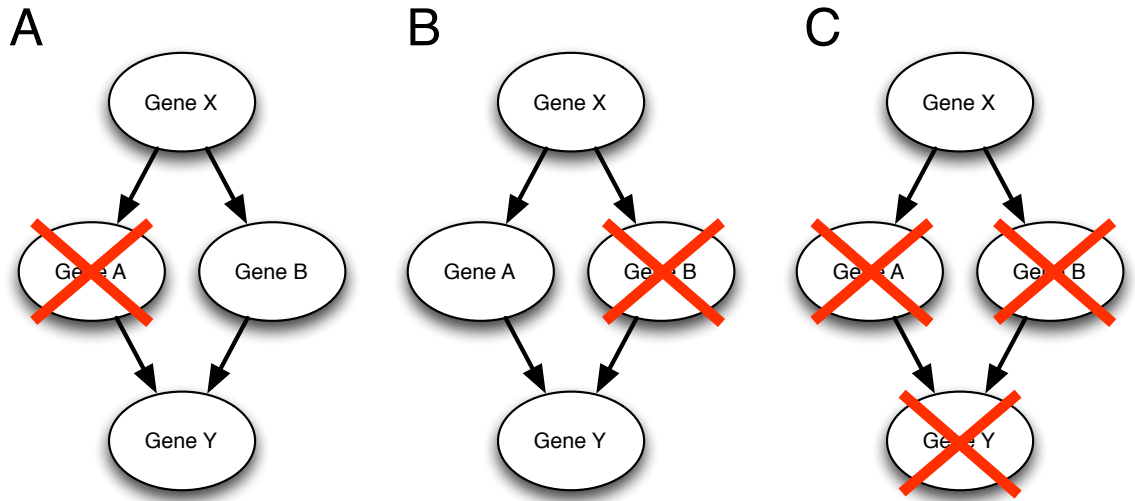


Figure 1 – Synthetic lethality. In this simple example gene Y is vital to the organisms survival and genes A and B are synthetically lethal in the pathway from X to Y. If either A or B are disabled (sub-figure A and B), either via knockout mutant or by exposure to a drug which targets them, the pathway continues to operate and the organisms survives. If both A and B are disabled the organisms dies (sub-figure C).

Somewhat surprisingly synthetic lethality is quite common in many pathways, a large survey of the yeast genome (Costanzo, et al., 2010) found that nearly all genes participate in at least one synthetically lethal interaction, with 5 being the most frequent number of interactions. Synthetic lethality may have evolved not only to provide redundancy, but also to allow evolution to take place. For example, if genes A and B are synthetically lethal in pathway 1, and gene A, which is also part of pathway 2, evolves so that it is more effective in pathway 2 but no longer able to carry out its function in pathway 1, the organism will continue to survive. If the survival disadvantage of losing redundancy in pathway 1 has less impact than the survival advantage of improved function in

pathway 2, the mutation has conferred an evolutionary advantage. It does not appear that synthetic lethality is an accidental byproduct of gene duplication: “Although ~13% of the yeast genome is a relic of an ancient duplication of the entire genome, duplicated genes do not seem to make a disproportionate contribution to buffering.” (Hartman, et al., 2001) (buffering is synthetic lethality for the purposes of redundancy).

Synthetic Lethality and Synthetic Genetic Array (SGA)

In the 1990s, the idea of performing a systematic double deletion of every pair of genes in the yeast genome was first proposed to produce an interaction network (Boone, 2014; Uetz, et al., 2000). A partial network consisting of a screen of 132 query genes against ~4700 deletion mutants was published in 2004 (Tong, et al., 2004). A more complete network, covering 30% of the yeast genome was published 6 years later, in 2010 (Costanzo, et al., 2010).

Synthetic lethality as double gene deletion has been extended to a gene-drug pair (Dunstan, et al., 2002; Parsons, et al., 2003); instead of knocking out a gene, a drug is used to silence a gene. For example, if a single knockout yeast colony is exposed to a drug, which is not toxic in wild-type yeast, and the colony dies, it can be surmised that the drug is targeting one of the genes that is synthetically lethal with the gene that was knocked out.

I make use of synthetic lethality and the Boone SGA data in the BioSpace project, where I use a set of yeast knockouts to build a library of drug

fingerprints that are used to predict modes of action and to predict drug targets directly.

Deletion of one a gene in synthetic lethal pair may reduce the growth rate of an organism, and deletion of both genes may not kill the colony, just significantly inhibit its growth. The loss of two genes that are not synthetically lethal may also inhibit growth, but less than the loss of two synthetically lethal genes (Giaever, et al., 2002). Therefore, it is necessary to decide if a reduction in growth is indeed a synthetic lethal pairing or just two independent growth reduction effects. To determine if two genes are synthetically lethal, we make use of Bliss Independence (Bliss, 1939).

Bliss Independence

It has long been known that drugs interact in biological systems (Fraser, 1872). Bliss grouped drug-to-drug interactions into classes of additive, synergistic, and antagonistic, based on the effects of two drugs in combination versus each drug singly (Figure 2). A synthetic lethal interaction is one that has more inhibition than is predicted by Bliss Additivity.

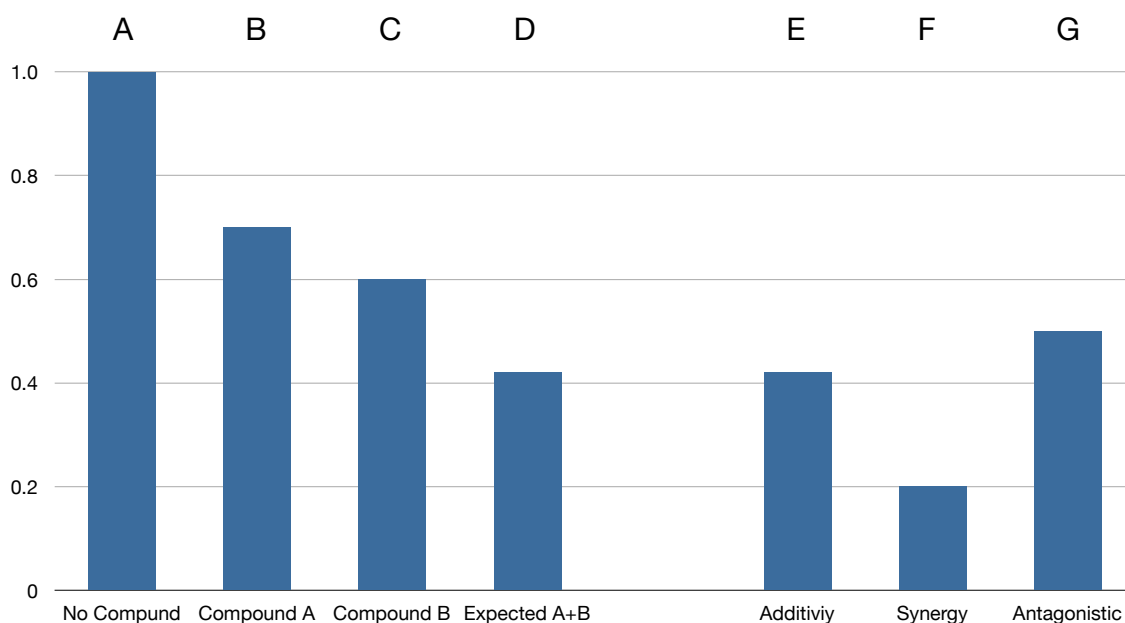


Figure 2 – Bliss Independence example. Percentage growth inhibition with no compound (Column A), with compounds A and B alone (columns B and C), left scale shows relative growth. Column D shows the predicted inhibition of compounds A and B combined calculated via compound A inhibition times compound B inhibition. If the two compounds are additive (aka Bliss Independent) the expected results would close to the expected value, indicated by column E. If growth inhibition is higher than expected based on the individual effects of each drug (column F) the compounds are said to be synergistic. If inhibition is less than expected (column G) the compounds are antagonistic.

As an example of drug knockout synthetic lethality we can look at this example of a growth time series of a *Saccharomyces cerevisiae* *cyk3* knockout in grown in the presence of Actinomycin compared to the growth of the same knockout without Actinomycin and the growth of wild-type with Actinomycin, figure 3.

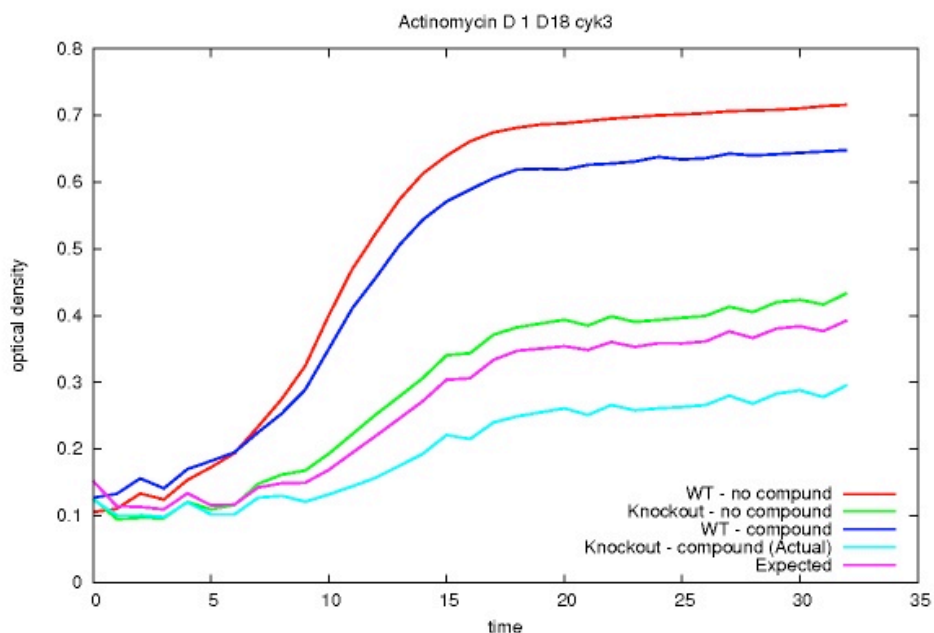


Figure 3 – An example of yeast knockout response to a drug. The top line (red) shows the growth of wild-type (WT) yeast over time (in hours) without a drug, the optical density scaling linearly with the amount of yeast present. The second line (blue) shows the presence of Actinomycin slightly reducing growth, the third line (green) shows the growth of the *cyk3* knockout growing without the drug. The expected growth, based on the Bliss model of independence, is shown with the fourth line (magenta). The actual growth of the mutant in the presence of Actinomycin is shown in the bottom (cyan) line. Since the growth is less than expected, we conclude that the *cyk3* gene is synthetically lethal with one or more genes targeted by Actinomycin.

1.3.1 D-Map

As previously discussed, pairs of drugs cause more growth inhibition than predicted by Bliss Additivity if the the genes they target are synthetically lethal with each other. D-map uses this concept by growing yeast (*S. cerevisiae*) exposed to pairs of compounds to establish fingerprints.

D-Map adds one compound to the agar upon which the yeast is grown, and pins a screening library of compounds into the agar (pinning transfers small amounts of a set of chemicals from liquid reservoirs onto the surface of the agar using hollow needles via capillary action). After sufficient doubling time, the plates are scanned, and the inhibition level of each pair of drugs is measured. The actual amount of inhibition is compared to the inhibition predicted by Bliss Additivity model resulting in a a vector of log ratios. This vector is used as a D-Map drug interaction fingerprint. D-Map fingerprints are compared to each other to determine drug similarity or to a library of previously gathered fingerprints of known drugs to predict the class to which and unknown compound belongs. Fingerprints can be extended and predictions made more accurate by adding additional plates pinned with different known compounds or pinning into a different microorganism, for example, *Schizosaccharomyces Pombe*.

The Lokey Lab yeast assay (Gassner, et al., 2007) was used to perform the lab work required for this method; however, as we'll see later there were limitations to this method that required some refinement (Woehrmann, et al., 2010).

1.3.2 BioSpace

As discussed below, there are practical issues with using a pair of compounds to establish fingerprints; for example, bioactive compounds generally target more than one gene, making fingerprints generated with them nonspecific. Using yeast mutants which have a gene knocked out makes it possible to directly control what genes are silenced and thereby control the synthetic lethal gene pairs that are used to build fingerprints.

If the situation were ideal, every gene has exactly one synthetic lethal partner and we have full knowledge of every one of these pairings, it would make determining the target of an unknown compound straightforward. Simply expose a complete set of yeast knockouts to the compound being investigated and the drug target will be revealed by which knockout does not grow. By using yeast knockouts that have the silenced gene replaced by a barcode (Pierce, et al., 2006) it would be possible to test the entire yeast genome in parallel, so only one experiment would have to be done. However, the actual situation is far from ideal, most genes have many synthetic lethal partners and, as mentioned earlier, we have knowledge of only about 30% of the synthetic lethal pairings. However,

as we will see in chapter 3.2, it is possible to use a library of yeast knockouts to make probabilistic predictions of drug targets.

2 Cytological Profiling

Cytological profiling (CP) (Perlman, et al., 2004) uses automated microscopy to generate a compound fingerprint based on physical attributes of cells, figure 4. Used as a screening technique, the mode of action of an unknown compound can be predicted by comparing its fingerprint to fingerprints of known drugs.

To perform cytological profiling HeLa cells are grown exposed to a compound for ~24 hours. The cells are then stained using one or more stains and scanned on a computer controlled microscope. The resulting images are analyzed by software for various features, such as the shape and size of the nucleus. These features are compared to control cells grown under the same conditions but without the compound being tested. The features make up a ‘fingerprint’ which can be compared to other fingerprints predict compound similarity or grouped with similar fingerprints to create classes of drugs. As was shown by Perlman and

colleagues these fingerprints are a good predictor of a compound's mode of action, at least for some modes of action.

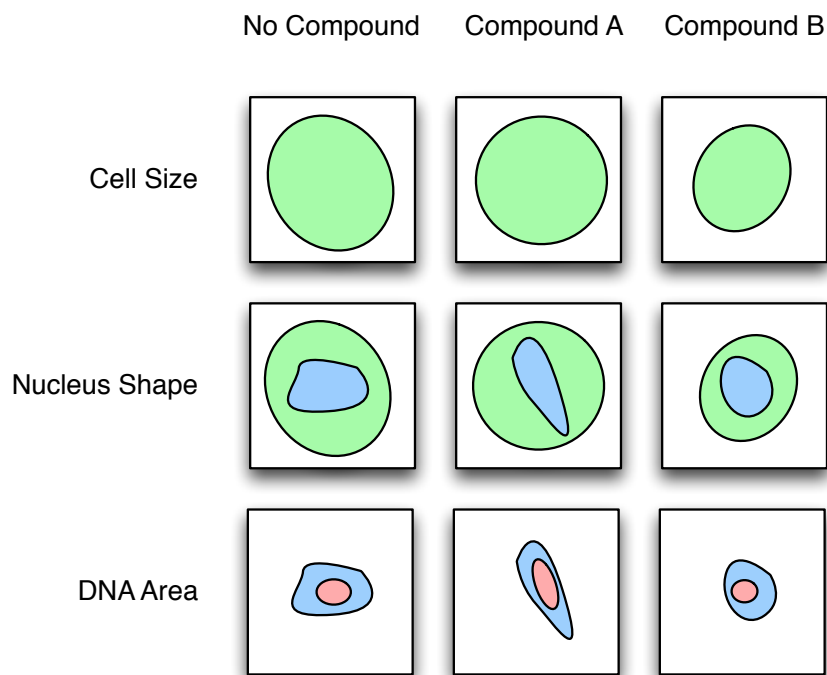


Figure 4 - Example of cytological profiling features. Cell grown exposed to compounds are compared to control cells grown without a compound; the vector of these comparisons is a cytological profiling fingerprint.

In cooperation with the UCSC Chemical Screening Center I developed a pipeline to perform cytological profiling (Woehrmann, et al., 2013). I tested the system by comparing compounds that were known to have a similar mode of action and found novel modes of action in compounds that were later show to be correct.

My contribution to the following paper were: designing the CP pipeline, establishing which features were biologically reproducible, modifying the dimensionless scoring method (initially developed by co-author J. K. Durbin),

creating the Mode-of-Action affinity rank score, calculating the method of action classes Kolmogorov–Smirnov scores and p-values, GSEA analysis, MINE analysis, and finally developing the maximum Pearson correlation (MPC) score, which allows dosage independent similarity comparisons to be made, a significant improvement over the Perlman developed titration-independence scoring system (TISS).

Large-scale cytological profiling for functional analysis
of bioactive compounds†

Cite this: *Mol. BioSyst.*, 2013,
9, 2604

Marcos H. Woehrmann,^a Walter M. Bray,^b James K. Durbin,^a Sean C. Nisam,^b
Alicia K. Michael,^b Emerson Glassey,^a Joshua M. Stuart*^a and R. Scott Lokey*^b

Cytological profiling (CP) is an unbiased image-based screening technique that uses automated microscopy and image analysis to profile compounds based on numerous quantifiable phenotypic features. We used CP to evaluate a library of nearly 500 compounds with documented mechanisms of action (MOAs) spanning a wide range of biological pathways. We developed informatics techniques for generating dosage-independent phenotypic “fingerprints” for each compound, and for quantifying the likelihood that a compound’s CP fingerprint corresponds to its annotated MOA. We identified groups of features that distinguish classes with closely related phenotypes, such as microtubule poisons vs. HSP90 inhibitors, and DNA synthesis vs. proteasome inhibitors. We tested several cases in which cytological profiles indicated novel mechanisms, including a tyrothostin kinase inhibitor involved in mitochondrial uncoupling, novel microtubule poisons, and a nominal PPAR- γ ligand that acts as a proteasome inhibitor, using independent biochemical assays to confirm the MOAs predicted by the CP signatures. We also applied maximal-information statistics to identify correlations between cytological features and kinase inhibitory activities by combining the CP fingerprints of 24 kinase inhibitors with published data on their specificities against a diverse panel of kinases. The resulting analysis suggests a strategy for probing the biological functions of specific kinases by compiling cytological data from inhibitors of varying specificities.

Received 23rd June 2013,
Accepted 20th August 2013

DOI: 10.1039/c3mb70245f

www.rsc.org/molecularbiosystems

Introduction

Multi-parametric phenotypic profiling has emerged as a powerful tool for characterizing small molecules and their effects on cells or model organisms. In particular, the union of high-throughput screening technologies with automated microscopy and image processing has enabled the development of a technique called “cytological profiling” (CP),^{1–3} in which cells are imaged in multi-well plate format using automated microscopy and the resulting images are quantified in terms of various descriptors, or “features”. Rather than focusing on a specific phenotype or biological endpoint, CP strategies employ multiple cytological probes selected to cover a wide range of biological pathways. The resulting images generate hundreds of quantifiable cytological features that capture well averages (e.g. overall fluorescence intensity) as well as distributions of measurements on individual cells (e.g. nuclear size).

Unbiased image-based screening approaches such as CP have been used to classify bioactive compounds by MOA based on their phenotypes,^{2,4,5} predict kinase inhibitors with novel target profiles,⁶ investigate cell cycle modulators,^{7–10} and classify biofilm inhibitors.¹¹ Our labs have also applied CP as a “function first” approach in the evaluation of natural product extracts, enabling prioritization of natural product extracts with novel biological activity prior to purification and structure elucidation.¹² While the reference libraries used for most of these studies were limited to relatively focused sets of compounds with a limited range of MOAs, in the present study we describe the application of CP to evaluate a library of nearly 500 compounds with documented MOAs spanning a wide range of biological pathways. We have developed a bioinformatics technique based on histogram differences, which combines several hundred diverse imaging features into comparable dimensionless scores for use in predicting MOA. Using a *t*-test, KS-, and Silhouette-based tests newly applied to these types of data, we validated that the dimensionless scores can be used to accurately predict known MOAs. We then show that a pipeline based on these scores can identify unanticipated classes for several drugs, even for classes where the distinguishing cytological features had no obvious biological connection to the predicted classes.

^a Department of Biomolecular Engineering, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA. E-mail: jstuart@ucsc.edu, slokey@ucsc.edu

^b Department of Chemistry and Biochemistry, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70245f

We tested several of these cases in which cytological profiles indicated novel mechanisms and confirmed the MOAs predicted by the CP signatures. Finally, we were able to identify correlations between cytological features and kinase inhibitory activities by combining our CP dataset with published data on the specificities of 24 kinase inhibitors. This analysis led to the identification of networks of kinases and CP features that recapitulated known interactions that would not have been detected based on the activity of any single inhibitor alone.

Results

We implemented a CP screening pipeline that incorporates different elements from published studies,^{2,4,13} in which automated fluorescence microscopy and computer-aided image analysis are used to detect and quantify various cellular level properties, or “features,” that change upon addition of a drug. Compounds were delivered robotically to two separate assay plates, each of which received a different set of fluorescent stains. Stain set 1 contained a DNA stain (Hoechst), an anti-phosphohistone H3 antibody (pH3) to mark mitotic cells, and 5-ethynyl-2'-deoxyuridine, an S-phase marker which is incorporated into DNA and visualized with rhodamine-azide using Cu^I-catalyzed “click” chemistry.¹⁴ Stain set 2 contained Hoechst and stains for microtubules (MTs) and actin filaments (Fig. 1). The plates were imaged using an automated microscope, and image analysis software was used to create and quantify various features for each fluorescence channel. Many features were derived from measurements of individual cells or nuclei and provided a distribution of values for a given well, *e.g.*, the intensity of EdU fluorescence in each nucleus. Other features were comprised of a single value representing well averages,

e.g., the overall intensity of a stain. Filtering the features based on technical reproducibility (see Methods) allowed us to prune the number of features from 430 down to 248, which were used in all subsequent analyses.

For our set of reference compounds, we selected a commercial library of 480 small molecules with diverse structures and MOAs. This library, called the ICCB collection, was manually curated by the Harvard Institute of Chemistry and Cell Biology (ICCB) to target a wide range of biological processes and targets, including the cytoskeleton, protein synthesis, kinases, nuclear hormone receptors, ion channels and G-protein coupled receptors, and specific classes of enzymes such as phosphodiesterases and nitric oxide synthases, among others. The compounds were arrayed into 384-well plates at 4 dilutions each in DMSO to create 1920 unique samples. Compounds were annotated with respect to their published target(s), function(s), and/or mechanistic class(es) and grouped into 114 classes, each class containing from 1 to 84 compounds (Table S1, ESI†).

Dimensionless measure to relate diverse CP features: the histogram-difference (HD) score

In order to condense all of the diverse cellular feature distributions into discrete values and to control for plate-to-plate variability, we created a generally applicable dimensionless measure, called the “histogram difference” (HD) score, for each feature that quantifies the extent to which that feature changes upon addition of a drug compared to the distribution for that feature in the aggregate of all DMSO control wells within the same plate. Thus, the method only requires measurements in drug and control conditions but the measurements type can take any form. In this analysis all CP measurements, whether continuous (*e.g.* cell size), ordinal (*e.g.* number of cellular nuclei), or categorical (*e.g.* inclusion/exclusion of dye) were converted into *quanta* in which every value was mapped to one of twenty bins. A histogram was then calculated for each feature by calculating the proportion of values falling within each bin. After smoothing the histograms, the HD score was computed as the sum of squared differences between the distribution obtained from the set of bins for the drug and the distribution obtained for the DMSO control. A CP fingerprint was defined as the vector of all HD scores across all of the features.

HD-based CP features detect known biological impacts of several positive control test classes of compounds

Clustering the full matrix of compounds and features revealed that many classes of compounds with known MOA relatedness have similar HD-based fingerprints (visual inspection of hierarchical clustering solution, Fig. 2; Fig. S1, ESI†). Some examples of these classes were the DNA damaging agents, proteasome inhibitors, F-actin cappers, mitochondrial uncouplers, and microtubule poisons. Further demonstrating the utility of the HD-scores, clustering of the features themselves (*i.e.*, the columns in the heat maps in Fig. 2 and 3) revealed sets of features whose variation across compound classes could be mapped, in many cases, to subtle phenotypic differences between those classes.

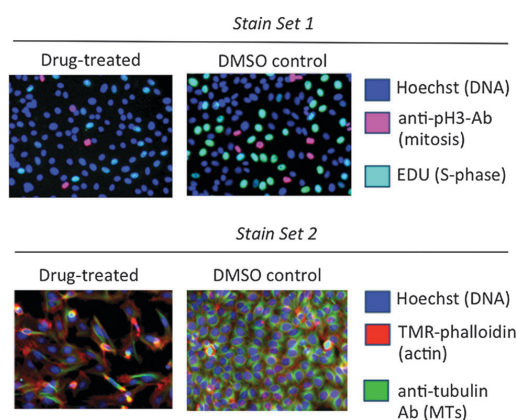


Fig. 1 Stains used for cytological profiling in this study. Two sets of fluorescent stains were used for each compound treatment. Examples of drug-treated wells and control wells are shown for each stain set. Stain set 1 is comprised exclusively of nuclear stains: Hoechst dye (DNA), anti-phosphohistone H3 antibody (mitotic marker), and EDU (clickable version of BrdU, a metabolically incorporated nucleoside analog used as an S-phase marker). Stain set 2 contains Hoechst dye, and cytoskeletal stains for actin and microtubules.

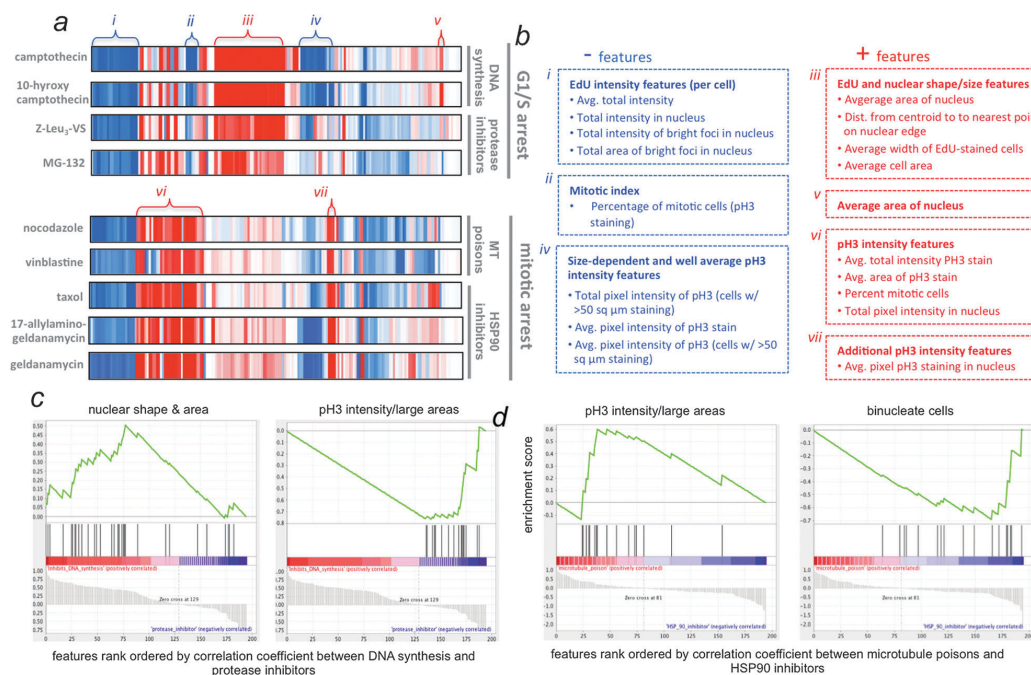


Fig. 3 Expanded heat maps showing cytological profiles of selected compound classes, and the features that allow compounds to be distinguished by MOA. Left. (a) Magnified view of cytological profiles from two major cell cycle phenotypes, G1/S and mitotic arrest. Drug concentrations: camptothecin (0.77 μM), 10-hydroxycamptothecin (3.7 μM), Z-Leu₃-vinyl sulfone (2.4 μM), MG-132 (2.8 μM), nocodazole (0.9 μM), vinblastine (1.5 μM), taxol (0.3 μM), 17-allylaminogeldanamycin (2.3 μM), geldanamycin (2.4 μM). Red features indicate positive values relative to control wells (based on HD analysis output), blue features indicate negative values relative to control, and white features are identical to those in the controls. (b) Groups of features that distinguish target/MOA classes within the same cell cycle phenotype summarized in red and blue boxes based on whether they are primarily increased or decreased relative to control cells, as highlighted with braces in part a. (c) GSEA analysis to identify feature categories that are positively (red) or negatively (blue) correlated with compound class, differentiating between (left) DNA synthesis inhibitors and protease inhibitors and (right) HSP90 inhibitors and MT poisons.

library arrested cells in metaphase with their spindles intact, giving rise to areas of intense pH3 staining that were smaller and more elongated than those in cells treated with MT poisons (Fig. 4b).

A concentration-independent compound similarity score

The phenotypic effects of drugs on cells are generally dose-dependent, ranging from no phenotype at the lower end to, in many cases, cell death at the higher end. At intermediate drug concentrations, multiple phenotypes can emerge depending on the dose. Initially, we clustered the concentrations for each compound independently. Inspection of the resulting heat maps showed that, in many cases, multiple doses of the same compound clustered together, indicating that these fingerprints contain information relevant to that compound's MOA (Fig. 2).

However, in order to quantify the ability of HD-based CP fingerprints to classify compounds according to their MOAs, we needed a dose-independent similarity score for each pair of compounds. In their approach to this problem, Perlman *et al.* developed a titration-independence scoring system (TISS), which

calculates compound similarities by comparing “titration sub-series” across a range of 13 concentrations for each compound.² Compounds that have different potencies but share the same MOA or target tend to have similar fingerprint profiles that are shifted relative to each other over a range of concentrations.² However, the TISS score requires a relatively large number of doses per compound (to generate multiple overlapping concentration windows). Because we had relatively few doses per compound, we developed a variation on the TISS concept based on the maximum Pearson correlation (MPC) among fingerprints in the drug/dose matrix for a given pair of drugs. The MPC of a pair of compounds simply represents the maximum Pearson correlation between the fingerprints of any two compounds within the 4 × 4 matrix of all dosage pairs for those two compounds. The MPC therefore identifies a dosage-independent similarity between two compounds that can highlight overlapping MOAs in cases where phenotypes may diverge at lower or higher doses (*e.g.*, in cases where two drugs have different specificity profiles at different doses).

Because compound-dose instances with very weak phenotypes look similar to each other, we eliminated these instances

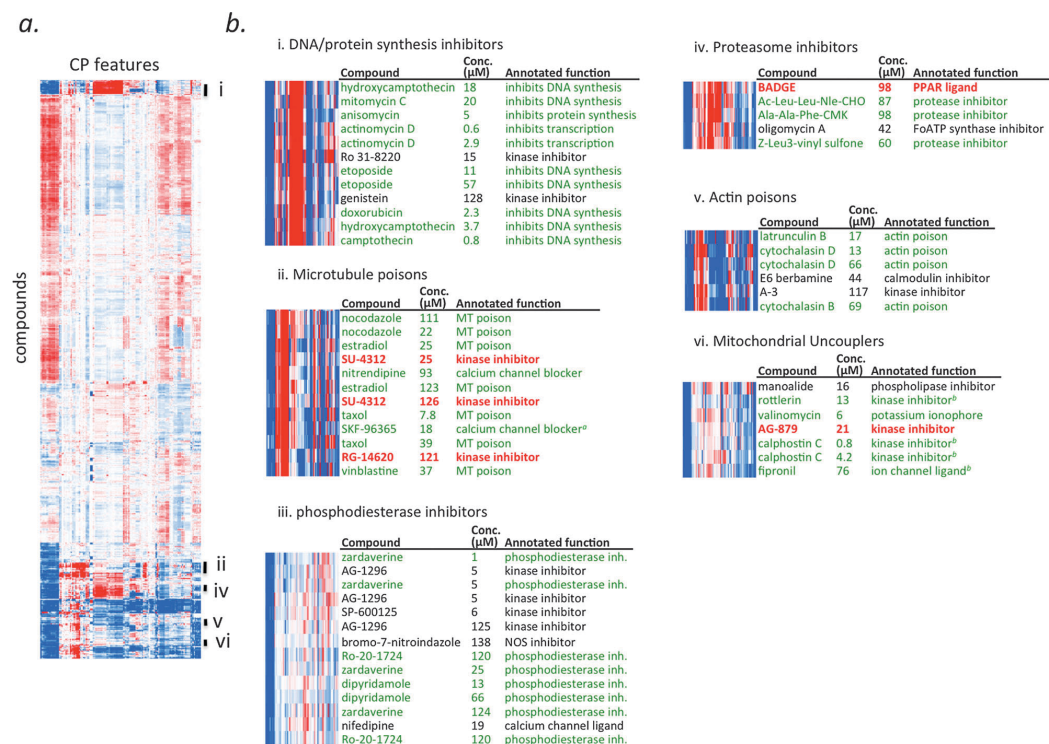


Fig. 2 Heat map of cytological profiles clustered by similarity. (a) Heat map showing all compounds at 4 concentrations each, with compounds in rows and cytological features in columns. Features and compounds were hierarchically clustered based on their pairwise Pearson coefficients. (b) Individual clusters expanded, showing compound names, concentrations, and annotated functions. Compounds in green are annotated with the MOA that defines the cluster. Compounds in black are annotated differently from the MOAs that define their respective clusters. Compounds in red were selected for further testing based on MOAs suggested by their respective clusters.

In many cases, these bins of covariate features tracked to distinct cytological phenomena in the images that reflected their known biological effects. For example, clustering of the DNA damaging agents such as camptothecin and 10-hydroxycamptothecin was driven primarily by (1) negative EdU intensity features due to the G1/S arrest induced by these compounds and (2) positive nuclear and EdU area features due to the corresponding increase nuclear size for cells in late G1 (Fig. 3).

The HD scores provide specificity between close, but distinct compound classes. Like the DNA damaging agents, the electrophilic protease inhibitors such as the proteasome inhibitor MG-132 also had strongly negative EdU intensity features, indicative of the G1/S arrest induced by these compounds. However, these two classes could be distinguished on the basis of their fingerprints due to differences in nuclear size parameters and pH3-related features such as mitotic index (Fig. 3 and 4a). Inspection of the images showed that the DNA damaging agents such as camptothecin produced nuclei that were larger and more rounded than control cells, while the protease/proteasome inhibitors had less of an effect on nuclear size and shape. In addition,

the DNA synthesis inhibitors showed almost no cells in mitosis, while the protease inhibitors had less of an effect on mitotic index. Indeed, these phenotypic differences were reflected in the cytological fingerprints of the compounds, allowing these two classes to be distinguished from each other despite the similarity in their cell cycle arrest phenotypes.

Another example of the HD score's specificity was exemplified in its ability to tease apart the different mitotic arrest-related compounds. The clustering of the microtubule (MT) poisons was driven by intensity and area features in the pH3 stain, corresponding to the arrest in mitosis induced by these compounds (Fig. 3). The HSP90 inhibitors geldanamycin and 17-allylaminogeldanamycin, which also cause a mitotic arrest, had similar cytological profiles to those of the MT poisons. However, unlike the MT poisons, which induce aberrant spindle morphologies by acting directly on microtubules, HSP90 inhibitors have been shown to activate the spindle assembly checkpoint *via* a polo-like kinase (PLK1)-dependent mechanism,^{15,16} causing some cell lines (including HeLa cells like those used in this study) to arrest in metaphase. Indeed, we observed that both of the HSP90 inhibitors in the reference

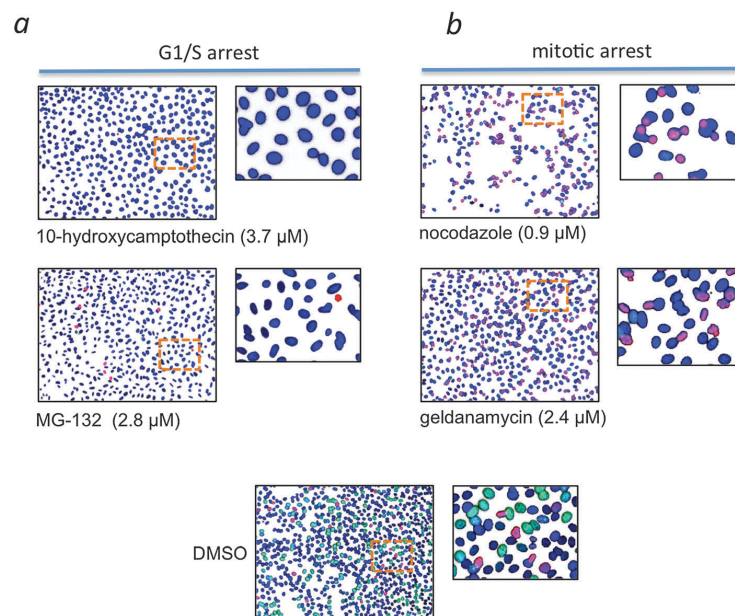


Fig. 4 Images of cells treated with compounds in MOA classes highlighted in Fig. 3. Images from stain set 1 of HeLa cells treated with DMSO (top), 10-hydroxycamptothecin (middle), or MG-132 (bottom). Highlighted regions in orange boxes are shown expanded to the right of each image. Color key: cyan (EdU), blue (DNA), magenta (pH3). Among compounds that cause G1/S arrest, features that distinguish DNA damaging agents (e.g., 10-hydroxycamptothecin) from proteasome inhibitors (e.g., MG-132) relate to nuclear size and total intensity of pH3 staining. For compounds that cause mitotic arrest, features that distinguish microtubule poisons (e.g., nocodazole) from HSP90 inhibitors (e.g., geldanamycin) relate to size and shape features in pH3 stain. Note the elongated metaphase spindles induced by geldanamycin, compared to the more rounded mitotic nuclei from the unstructured spindles induced by nocodazole.

from the MPC analysis. Taking the vector magnitude (*i.e.*, the square root of the sum of the squares of all HD feature scores) for a compound-dose instance yielded an effective measure of its phenotypic strength. Plotting the rank-ordered CP vector magnitudes for all compound-dose instances in the library (Fig. S2, ESI†) shows an elbow at 2.4 in the plot above which distinct phenotypes emerge among the instances (the theoretical maximum vector magnitude is 15.7). The ~75% of the compound-dose instances that fell below this cutoff correspond to a relatively featureless region of the heat map. Eliminating compound-dose instances below this cutoff left 460 instances, representing 239 compounds. Most of the compounds that did not meet the minimum activity threshold even at their highest doses (and were therefore discarded from future analysis) were in the “bioactive lipids” category, which included fatty acids, prostaglandins, and steroid derivatives.

Likewise, we eliminated compound-dose instances that were highly toxic, defined as wells in which there were fewer than 10% of cells remaining after compound treatment, since these instances were similarly uninformative with respect to mechanistic comparisons. This eliminated a further 61 compound-dose instances, resulting in 17 additional compounds being removed from the library, (16 compounds were removed because at the higher concentrations they were too toxic and at the lower concentrations showed no activity, and one compound, calyculin

A, was removed because it was too toxic at all dilutions tested). After filtering out compound-dose instances that were below the phenotypic magnitude threshold and above the toxicity cutoff, we were left with 222 compounds (399 compound-dose instances), representing 46% of the library (Table S1, ESI†).

HD-based CP fingerprints classify compounds with known MOAs for 14 mechanistic classes

To quantify the extent to which the CP data obtained from the reference library was able to classify compounds by MOA, we asked whether the fingerprints of compounds within the same class were more similar to each other than to fingerprints of compounds in different classes. To this end, we collected the MPC scores (as determined above) of all pairs of compounds in the same annotated class (the “within-class” group) and all pairs between compounds of different classes (the “between-class” group). Next we compared the distribution of correlations obtained for the within-class pairs to the between-class pairs, using both the *t*-test and the non-parametric Kolmogorov–Smirnov (KS) test to measure the separation in the two distributions. For 15 out of the 31 classes containing more than one compound, the within-class pairs had significantly higher Pearson correlations (p -value ≤ 0.05) than the pairs between classes using either the *t*-test or KS-test statistic (Table 1). Nine of the classes has both *t*-test and KS-test statistics with significant

Table 1 MOA classes with significant in-class vs. outside-of-class Pearson correlations, their K-S statistics and *p*-values, and the feature categories that are significantly enriched (nominal *p*-value < 0.001) in features that are increased (in italic) or decreased (in bold) in that compound class relative to all other compounds in the library

Class	# cpds in class	KS	<i>p</i> -Value	Distinguishing feature classes based on GSEA analysis
Inhibits DNA synthesis	8	0.84	1.1×10^{-34}	<i>Nuclear area/shape; cell area/shape</i> pH3 intensity/large areas; mitotic index; pH3 area
Calcium channels	14	0.39	2.3×10^{-24}	<i>pH3 intensity/large areas; cell count</i> Nuclear area/shape; binucleated cells
Kinase inhibitor	50	0.11	1.0×10^{-23}	<i>pH3 intensity; fluctuation/gradient</i> Cell count
Microtubule poison	4	0.97	2.5×10^{-10}	<i>pH3 intensity</i> Binucleated cells; cell count
Protease inhibitor	12	0.26	3.3×10^{-08}	<i>Nuclear area/shape; pH3 intensity; cell area/shape</i> EdU intensity; fluctuation/gradient; tubulin intensity
Bioactive lipid	11	0.26	8.6×10^{-07}	<i>EdU intensity; fluctuation/gradient; nuclear area/shape; cell count</i> Nuclear area/shape; cell area/shape
Phosphodiesterase inhibitor	8	0.34	6.9×10^{-06}	<i>Binucleated cells; pH3 intensity/large areas; cell count</i> Actin intensity
Inhibits protein synthesis	3	0.94	5.2×10^{-05}	<i>Nuclear area/shape</i> Fluctuation/gradient
Prolyl <i>cis-trans</i> isomerase inhibitor	3	0.93	6.0×10^{-05}	<i>pH3 intensity/large areas; pH3 intensity</i>
Actin poison	3	0.82	5.8×10^{-04}	<i>pH3 intensity/large areas; fluctuation/gradient</i> Nuclear area/shape; cell area/shape
Calmodulin inhibitor	3	0.60	2.5×10^{-02}	<i>pH3 intensity/large areas</i> Binucleated cells
HSP-90 inhibitor	2	1.00	3.7×10^{-02}	<i>Mitotic index; pH3 area</i> pH3 intensity/large areas
PAF receptor ligand	2	1.00	3.8×10^{-02}	<i>Tubulin foci/1–4 nM; fluctuation/gradient; EdU intensity</i> Nuclear area/shape; cell area/shape; actin intensity
Inhibits transcription	2	0.98	4.4×10^{-02}	<i>Nuclear area/shape; cell area/shape; binucleated cells</i> pH3 intensity/large areas; fluctuation/gradient; mitotic index

p-values (three of the classes contained only two compounds, so the *t*-test cannot be used).

The distance matrix of pairwise MPC similarities (clustered by their Euclidean distances) is shown in Fig. 5 along with six highlighted clusters of compounds and their annotated functions/targets. Cluster i contains actin poisons as well as the myosin II inhibitor blebbistatin, and myosin light chain kinase (MLCK) inhibitors ML9 and A-3. Cluster ii contains the phosphodiesterase inhibitors. Cluster iii contains the MT poisons as well as a number of compounds not annotated as MT poisons but that show distinct anti-mitotic phenotypes. The supercluster that comprises clusters iv, v and vi contain primarily agents that cause a G1-S phase arrest by various mechanisms. Clusters iv and vi contain compounds that impinge directly or indirectly on protein synthesis, including DNA anti-metabolites, topoisomerase poisons, transcription inhibitors, and translation inhibitors. In the middle of this supercluster is cluster v, which contains several known electrophilic proteasome inhibitors including MG-132. The full heat map of compounds clustered by their CP fingerprints can be found in Fig. S1 and Table S1 (ESI[†]).

CP feature classes that distinguish compound classes

Based on inspection of the CP fingerprints and corresponding microscopy images, we were able to identify features that distinguish the DNA damaging agents from other agents that cause G1/S cell cycle arrest (*e.g.*, protease inhibitors), and the MT stabilizers from other anti-mitotic agents (*e.g.* HSP-90 inhibitors). Further, K-S analysis of the within- vs. between-class compound pairs showed that for 17 compound classes, CP

was able to accurately differentiate compounds by class (Table 1). We also undertook a more general and quantitative approach for determining which cytological phenotypes underlie the ability to differentiate any compound class from the rest (or one class from another). We grouped the CP features into 23 natural “feature categories” based on the broad phenotypic signatures on which they report (*e.g.*, nuclear shape, pH3 intensity, cell count, mitotic index, *etc.*). Then we used the technique of “gene sets enrichment analysis” (GSEA)^{17,18} to determine which feature categories are enriched in features whose values correlate strongly with a particular compound class. GSEA was developed to identify *gene sets* (*i.e.*, sets of genes grouped by function, biological process, co-regulation, *etc.*) that are enriched in *genes* whose up- or down-regulation are correlated with a particular *phenotype* (*e.g.*, of metastatic vs. non-metastatic) among a set of *samples* (*e.g.* tissue samples from different patients). Applying GSEA to our dataset, we replaced *genes* with CP features, *gene sets* with feature categories, *phenotypes* with annotated MOA classes, and *samples* with compound-dose instances. This allowed us to ask, for each annotated compound class, what CP feature categories correlate significantly with the distinction between that MOA and the rest of the library (Table 1 and Table S2, ESI[†]). We also used GSEA to identify CP feature categories that differentiate between two classes of compounds with similar overall phenotypes, for example, between DNA synthesis inhibitors and protease inhibitors, and between MT poisons from HSP90 inhibitors. This analysis revealed the same sets of features that were identified by visually comparing fingerprints and microscopy images (as in Fig. 3a and b), and also helped identify feature categories that were not found in the manual analysis (*e.g.* binucleate cell features that

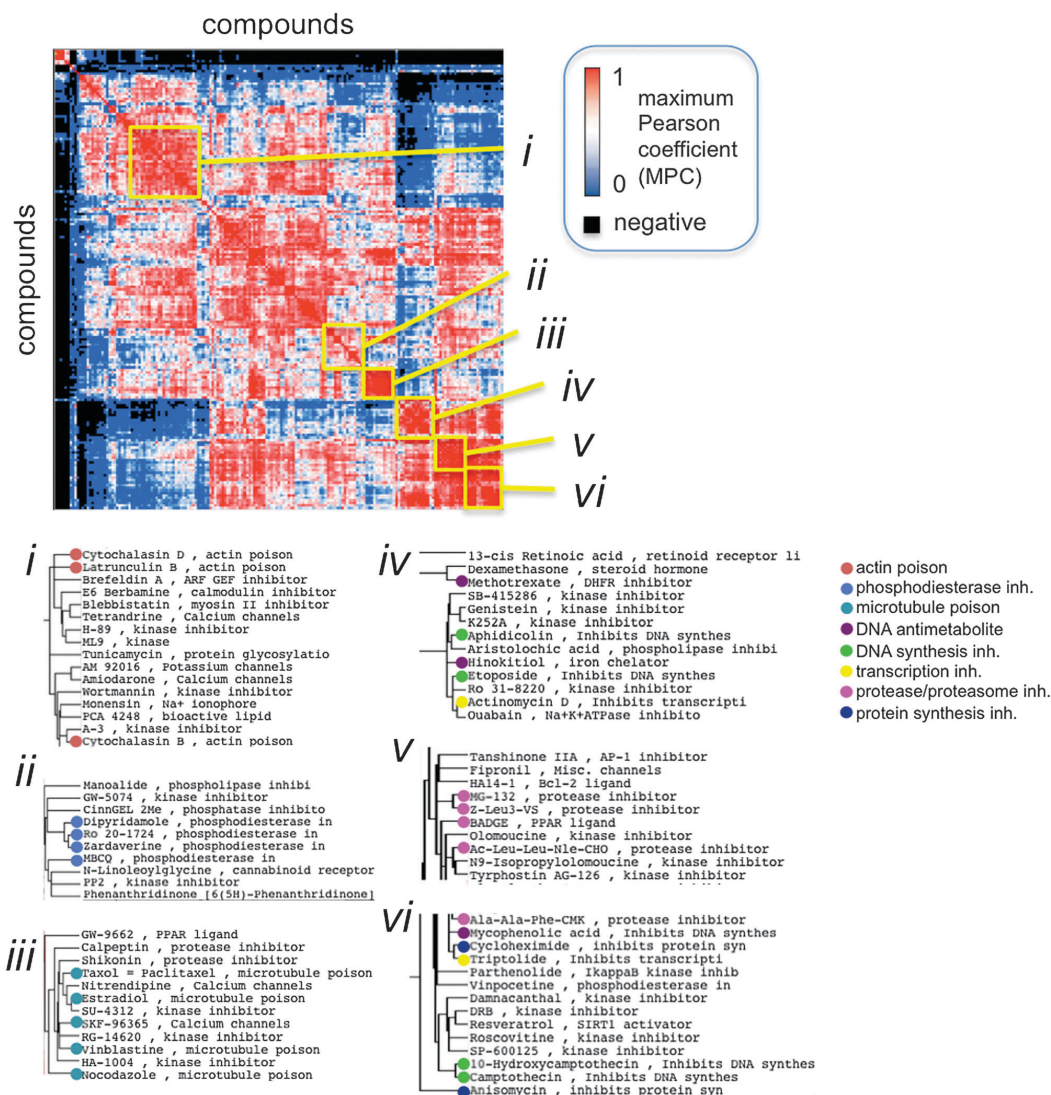


Fig. 5 Distance matrix of maximum Pearson coefficient (MPC) scores between all pairs of compounds that passed the weak phenotype/toxicity tests. Six of the major clusters are shown representing many of the MOA classes in the library.

distinguish between MT poisons and HSP90 inhibitors; Fig. 3d). We have made the full GSEA analysis for all compound classes vs. the rest of the library available, as well as the input data needed to perform GSEA on any pair of compounds (Table S2, ESI†).

Predicted compound class memberships of known and unanticipated MOAs

We determined which compounds were highly similar to other compounds of their expected class and, conversely, those with unanticipated similarity to compounds outside the expected class.

To do this, we calculated a *mode-of-action affinity rank (MAR)* score to measure the extent to which a compound's HD-based fingerprint was similar to other compounds in its own class relative to compounds of a different class (see Methods). First, the centroid of each class was calculated based on the CP fingerprints of all annotated members of that class. The MAR score for each compound was then calculated as the normalized rank of the class annotated for that compound among all classes, ordered by the distance from that compound's fingerprint to each class's centroid (normalized from 1 to 0). Out of 315 compound-dose

instances in classes containing more than one compound, 106 (34%) had MAR scores of 1, indicating that their CP fingerprints are closest to the centroid of their annotated classes than to those of any other class. For compounds whose MAR scores were less than 1, higher-ranking classes (*i.e.*, whose centroids are closer to the compound's CP fingerprint than it's annotated class) provided hypotheses into alternate MOAs for that compound. 209 such unanticipated associations were uncovered (Table S3, ESI[†]).

Validation of predictions for unanticipated MOAs

We followed up on three groups of compounds where several compounds were predicted to be more similar to compounds of another class than the anticipated class (MAR scores less than 1 and/or clustered well outside their annotated class; Fig. 6). First, the known microtubule (MT) poisons nocodazole, vinblastine, and taxol all had MAR scores of 1. There were also a number of compounds in this cluster that were not annotated as MT poisons, including the SU-4312 (VEGF inhibitor), SKF-96365 (calcium channel blocker), GW-9662 (PPAR ligand), RG-14620 (kinase inhibitor) and nitrendipine (calcium channel blocker), whose fingerprints had much higher similarities to the known MT poisons than to other compounds in their classes. Although SKF-96365 has been widely used as a calcium

channel blocker, one study found that the compound also depolymerizes MTs *in vitro*.¹⁹ We tested the other compounds in this series for their effects on MT polymerization *in vitro* and found that SU-4312 inhibits polymerization significantly at 90 μM , while nitrendipine and RG-14620 had no effect (Fig. 6a). Compound GW-9662 showed a small but significant increase in MT polymerization at a high concentration (180 μM).

Second, the "tyrphostins" are a class of broad-spectrum tyrosine kinase inhibitors of moderate potency and selectivity, which include tyrphostin 8 and AG-879. Many tyrphostins are electron deficient phenols, which can act as proton shuttles and uncouple mitochondrial respiration from ATP synthesis. Indeed, a number of tyrphostins have been shown to act as uncouplers,²⁰ and not surprisingly many of these compounds in the library cluster with the known uncouplers FCCP and valinomycin, both of which have MAR scores of 1. Among this cluster was AG-879, a tyrphostin that, while annotated as a kinase inhibitor, also contains an electron-deficient hydroxyl group. In a standard assay for uncoupling that measures the rate of O₂ consumption (QO₂) in isolated rat mitochondria, we found that AG-879 showed a QO₂ similar to that of FCCP (Fig. 6b). By contrast, tyrphostin 1, which lacks electron-withdrawing substituents on the benzene ring and does not cluster with the uncouplers in the CP analysis, had no effect on QO₂ in mitochondria.

Third, the commercial compound bisphenol A diglycidyl ether (BADGE), a component of epoxy resins and a common environmental contaminant found in plastic beverage containers, clustered with known proteasome inhibitors such as MG-132 and Ac-Leu-Leu-Nle-CHO. BADGE is annotated in the library as a ligand of peroxisome proliferator-activated receptor (PPAR) gamma, based on reports that it binds to this receptor and blocks PPAR-gamma-induced adipocyte differentiation.²¹ Consistent with the similarity of its CP fingerprint to those of known protease inhibitors, we found that BADGE inhibits proteasome activity in a fluorescence-based *in vitro* assay. Although its EC₅₀ in the proteasome assay was high (~100 μM), this is similar to the IC₅₀ reported for the interaction of BADGE with PPAR-gamma (Fig. 6c).

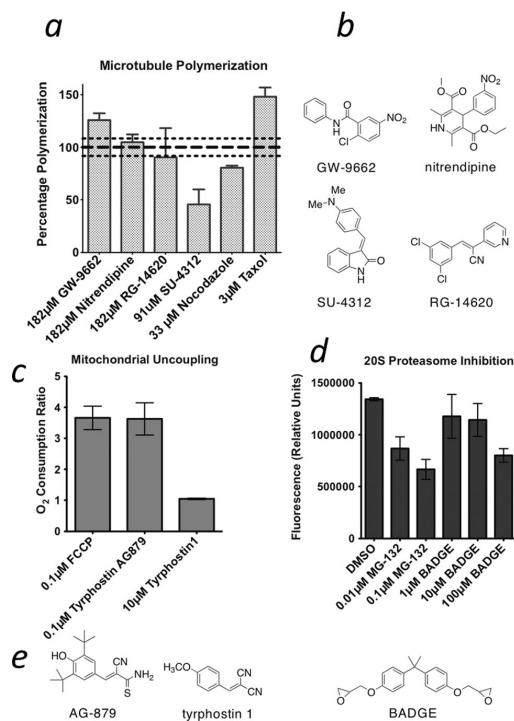


Fig. 6 Follow-up studies to investigate predicted activities of selected compounds based on their CP profiles. (a) *In vitro* MT polymerization assay. (b) Structures of compounds used in (a). (c) Mitochondrial uncoupling assay. (d) Proteasome inhibition assay. (e) Structures of compounds tested in (c) and (d).

A kinase-feature network

Kinases play central roles in most cellular processes and are important drug targets in many diseases, from autoimmune disorders to cancer.²² And yet, because most known kinase inhibitors have relatively broad specificity profiles, their utility as tools to probe the biology of individual kinases has often been called into question.^{23–26} Even the most specific inhibitors target several or more kinases in the cell,²⁶ thus driving the development of chemical genetic tools for probing the role of kinases in biology.^{27,28} Here we use maximal information statistics to combine the cytological profiles of 24 kinase inhibitors from the ICCB library with their inhibition profiles against a panel of kinases, resulting in a network connecting kinases to cytological features. The results suggest a method for using the combined data from many non-specific kinase

inhibitors to deduce functional phenotypic information on individual kinases.

To map the visible cellular-level changes induced upon treatment of particular kinase inhibitors, we selected 24 of the 44 kinase inhibitors from the ICCB reference library for which quantitative, systematically determined kinase inhibition data could be found.^{23–25} For most of these early generation kinase inhibitors, their specificities were quite broad even among the relatively small panel of 24 kinases tested in these studies. Still, we hypothesized that by joining CP features and kinase inhibitory activities into the same dataset, kinases and features could be connected through the aggregate data from all of the inhibitors. In this way, cellular phenotypes can be ascribed to individual kinases through correlations in their perturbations over the entire panel of kinase inhibitors, independent of the specificity of any individual inhibitor.

Before merging the datasets, we determined whether kinase specificity correlated significantly with the CP profiles of the 24 kinase inhibitors by comparing the pairwise Pearson coefficients between the two datasets. In particular, we clustered the

symmetric distance matrix of pairwise Pearson correlation coefficients for each pair of inhibitors (at their maximum-correlation doses) with respect to their CP fingerprints (Fig. 7a). Then, using the same ordering of compounds, we created a distance matrix based on these compounds' published kinase specificity profiles (Fig. 7b). The resulting heat map shows two clusters of similarity, one around kinases ML9, ML7 and Y27632, and another one around PP1, PP2, SB203580, and SB202190. The significance of the observed correlation between CP phenotype and kinase specificity (p -value = 0.02).

Next, we merged the CP and kinase inhibition datasets for the 24 inhibitors and determined the maximal information coefficient (MIC) score between all pairs of CP features and kinase IC₅₀ values across the entire dataset using MINE.²⁹ The MIC is a nonparametric statistic for exploring complex functional relationships between variables in large datasets. We identified three subnetworks connecting kinases with CP features in which the pairs of nodes (kinases or features) had MIC similarity scores of 0.9 or greater (p -value = 0.0002 (Fig. S3, ESI†)) (Fig. 7c). The largest of these networks contained AMPK

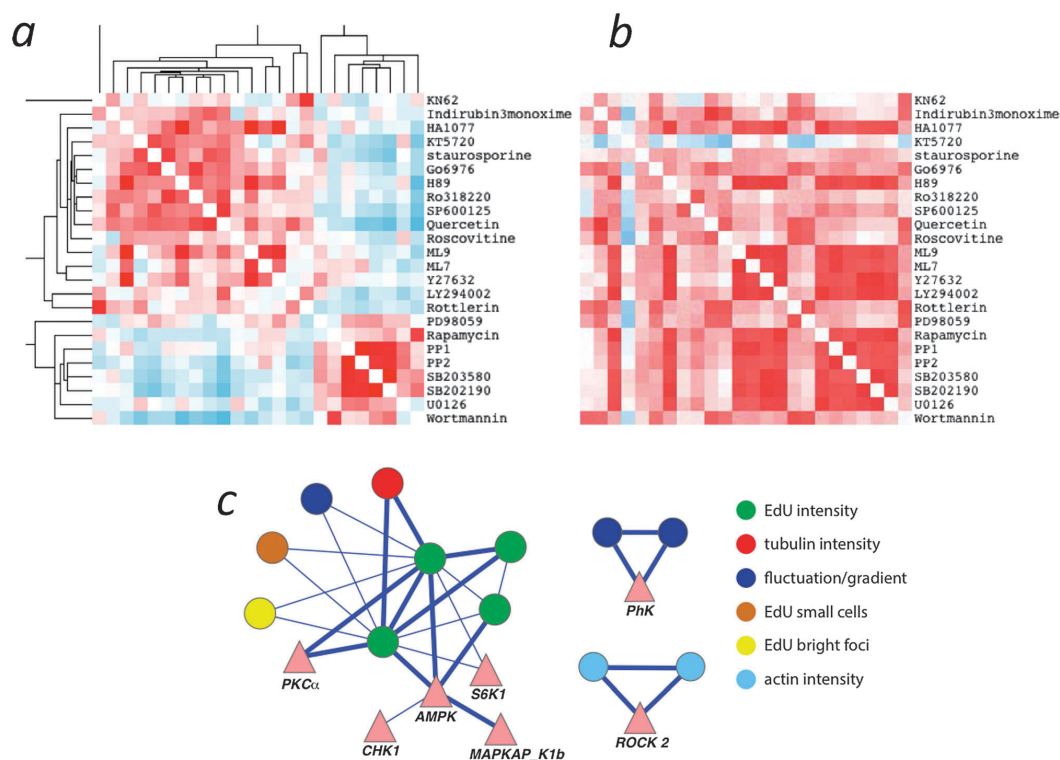


Fig. 7 Connecting kinases to cytological features by comparing *in vitro* kinase inhibition data with CP fingerprints using maximal information coefficient (MIC) analysis. (a) Pairwise clustering of kinase inhibitors by the similarities of their cytological profiles. (b) The pairwise similarities of the same inhibitors against a panel of kinases (published data), in the same order derived from the clustering of the CP similarities. (c) Networks linking kinases (triangles) and CP features (circles) resulting from MIC analysis of feature and inhibition data from 24 of the kinase inhibitors in the ICCB collection. Edges represent MIC scores greater than 0.9, and edge width corresponds to MIC score with the thickest lines representing scores of 1.0. Color coding of features is the same as in Fig. 6.

and PKC α as the major kinase nodes, both of which were strongly connected to features related to EdU intensity (MIC score = 1.0, p -value < 0.00001). These features have strong negative values for compounds that cause G1- and/or S-phase cell cycle arrest such as the DNA damaging agents and anti-metabolites. Although PKC isoforms have a multitude of cell cycle effects depending on the isoform and cell type, a strong link has been established between PKC activity and the phosphorylation state and function of the cell cycle regulatory protein pRb, thereby implicating PKC as a major player in the G1-S transition.³⁰ Likewise, the other major kinase node in this network, AMP-activated protein kinase (AMPK), is involved at the G1-S transition as part of the energy-sensing metabolic checkpoint.^{31,32}

The other two networks revealed by this analysis connect phosphorylase kinase (PhK) with texture- and gradient-related cytological features, and Rho-associated kinase (ROCK) II with actin intensity features. While we can find no obvious connection between the known biological functions of PhK and the phenotypic variables most associated with texture and gradient features (*e.g.* cytoskeletal morphology), the connection between ROCK II and actin is firmly established. The ROCK kinases play major roles in processes related to actin structure and dynamics, including cell migration, maintenance of cell polarity, stress fiber formation, and cytokinesis (reviewed in ref. 33). It is noteworthy that visual inspection of the kinase specificity data for ROCK II alone does not point toward actin features in particular; indeed, it is only through MIC analysis of the entire dataset that these relationships are revealed.

Discussion

This report describes a cytological profiling screening pipeline based on automated microscopy and image analysis. At the heart of the approach we describe a general purpose dimensionless score to quantify cytological features for use as fingerprints to study small molecule MOAs. Application of a concentration-independent correlation approach showed that compounds with the same or related MOAs had similar CP fingerprints. For 15 distinct annotated MOA classes, a t -test or Kolmogorov-Smirnov distributional test showed significantly higher pairwise similarities among compounds within the same class than between-class pairs. Further, novel associations could be identified using a class-affinity MAR score, which revealed testable predictions of novel mechanistic classes for several compounds.

The MT poisons formed a tight cluster whose fingerprints were dominated by an increase in the number of mitotic cells and a decrease in the intensity of tubulin staining. Clustering with the known MT poisons was SKF-96365, which, although annotated as a calcium channel blocker, has also been shown to inhibit MT polymerization *in vitro*.¹⁹ These authors suggest that the MT depolymerizing activity of SKF-96365 likely accounts for the observation that in addition to blocking calcium influx, SKF-96365 also causes a mitotic arrest in leukemia cells.³⁴ While nitrendipine and RG-14620 also clustered tightly with the MT

poisons, they showed no activity against MTs *in vitro*. This suggests that these compounds may be acting on the mitotic apparatus or spindle assembly checkpoint by a mechanism that does not involve direct interaction with microtubules. Nitrendipine is a member of the dihydropyridine class of calcium channel blockers and is a widely prescribed antihypertensive agent. The mitotic block induced by nitrendipine could be a direct result of its inhibition of L-type calcium channels, although it may have an additional target or targets that underlie its antimitotic phenotype. Likewise, while RG-14620 is annotated as a protein tyrosine kinase inhibitor, given the lack of specificity characteristic of this and similar typhostins it seems likely that its antimitotic phenotype is the result of its inhibition of another target, possibly a kinase involved in the metaphase-to-anaphase transition. It is noteworthy that closer inspection of the images revealed that, like the HSP-90 inhibitors, this compound arrests cells in mitosis with their spindles intact and metaphase plates properly aligned, suggesting that its effect is downstream of microtubule polymerization.

Our analysis also implicated the environmental toxin BADGE as a proteasome inhibitor, in addition to its annotated activity as a PPAR-gamma inhibitor. BADGE has been shown to antagonize PPAR-gamma *in vitro* (with an IC₅₀ of ~100 μ M), and also to inhibit adipogenesis in cells,²¹ a phenotype which was assumed to relate to its PPAR-gamma-inhibitory activity. However, in another study BADGE was found to induce PPAR-gamma activation resulting in localization of the receptor to the nucleus and subsequent apoptosis.³⁵ BADGE has also been shown to induce apoptosis in a PPAR-gamma-independent fashion. Here we found that, in HeLa cells, BADGE clustered with proteasome inhibitors rather than other known PPAR-gamma ligands, *e.g.* ciglitazone. While this and previously observed discrepancies in the MOA of BADGE may simply be due to the different cell lines used in these studies, the previously observed effects on adipogenesis (and apoptosis) upon BADGE treatment could also be a result of its proteasome-inhibitory activity rather than specific binding to the PPAR-gamma receptor. Ligand-induced activation of many nuclear hormone receptors, including PPAR-gamma, is tightly associated with receptor degradation by the proteasome. Paradoxically, interference in this feedback loop by proteasome inhibitors can decrease transcriptional activity of these receptors, possibly due to the requirement for continuous receptor turnover in the maintenance of transcriptional activation.^{36,37} Indeed, inhibition of the proteasome by lactacystin decreases adipogenesis³⁸ in some cell lines, although whether the effect of BADGE on adipogenesis is related to its proteasome-inhibitory activity remains to be tested.

We also found that the electron-deficient phenolic typhostin, AG-879, acts as a mitochondrial uncoupler, consistent with the similarity to other known uncouplers in the reference library. Other compounds in this cluster include the well known uncouplers FCCP and valinomycin, in addition to other compounds that were annotated with various other MOAs but had also shown uncoupling activity in other published studies.^{20,39} Our study supports the conclusion from these studies

that the use of these and compounds as kinase inhibitors in cells must be approached with caution, since their phenotypic effects are driven primarily by their mitochondrial uncoupling activity rather than their annotated kinase inhibitory activities.

To investigate the subtle signals in the CP data that help distinguish a compound class from all other compounds we adapted Gene Set Enrichment Analysis (GSEA). By creating groups of related CP features we were able to produce a map of CP features associated with the compound classes. For example, nuclear area distinguishes DNA synthesis inhibitors whereas a combination of binucleated cell, pH3 intensity, cell count, and actin intensity features distinguish the phosphodiesterase inhibitors. The analysis was also able to reveal features that could help distinguish between any two classes. For example, the protease inhibitors can be distinguished from DNA synthesis inhibitors by considering, in part, the pH3 intensity in large areas. This information could be used, in principle, to screen a larger library for compounds of a desired class or property by selecting only those few discriminative CP features of interest.

The 44 compounds annotated as kinase inhibitors in the ICCB library are primarily 1st-generation compounds whose selectivities for individual kinases are modest at best. Indeed, a series of studies that tested 24 of these inhibitors in a panel of 24 kinases showed that most of them had very broad specificities, with none being highly selective for any one kinase. However, using the maximum information coefficient (MIC) analysis we were able to identify correlations between the CP fingerprints of these 24 inhibitors with their kinase specificity profiles, allowing us to identify networks connecting specific kinases to cytological phenotypes. Only a handful of the 24 kinases met the stringent threshold for similarity ($MIC > 0.9$) that we set for inclusion in the network. Among these, we correctly identified the connection between ROCK II and actin-related features, and between PKC and AMPK and EdU features that relate to G1-S cell cycle progression. Expansion of this dataset to include more kinase inhibitors and more cytological features would presumably increase the resolving power of the MIC analysis and allow us to develop more robust and inclusive networks relating specific kinases to cellular phenotypes. Importantly, the specificity of any given inhibitor included in this analysis need not be high; rather, the specificity profiles among the inhibitors should be diverse enough to capture some information for each kinase that, in aggregate, could generate meaningful associations between kinases and phenotypic signatures. The use of large-scale "kinome-wide" binding⁴⁰ and inhibition²⁶ data should also greatly expand the power of this type of analysis. Inclusion of additional times, concentrations, cell lines, and perturbations would also likely increase the number of kinases that pass the stringency test.

Conclusions

The automation of microscopy-based observations offers a powerful method for quantifying and classifying the action of perturbagens using the observable affects they induce on cellular

level phenotypes. We describe a pipeline useful for quantifying various effects, made possible by a general-purpose dimensionless score applicable to a wide range of high-content microscopy measures. The advance lays the groundwork to support an expanding list of cellular and sub-cellular phenomena quantifiable by HT microscopy using new lenses, stains, and cell cycle blocks. The adoption of such approaches will significantly impact the growth of CP-based screening into an established functional and chemical genomics platform for biological discovery. Incorporation of additional cellular models (new cell lines, xenografts, and tissue cultures), cellular perturbations such as RNA-interference technologies or ligand presentations or combinations of such exposures, will help shed light on drug MOA for the manipulation or treatment of cellular- and tissue-level outcomes.

Materials and methods

Robotic assay

HeLa cells were plated into two clear-bottom 384-well plates (Corning) at a density of 2500 cells per well in 25 μ l Dulbecco's modified Eagle's medium (DMEM) with 10% Fetal Bovine Serum. The plates were then placed in a 37 °C incubator under 5% CO₂. After 24 h of incubation 150 nl of compound was then added to the wells using a pin-tool attachment on the Janus MDT (PerkinElmer). The plates were incubated once more at 37 °C under 5% CO₂ for 19 h. After the 19 h incubation each plate was stained with a different stain set a nuclear stain set or a cytoskeletal stain set. The nuclear stain set is given pulse of the thymidine analog 5-ethynyl-2'-deoxyuridine (EdU) for 1 h at 37 °C. Cells were fixed with 4% formaldehyde for 20 min, washed with PBS, treated with 0.5% Triton X-100 for 10 min, and washed with PBS. Plates were blocked with a 2% BSA PBS solution for 20 min, washed with PBS, and stained with rhodamine-azide using click chemistry by incubation with 4 mM CuSO₄, 2 mg ml⁻¹ sodium ascorbate, and 1 mg ml⁻¹ rhodamine azide in 100 mM TRIS buffer for 30 min at 25 °C in the dark. After washing with PBS, rabbit antiphospho-histone H3 (Millipore) in 2% BSA PBS was added and the plate was incubated overnight at 4 °C. Plates were rinsed with PBS, and Cy5 conjugated goat anti-rabbit antibody and Hoechst stain were added for 1 h in 2% BSA in PBS at 25 °C. The plates were then left in 0.1% azide in PBS. For the cytoskeletal stain set, cells were fixed with 4% formaldehyde for 20 min, then washed with PBS. After treatment with 0.5% Triton X-100 for 10 min, the plates were washed with PBS and blocked with a 2% BSA PBS solution. After washing with PBS, plates were incubated overnight at 4 °C with FITC-conjugated mouse anti-(α -tubulin and rabbit antiphospho-histone H3 antibodies in 2% BSA PBS. After washing with PBS, Cy5-conjugated goat anti-rabbit antibody, rhodamine-labeled phalloidin, and Hoechst stain in 2% BSA PBS were added and plates were incubated for 1 h at 25 °C. The plates were then left in 0.1% azide in PBS.

Microscopy technique

Plates were imaged using a 10 \times Nikon objective on an ImageXpress Micro epifluorescence microscope (Molecular Devices).

We captured four images per well for each wavelength in a plate resulting in 4608 images for the nuclear stain set and 6144 images for the cytoskeletal stain set. Images were analyzed with MetaXpress 3.1 software (Molecular Devices). Measurements were taken using built-in morphometry metrics, the multi-wavelength cell scoring, transfluor, and micronuclei modules.

Image feature quantification

Cells exposed to a compound at a particular concentration were fixed and post-processed for automated microscopy to record various imaging features $i = 1 \dots I$. For example, the imaging feature Nuclear_Total_Area_Transfluor_EdU is the total area of the nuclei in a well of a nuclear stain set plate as measured by the Transfluor module. Each experiment captures a “snapshot” of a whole population of cells, producing a collection of measurements for each imaging feature. We denote the set of values detected for feature i in the experiment as, where K is the total number of values detected in the experiment. In order to detect concentration-dependent modes of action we treat each concentration of a compound separately and so would have a different set E for each different compound concentration. In what follows we use the term compound as a short-hand for a distinct compound-dose instance.

Filtering CP features based on biological reproducibility

Cellular properties contain both technical and biological sources of noise that can lead to feature-specific variability in their detection and quantization. To identify CP features providing measurements with high signal-to-noise, we measured the variability of each feature across biological replicates. Specifically, we took the two highest concentrations of each compound and repeated the experiment on two different days for each of the two compound concentrations, generating a technical replication of the experiment. For each CP feature, we collected all of the measurements across all of the compounds, forming a vector of results collected on the first day and another vector for the second day. We calculated the Pearson correlation between these two vectors for each CP feature and eliminated any CP feature with a correlation lower than 0.5.

We were interested in quantifying the differential affect that a compound has on cells as measured by a single imaging feature. Therefore, a negative control set of values was also collected for every experiment in which only buffer was added to the cells. We denote this background set of values detected in buffer as, where L is the total number of levels recorded by imaging the cell population in buffer. Note that L does not in general equal K because the number of quantified points in the experiment and control will not be the same. One could imagine computing any number of statistics to contrast the two sets of values collected in $E(i)$ against those collected in $B(i)$. For example, a two-sample unpaired t -test could be used to contrast, the mean of the experimental population of values, to, the mean of the background set of values that would factor in an estimate for the standard errors. This approach would work well for those particular imaging features that are continuous-valued and well approximated by Normal distributions. However, because

we require a method that can be applied to hundreds of different imaging features, some of which are continuous, some ordinal, and others discrete, parametric statistical tests, like the t -test, that assume particular distributions will not work in general. We therefore describe an *ad hoc*, non-parametric method we call histogram difference (HD) that calculates the overall dissimilarity between empirically estimated distributions for $E(i)$ and $B(i)$. The approach is similar in motivation and formulation to other empirical approaches such as the Kolmogorov–Smirnov test (KS), which measures the maximum differences between two cumulative distributions. HD produced higher correlations for unique compounds within a class as compared to KS (data now shown).

Cytological profiling fingerprints

We developed a dimensionless scoring method to contrast any of various features measured in the presence of a compound relative to their measures detected in background control. For every CP image feature i and every compound, a histogram $H_{E,i}$ was constructed by calculating the proportion of values in $E(i)$ observed in the presence of the compound falling within a set of Q equally spaced quanta. For example, for $i = \text{cellular_nuclei}$ $H_{E,i}(2)$ would reflect the number of cells detected to have nuclei in the 1/20th to the 1/10th highest range. To decrease the random fluctuations in the estimates due to small sample sizes the counts in each quanta were updated by averaging-in the counts from neighboring quanta. A kernel smoothing approach was used in which the counts found for quanta q were corrected by taking a running weighted average between q and all other quanta q' in which the weights were proportional to the exponentially dampened distance between the centers of quanta q and q' . The smoothed count estimates can then be expressed as:

$$H_{E,i}'(q) = \sum_{q'=1}^Q 2^{-\lambda|q-q'|} \cdot H_{E,i}(q')$$

where $H_{E,i}(q)$ is the number of items with a feature score in the range of quanta q , λ is a bandwidth parameter controlling how much impact neighboring quanta have on the center quanta, and $D(q,q')$ is the distance between the centers of the two quanta. This formulation allows for operations on histograms with unequally distributed centers. However, in our application, we used equal spaced quanta. This allows the distances between quanta to be expressed as a simple difference between the indices of the quanta themselves, *i.e.* $D(q,q') = \beta|qq'|$ for some appropriately chosen constant β . The β parameter reflects the range of the data for the particular cytological profiling feature. Setting a new parameter $\alpha = \lambda \times \beta \times \ln(2)$ then gives the simplified smoothing operation:

$$H_{E,i}'(q) = \sum_{q'=1}^Q e^{-\alpha D(q,q')} \cdot H_{E,i}(q')$$

Letting $\alpha = 1$ is equivalent to letting a quanta have half of the weight influence its own smoothed estimate and the other half derived from all of the other 19 quanta. In this study, we set the center quanta's weight to 2, its two immediate neighbors to 1,

and the rest of the neighbors to 0, which closely approximates this function and provides fast processing of the data. Using a completely analogous procedure, a smoothed histogram for the background set was also calculated and denoted $H_{B,i}$.

A dimensionless score for feature i could then be calculated by computing a directional difference between the two smoothed histograms:

$$F(i) = (-1)^{\bar{B}_i - \bar{E}_i} \sum_{q=1}^Q (H_{E,i}(q) - H_{B,i}(q))^2$$

where \bar{E}_i and \bar{B}_i are the average levels detected in the experiment $E(i)$ and buffer $B(i)$ as described previously. This score becomes more extreme as the $E(i)$ and $B(i)$ distributions separate and take on positive scores when $E(i)$ is higher than $B(i)$ on average.

CP fingerprints were formed by collecting a compound's HD scores across all of the features into a single fingerprint vector:

$$F = [F(1), F(2), \dots, F(m)]^T,$$

where m is the number of total CP imaging features measured and T indicates vector transpose.

Mode-of-action affinity rank (MAR) score

We used a non-parametric measure of concordance to evaluate the ability of CP fingerprints to relate compounds of like mode-of-action class. For each class, we calculated the class' centroid by averaging all CP fingerprints across compounds in the class. Then, for each compound, we calculated the Pearson correlation between the compound's CP fingerprint and all classes' centroids. After sorting all classes by their similarity to the compound we recorded the rank of the compound's class relative to other classes and then computed a rank score $s = (r - 1)/(K - 1)$ where K is the number of total classes and r is the rank of the known class. The score assigns values on a scale from 0 (lowest, most dissimilar rank) to +1 (highest, most similar rank). A value of 0.5 corresponds to a compound that has a similarity to its class at the median level, with half of the classes achieving higher similarity levels.

Within- versus between-class comparisons using the *t*-test and Kolmogorov-Smirnov test

The amount of information in the CP fingerprints were estimated by comparing the similarities of compounds that are in the same class vs. similarities of compounds that are in different classes. Pearson correlations for each pair of compounds were calculated and then grouped according to their class membership. Each Pearson correlation r was transformed using the Fisher Z-transform $z = 1/2 \ln((1 + r)/(1 - r))$ to give a more Normal-like distribution. Both the Student's *t*-test and the Kolmogorov-Smirnov (K-S) test were used to compare the within-class to the between-class correlations. The higher the *t*-test or K-S test value the larger the difference between the two distributions and the less likely each was drawn from the same distribution. Note that compounds which are identical but with different concentrations are not included in either set to avoid skewing the within-class distribution.

Mitochondrial respiration assay

The mitochondrial respiration assay was performed essentially as described in ref. 41. Mouse liver was homogenized in 10 mL isolation buffer (10 mM Tris-MOPS, 1 mM EGTA, 200 mM sucrose, pH 7.4). The resulting homogenate was centrifuged at 600g for 10 minutes at 4 °C. The supernatant was kept and centrifuged at 7000g for 10 minutes at 4 °C. The resulting supernatant as discarded and the pellet was resuspended in 5 mL isolation buffer followed by centrifugation at 7000g for 10 minutes at 4 °C. A Hansatech DW1 oxygen electrode chamber was used for all oxygraph experiments. The oxygraph was calibrated using N₂ prior to each experiment in 1 mL experimental buffer (125 mM KCl, 1 μM EGTA-Tris, 10 mM Pi, 1 mM Tris-MOPS) at 37 °C. After a stable baseline was achieved, 20 μL of the mitochondria extract was added and oxygen consumption recorded for 1-3 minutes until a stable rate was observed. 10 μL of compound was then added to the reaction vessel and the oxygen consumption rate was recorded until a stable rate was established.

Proteasome inhibition

20S proteasome activity was measured in HeLa cells using a kit from Cayman Chemicals. HeLa cells were plated in a clear bottom 96-well plate (Corning) and incubated at 37 °C under 5% CO₂ for 24 h. The cells were then treated with the compound and incubated for 20 h before lysis and addition of proteasome substrate. Cleaved substrate was measured by monitoring the fluorescence intensity of each well with an EnVision plate reader (PerkinElmer) at an excitation of 355 nm and an emission of 490 nm.

Tubulin polymerization assay

Inhibition of tubulin polymerization was assayed using a Tubulin Polymerization Kit from Cytoskeleton. Polymerization was measured in a 96-well plate using an EnVision plate reader (PerkinElmer) at an excitation of 360 nm and emission of 460. Tubulin was incubated in with compound at 37 °C for 2 h and fluorescence intensity was measured every 10 minutes during that incubation. Rate of polymerization was determined as the first derivative of the plot of fluorescence intensity vs. time. Percent polymerization was calculated as the rate of polymerization in the presence of compound divided by the rate of polymerization in the presence of DMSO multiplied by 100.

Acknowledgements

We acknowledge funding from NIH (R.S.L., 5R01GM084530), California QB3, and NSF CAREER (J.M.S.).

References

- 1 T. J. Mitchison, *ChemBioChem*, 2005, **6**, 33–39.
- 2 Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu and S. J. Altschuler, *Science*, 2004, **306**, 1194–1198.
- 3 J. Lorang and R. W. King, *Genome Biol.*, 2005, **6**, 228.
- 4 D. W. Young, A. Bender, J. Hoyt, E. McWhinnie, G. W. Chirn, C. Y. Tao, J. A. Tallarico, M. Labow, J. L. Jenkins, T. J. Mitchison and Y. Feng, *Nat. Chem. Biol.*, 2008, **4**, 59–68.

- 5 T. R. Jones, A. E. Carpenter, M. R. Lamprecht, J. Moffat, S. J. Silver, J. K. Grenier, A. B. Castoreno, U. S. Eggert, D. E. Root, P. Golland and D. M. Sabatini, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 1826–1831.
- 6 M. Tanaka, R. Bateman, D. Rauh, E. Vaisberg, S. Ramachandani, C. Zhang, K. C. Hansen, A. L. Burlingame, J. K. Trautman, K. M. Shokat and C. L. Adams, *PLoS Biol.*, 2005, **3**, e128.
- 7 J. J. Sutherland, J. Low, W. Blosser, M. Dowless, T. A. Engler and L. F. Stancato, *Mol. Cancer Ther.*, 2011, **10**, 242–254.
- 8 F. Gasparri, A. Ciavolella and A. Galvani, *Adv. Exp. Med. Biol.*, 2007, **604**, 137–148.
- 9 D. L. Towne, E. E. Nicholl, K. M. Comess, S. C. Galasinski, P. J. Hajduk and V. C. Abraham, *J. Biomol. Screening*, 2012, **17**, 1005–1017.
- 10 J. Low, A. Chakravarty, W. Blosser, M. Dowless, C. Chalfant, P. Bragger and L. Stancato, *Curr. Chem. Genomics*, 2009, **3**, 13–21.
- 11 K. C. Peach, W. M. Bray, N. J. Shikuma, N. C. Gassner, R. S. Lokey, F. H. Yildiz and R. G. Linington, *Mol. BioSyst.*, 2011, **7**, 1176–1184.
- 12 C. J. Schulze, W. M. Bray, M. H. Woerhmann, J. Stuart, R. S. Lokey and R. G. Linington, *Chem. Biol.*, 2013, **20**, 285–295.
- 13 L. H. Loo, L. F. Wu and S. J. Altschuler, *Nat. Methods*, 2007, **4**, 445–453.
- 14 A. Salic and T. J. Mitchison, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 2415–2420.
- 15 G. de Carcer, *Cancer Res.*, 2004, **64**, 5106–5112.
- 16 S. K. Lyman, S. C. Crawley, R. Gong, J. I. Adamkewicz, G. McGrath, J. Y. Chew, J. Choi, C. R. Holst, L. H. Goon, S. A. Detmer, J. Vaclavikova, M. E. Gerritsen and R. A. Blake, *PLoS One*, 2011, **6**, e17692.
- 17 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15545–15550.
- 18 V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler and L. C. Groop, *Nat. Genet.*, 2003, **34**, 267–273.
- 19 M. Mitsui-Saito, N. Nakahata and Y. Ohizumi, *Jpn. J. Pharmacol.*, 2000, **82**, 269–271.
- 20 S. P. Soltoff, *J. Biol. Chem.*, 2004, **279**, 10910–10918.
- 21 H. M. Wright, C. B. Clish, T. Mikami, S. Hauser, K. Yanagi, R. Hiramatsu, C. N. Serhan and B. M. Spiegelman, *J. Biol. Chem.*, 2000, **275**, 1873–1877.
- 22 P. Cohen, *Nat. Rev. Drug Discovery*, 2002, **1**, 309–315.
- 23 J. Bain, L. Plater, M. Elliott, N. Shpiro, C. J. Hastie, H. McLauchlan, I. Klevernic, J. S. Arthur, D. R. Alessi and P. Cohen, *Biochem. J.*, 2007, **408**, 297–315.
- 24 S. P. Davies, H. Reddy, M. Caivano and P. Cohen, *Biochem. J.*, 2000, **351**, 95–105.
- 25 J. Bain, H. McLauchlan, M. Elliott and P. Cohen, *Biochem. J.*, 2003, **371**, 199–204.
- 26 T. Anastassiadis, S. W. Deacon, K. Devarajan, H. Ma and J. R. Peterson, *Nat. Biotechnol.*, 2011, **29**, 1039–1045.
- 27 M. S. Cohen, C. Zhang, K. M. Shokat and J. Taunton, *Science*, 2005, **308**, 1318–1321.
- 28 K. M. Specht and K. M. Shokat, *Curr. Opin. Cell Biol.*, 2002, **14**, 155–159.
- 29 D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, *Science*, 2011, **334**, 1518–1524.
- 30 J. D. Black, *Front. Biosci.*, 2000, **5**, D406–D423.
- 31 R. G. Jones, D. R. Plas, S. Kubek, M. Buzzai, J. Mu, Y. Xu, M. J. Birnbaum and C. B. Thompson, *Mol. Cell*, 2005, **18**, 283–293.
- 32 J. Liang, S. H. Shao, Z. X. Xu, B. Hennessy, Z. Ding, M. Larrea, S. Kondo, D. J. Dumont, J. U. Gutterman, C. L. Walker, J. M. Slingerland and G. B. Mills, *Nat. Cell Biol.*, 2007, **9**, 218–224.
- 33 M. Amano, M. Nakayama and K. Kaibuchi, *Cytoskeleton*, 2010, **67**, 545–554.
- 34 T. Nordstrom, H. A. Nevanlinna and L. C. Andersson, *Exp. Cell Res.*, 1992, **202**, 487–494.
- 35 D. Bishop-Bailey, T. Hla and T. D. Warner, *Br. J. Pharmacol.*, 2000, **131**, 651–654.
- 36 A. P. Dennis, D. M. Lonard, Z. Nawaz and B. W. O'Malley, *J. Steroid Biochem. Mol. Biol.*, 2005, **94**, 337–346.
- 37 D. M. Lonard, Z. Nawaz, C. L. Smith and B. W. O'Malley, *Mol. Cell*, 2000, **5**, 939–948.
- 38 S. De Barros, A. Zakaroff-Girard, M. Lafontan, J. Galitzky and V. Bourlier, *J. Pharmacol. Exp. Ther.*, 2007, **320**, 291–299.
- 39 C. Vidau, R. A. Gonzalez-Polo, M. Niso-Santano, R. Gomez-Sanchez, J. M. Bravo-San Pedro, E. Pizarro-Estrella, R. Blasco, J. L. Brunet, L. P. Belzunces and J. M. Fuentes, *Neurotoxicology*, 2011, **32**, 935–943.
- 40 M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka and P. P. Zarrinkar, *Nat. Biotechnol.*, 2008, **26**, 127–132.
- 41 C. Frezza, S. Cipolat and L. Scorrano, *Nat. Protocols*, 2007, **2**, 287–295.

Supplemental Information.

Large-scale cytological profiling uncovers novel functions for hundreds of compounds

Marcos H. Woehrmann, Walter M. Bray, James K. Durbin, Sean C. Nisam, Alicia K. Michael, Emerson Glassey, Joshua M. Stuart, R. Scott Lokey

Inventory of Supplemental Information.

Figure S1. High-resolution heat map of cytological profiles clustered by similarity (Euclidean distance, average linkage). Related to Figure 2.

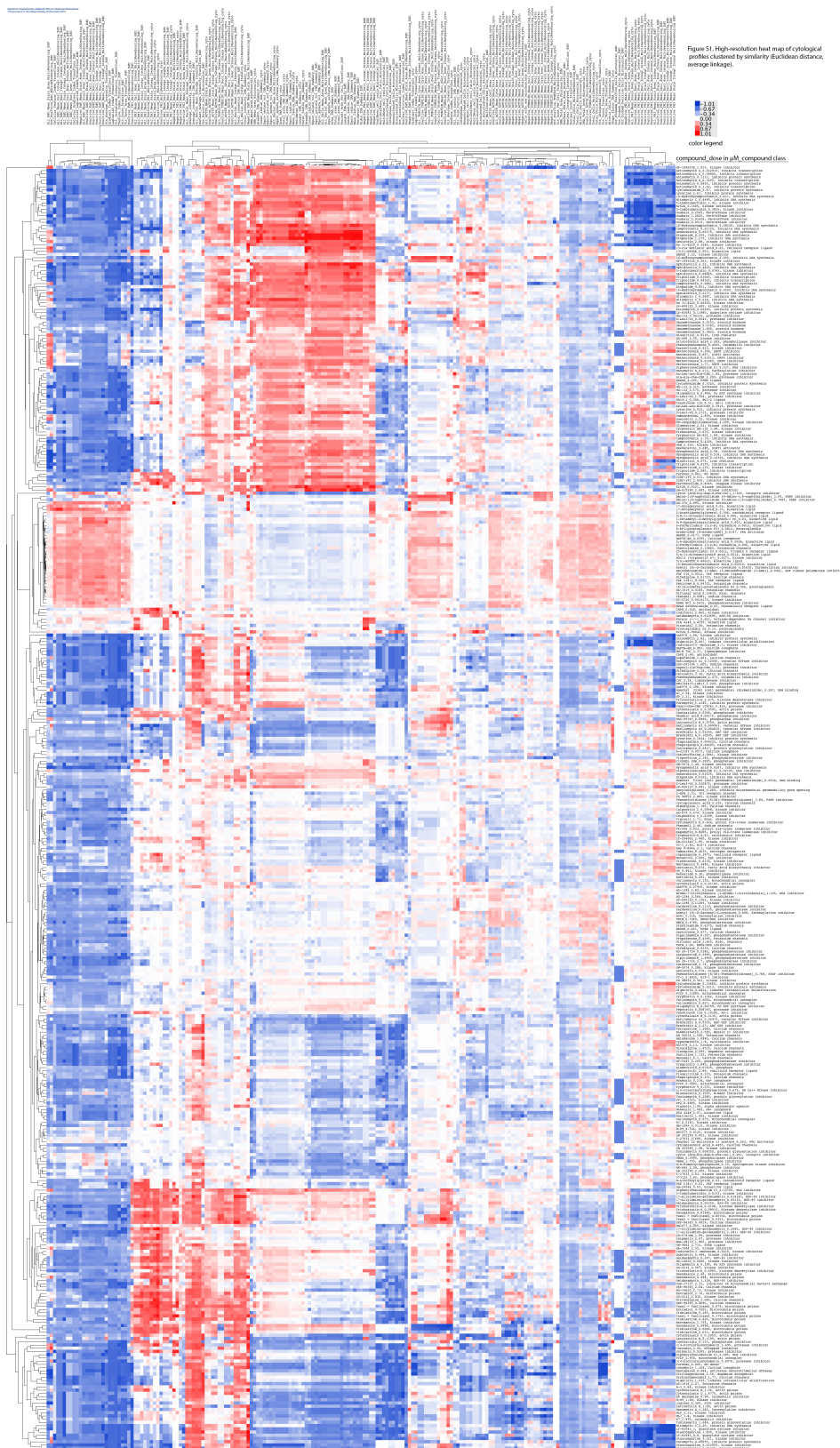
Table S1. Cytological profiling data (as HD scores) for 399 compound-dose instances after eliminating instances due to weak phenotypes and toxicity. List of cytological features, their corresponding feature categories, compounds, and corresponding compound classes. Related to Figure 2.

Table S2. Data used for input into GSEA analysis. Sheet 1: Cytological profiling data for input into GSEA (.gct file). Sheet 2: Compound classes (.cls file). Sheet 3: Cytological feature categories (.gmt file). Related to Figure 3.

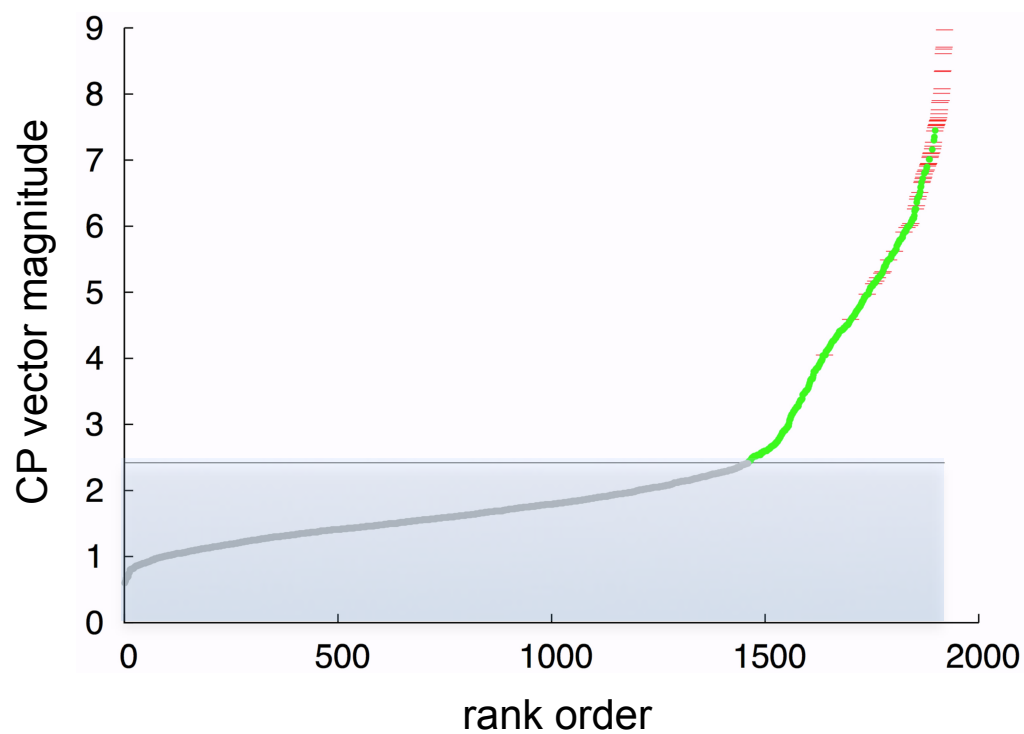
Figure S2. Chart showing compound-dose instances rank ordered by phenotypic vector magnitude, showing elimination of weak phenotypes and toxic instances. Related to Figure 5.

Table S3. Mode of Action Activity Rank (MAR) scores. Related to Figure 6.

The supplemental tables listed above can be found on the publisher's web site under DOI 10.1039/C3MB70245F.



Supplemental Figure 1. High-resolution heat map of cytological profiles clustered by similarity (Euclidean distance, average linkage).



Supplemental Figure 2. Elimination of weak phenotypes and toxic doses. The shaded region represents compound-dose instances whose CP vector magnitudes fell below the cutoff, set to 2.4 (460 instances). Red hashes represent instances that were eliminated due to toxicity tests (61 instances). Green marks represent instances that passed both the weakness and toxicity tests (399 instances; 222 compounds).

3 Screening in microbes

While cytological profiling has been shown to be a valuable method to predict drug target classes, it has drawbacks: it is costly, difficult to do correctly (even skilled technicians can have difficulty growing HeLa cells) there is a risk of contamination, and it is computationally expensive. These issues can be eliminated by screening in a microbial such as yeast. Yeast has been used in labs for many years and using it in assays is well understood. Yeast is inherently easier to grow than HeLa cells and as seen below it is possible to use a simpler assay when evaluating results; rather than using microscopy to examine individual cells the overall reduction in growth of the yeast colony is found by a simple measurement of optical density.

Though quite different in appearance, there is a large degree of conservation on the cellular level in basic biology between yeast and humans (Barberis, et al., 2005; Hughes, 2002; Mager and Winderickx, 2005; Parsons,

et al., 2003; Simon and Bedalov, 2004; Smith and Snyder, 2006; Steinmetz, et al., 2002).

3.1 D-Map

This section discusses my research in using chemical-chemical synthetic lethality to build fingerprints of known compounds and using these fingerprints to predict the mode of action of unknown compounds.

For this work I used *S. cerevisiae* (budding yeast) grown on an agar medium. There is a large degree of conservation in many essential pathways in eukaryotes including between yeast and human, allowing yeast to be used as a model organism in High-Throughput Screens (HTS) (Barberis, et al., 2005; Hughes, 2002; Mager and Winderickx, 2005; Parsons, et al., 2003; Simon and Bedalov, 2004; Smith and Snyder, 2006; Steinmetz, et al., 2002).

Screening of compounds for the ability to kill or inhibit the growth of microorganisms has long been done using a halo assay. A halo assay is done by soaking a small disk of filter paper in a compound under test which is then placed onto an agar solution containing a microbial. After a time the toxicity

of the compound can be determined by the area of inhibition (“halo”) seen around the filter paper (figure 5). Because the traditional halo assay is done by hand the limit on the number of experiments that can be performed is on the order of several dozen per day per technician. Additionally, the data is less quantitative and more qualitative; the size of the halo giving a general idea of the toxicity of the compound.

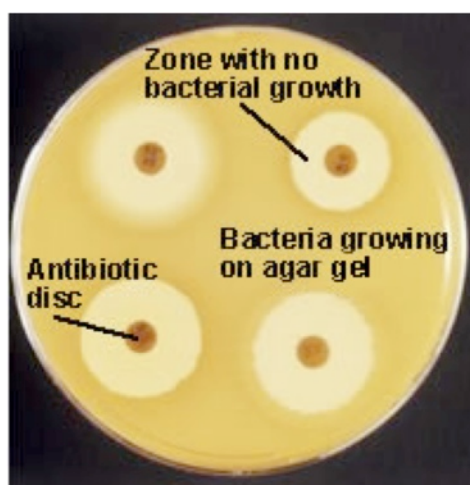


Figure 5 – Traditional Agar plate with filter discs that have been soaked in antibiotics. As the microorganism seeded into the agar grows it’s growth is inhibited by the antibiotic that diffuses out of the filter disc (ABPI, 2009).

The assay I used for D-Map is an automated, high-throughput yeast halo assay (Gassner, et al., 2007) that replaces the compound soaked filter discs with a small amount of up to 384 different compounds robotically transferred into the agar by hollow pins.

The set of pins is dipped into a 384 well plate where each well contains a compound dissolved in DMSO. A small amount of this solution is drawn

into the pin via capillary action; when the pins are then inserted into the agar the solution is dispensed. The pins are then cleaned and blotted in preparation for the next set of compounds. By making use of robots dozens of plates per day can be produced; allowing thousands of compounds to be tested.

After pinning the yeast is allowed to incubate at room temperature for 15 to 20 hours. The plates are then scanned by a digital imaging microscope, the amount of light being transmitted through the agar at each point giving a numerical indication of the toxicity of the compound (figure 6).

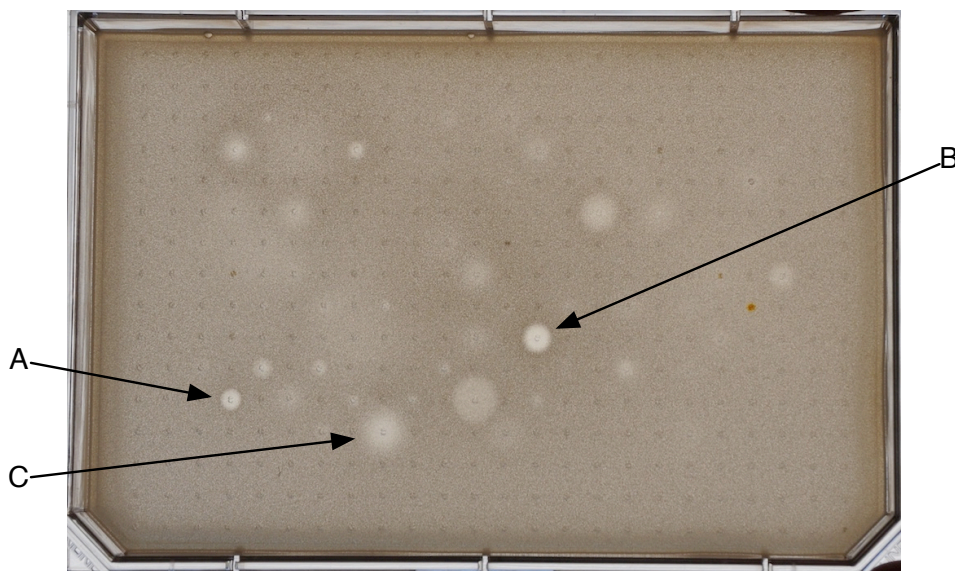


Figure 6 – Agar plate showing inhibition zones. Each of the barely visible small black dots is a location where a compound has been pinned (some locations are control spots and are pinned with only the solvent used to dissolve the compounds (DMSO)). Most pinning locations do not show any reduction in growth, some show a small, intense reaction and others a large, mild reaction. The red dot vertically centered and near the right hand edge of the plate is ay compound that is not optically clear. The letters are positions that are referenced in the text.

By adding an unknown compound to the agar as the plate is prepared, and comparing the results to a control plate, the amount of growth inhibition caused by the combination of each of up to 383 compounds (at least one position needs to be used as a control) is used to generate a D-Map fingerprint based on the Bliss additivity model of the compound pairs. D-Map fingerprints can then be compared to determine the similarity compounds. Additionally, a compendium of D-Map fingerprints can be built that can be used to predict a mode of action of an unknown compound (figure 7).

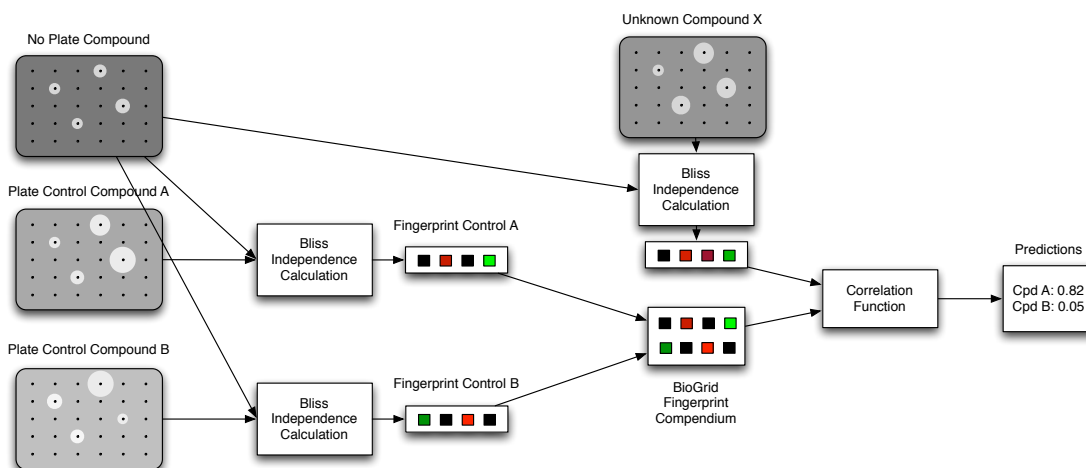


Figure 7 – D-Map method overview. The darkness of each plate represents the amount of yeast lawn growth; dark indicating more growth. Small black dots on the plates indicate pinning location and the size of the halos indicates the level of growth inhibition, the larger halo indicating greater inhibition. Using formula 3, a log Bliss ratio for each pinning location is calculated, the vector of these values form a fingerprint. Combinations of compounds that have less inhibition than expected from the inhibition of each compound by itself show as a red spot in the fingerprints, those with more inhibition show as green. The intensity of the color indicating the degree of variation from the expected inhibition..

The set of pinning compounds is chosen by their modes of action to create a diverse fingerprint. Tuning of the dilution of the pinning compounds is critical, they need to create a small but measurable halo when pinned into a plate containing no compound so that the change in toxicity of the compound pair can be determined. The selection and dilution tuning of the screening library of 109 compounds was previously done (Tamble, 2008) and is shown in Appendix 1, Table 3.

D-Map method

Plates containing wild-type budding yeast (*S. cerevisiae* strain BY4743) were prepared as described in (Gassner, et al., 2007). A total of 67 different compounds were used as plate compounds at the listed concentrations (Appendix 1, Table 5). After the liquid agar and yeast was poured the plates were allowed to cool for ~30 minutes. The 109 screening compounds (Appendix 1, Table 3) were then pin-transferred into the agar as described in (Gassner, et al., 2007). Only 109 screening compounds were used because compounds were placed in a chess board pattern, with the remaining locations left empty to reduce overlapping effects, also no compounds were placed in locations at the edge of the plate due to edge effects (yeast grown at the edges of a plate tends to grow better than in the middle due to a lower competition for nutrients (Kuzmin, et al., 2014)).

Plates were incubated at room temperature for ~24 hours and then absorbance readings were taken using a plate reader at 544nm. For each pinning location 5 readings were taken; one at the pinned location, and one each immediately above, below, to the left, and to the right of the pinned location. An example of a scanned plate can be seen in figure 8.

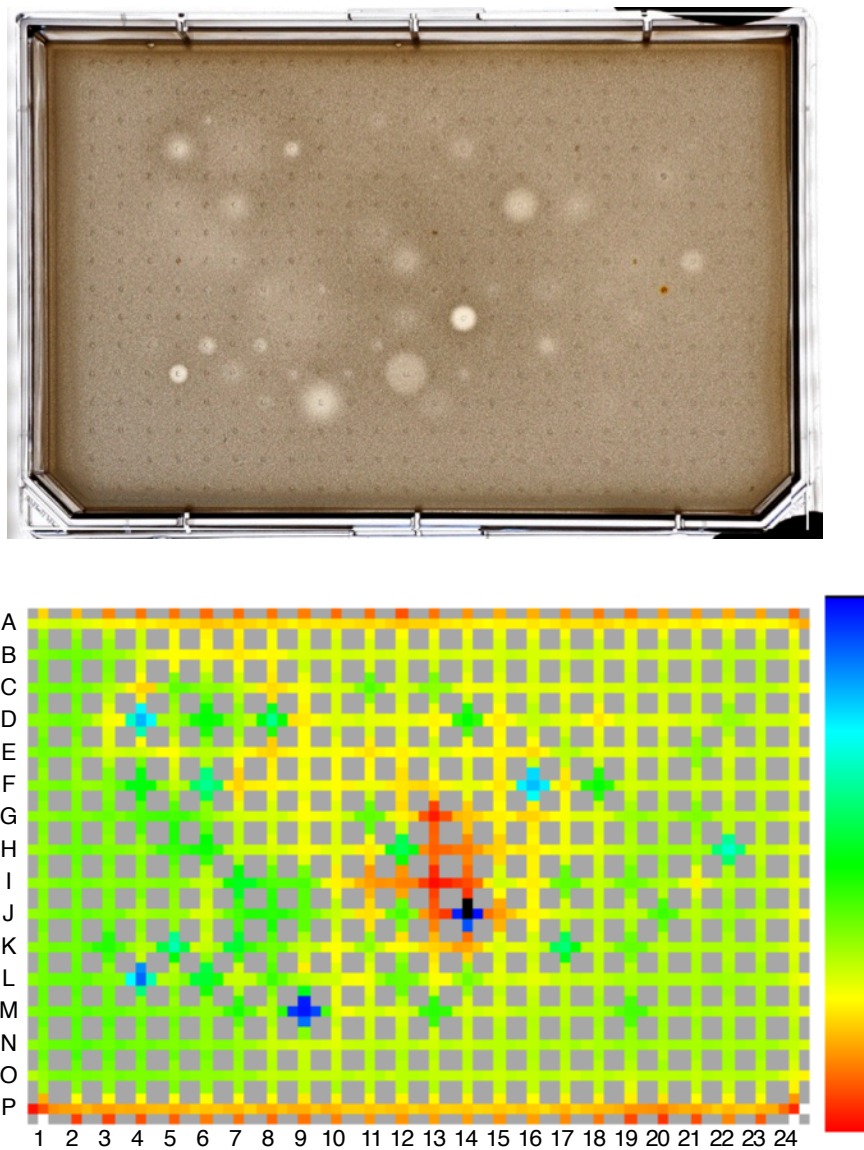


Figure 8 - Example D-Map plate with resulting false-color scanned image. . . Cooler colors indicate more inhibition.

All compounds were run as technical duplicates in daily batches of between 3 and 10 compounds in a total of 13 batches (each batch consisted of 8 to 22 plates, including two, blank control plates per bath).

The log ratio of the expected by Bliss Independence versus actual amount of inhibition for compound c was calculated by:

$$B_c = \log_2 \left(\frac{P_c W_x}{P_c W_0} \cdot \frac{P_0 W_0}{P_0 W_x} \right) \quad (3)$$

where $P_c W_x$ is the growth of plate compound c combined with pin compound x , $P_c W_0$ is growth of plate compound c by itself, $P_0 W_x$ is the growth of pin compound x in a plate with no compound, and $P_0 W_0$ is the growth of yeast without a plate compound nor a pin compound.

To increase accuracy the mean of the 5 measurements is used to determine $P_0 W_x$ and $P_c W_x$. The value of $P_c W_0$ was calculated as the mean of the four closest pinning locations that did not contain a compound (i.e. those immediately above, below, to the left, and to the right of the pinning location). Similarly $P_0 W_y$ and $P_0 W_0$ were calculated using values from the two control plate that did not contain a compound in the agar.

The values of B_c make a vector which I call the D-Map fingerprint for compound c . Note that it is possible to transpose the function of the pinned and plate compounds, using the plate compounds to build a fingerprint for the pinned compounds, however this has the disadvantage that if a new unknown

compound needs to be screened each of the plate compounds will need to be run again.

To determine if D-Map is able to differentiate classes of drugs I clustered them using the Eisen cluster software (Eisen, et al., 1998) to produce a heatmap, figure 9. Note that plates that showed little or no yeast growth across the entire plate were eliminated before clustering; these plates would have clustered with each other and not with plates containing other compounds in their class. Qualitatively the results are promising, in almost all cases the two technical replicates cluster with each other and several compound classes appear adjacent to each other or in tight groups (see figure 9 caption for details).

To quantitatively compare the effectiveness of D-Map's ability to identify different classes of drugs, I pairwise compared the fingerprints of drugs using Pearson correlation (Pearson, 1895). Compounds were not compared to themselves, either from technical replicates or from different dilutions. Similarly compounds which had no documented mode of action were not included.

Figure 10 shows the distribution of Pearson scores for all compounds within the same class and compounds known to be in different classes overall (part A) and for those classes that showed a higher correlation than background (parts B-D). The DNA disruption classes (2A and 2B) did not perform as well as I expected based on their heatmap clustering; this was

likely due to the two classes being grouped into several small clusters rather than one large cluster. The distribution of classes 2A and 2B (not shown) does show a number of compound pairs with a higher than expected correlation, but overall the pairwise correlation is worse than for compounds in different classes.

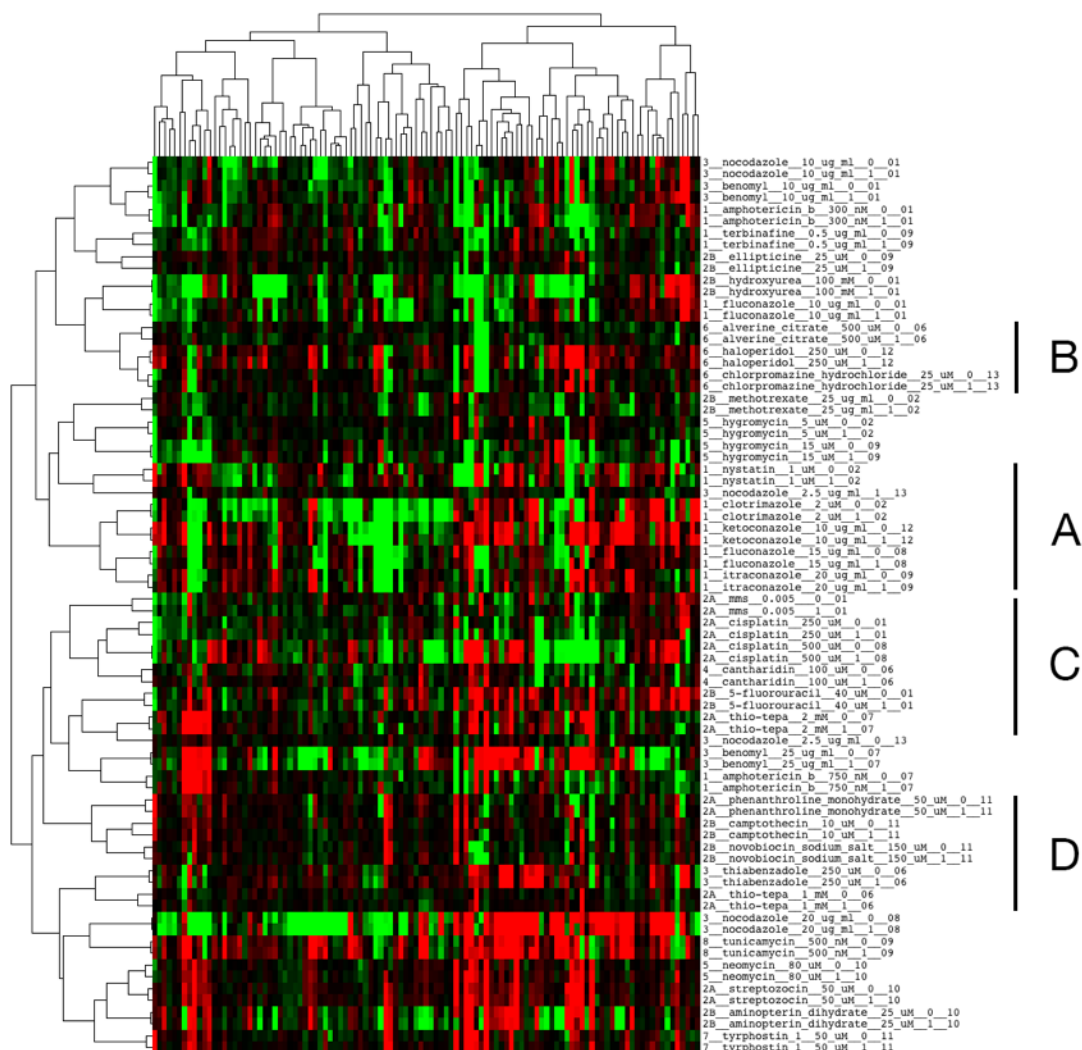


Figure 9 - Clustering of known D-Map test compounds. Compound labels (right column) are composed of class number (Appendix 1, Table 4), compound name, concentration, technical replicate, and batch number. With few exceptions compounds cluster with their technical replicates, as expected. Interesting groups are various Cell Wall Disrupters–Class 1 (A) and the Mammalian Neurological Activity–class 6 (B). The DNA Disruption classes (2A and 2B) are intermixed and split across (C) and (D) and other locations. A heatmap that includes unknown compounds is found in Appendix 1, Figure 22.

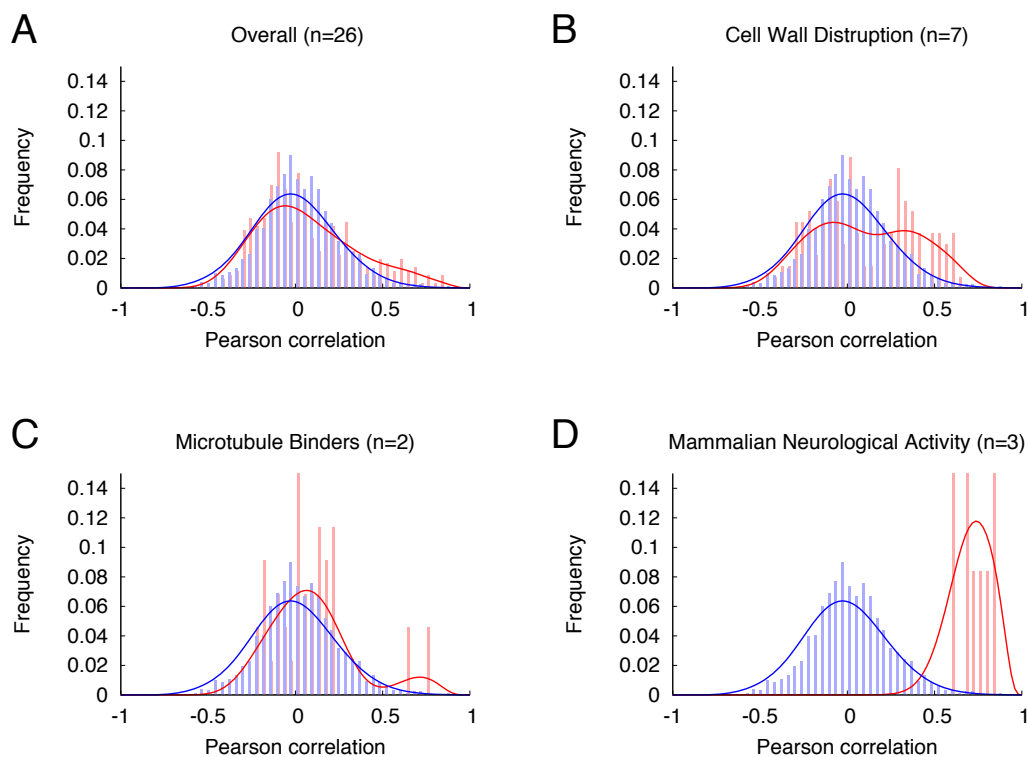


Figure 10 - Distribution of Pearson Correlation Scores overall (A) and for the 3 classes (B-D) that had a statistically significant higher correlation score for compounds within the class (red) versus compounds in differing classes (blue).

Compound counts shown do not included multiple concentrations of the same compound, but data from all concentrations is included in the histograms. Plots B and C show a distinct bi-modal distribution, suggesting that some compounds correlate with others in the same class but others do not.

To determine if the differences are statistically significant I applied the Kolmogorov–Smirnov test (Massey Jr, 1951) to the various classes; table 1 shows those classes for which the calculated P-value is less than 0.05. The high correlation of compounds in the mammalian neurological activity class was surprising, but a literature search showed that yeast has been shown to

react to psychoactive drugs such as Paxil, Haldol, and Prozac (Ericson, et al., 2008) and has been used as a model organism for studying neurodegeneration (Khurana and Lindquist, 2010).

Class Number	Description	Compound Count	KS	P-Value
	Overall	26	0.140	9.27 E-06
1	Cell Wall Distruption	7	0.310	3.36 E-11
3	Microtubule Binders	2	0.246	1.06E-02
6	Mammalian Neurological Activity	3	0.994	1.14E-10

Table 1 - KS and P-Values for significantly correlated compound classes. Note that only 6 classes contained more than one compound, so the 3 classes listed represent 50% of the total number of classes.

I examined the raw plate data for those compound classes that did not reliably cluster together and determined that many of them included halos that were so large that they were not accurately measured by using five scanned points. Similarly other halos were so small that their inhibition levels were inaccurately being determined because averaging the 4 adjacent readings caused the calculated values to be nonlinear in relation to the actual inhibition.

I determined that current assay (Gassner, et al., 2007) has limitations in the dynamic range that can be measured and has difficulties in accurately

recording diffuse inhibition halos. As can be seen in figure 6 there are several pinned locations that have significantly inhibited yeast growth but halos of varying sizes (indicated by A and B). The 5 point scanning method incorrectly determines a similar level of growth inhibition for these locations. Similarly halos that are not well defined, such as (C), are not accurately read. Additionally scanning at just 5 points does not allow us to determine if there are overlapping halos caused by excessively high compound concentrations that generate false positives.

To overcome these limitations I developed the HALO384 (Woehrmann, et al., 2010) method (appendix 2) that uses 9 scanning positions across each pinning location to integrate the optical density over a large area to more accurately determine growth inhibition (figure 11).

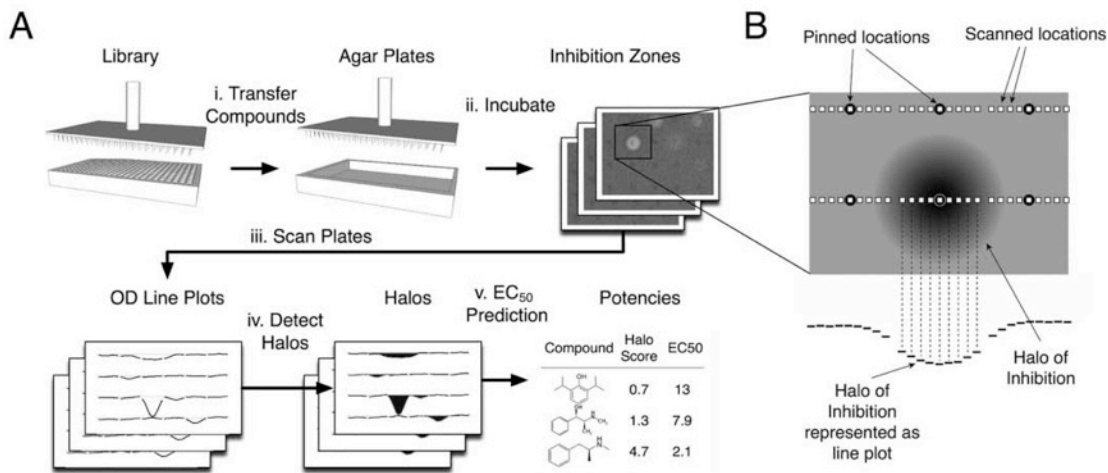


Figure 11 - (A) high-throughput pipeline for drug potency prediction. a library of compounds is transferred from standard 384-well plates into agar using a pinning robot (step i). following incubation at room temperature (step ii), optical density measurements are scanned from the resulting plates, producing optical density (OD) line plots (step iii) from which halos are detected and quantified (step iv). ec50 potencies are predicted for all compounds producing halos and recorded (step v). (B) overview of agar-based pinning, reading, and halo detection strategy. compounds are pinned into soft agar in a grid pattern (small black circles). a plate reader records 9 optical density readings across each pinning location (white squares) that can be viewed in cross section to visualize the pattern of growth inhibition as a function of the distance from pinning (OD line plot; bottom). Toxic compounds show a zone of clearing centered on the location where a compound has been pinned into soft agar (large shaded circle)

The HALO384 method produces potency predictions over a large range of concentrations for a given compound and these potency predictions are highly correlated with EC₅₀ values obtained from liquid culture (figure 12), considered the standard measuring technique for determining drug inhibition levels.

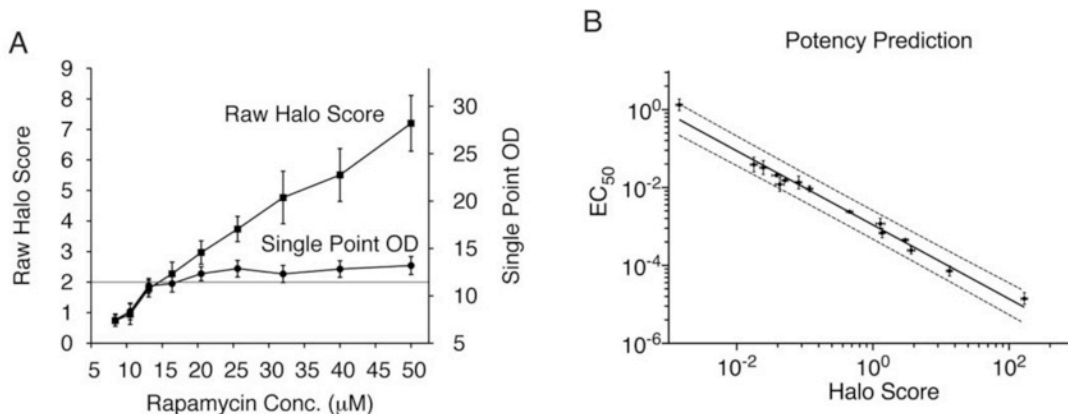


Figure 12 - (A) The dynamic range of the halo score is larger than single-point optical density (SPOD). Raw (pre-normalized) halo scores (left *y*-axis) and a well's SPOD reading (*right y*-axis) plotted against the pinned concentration of rapamycin (*x*-axis). Both the halo score and SPOD increase linearly with rapamycin concentration, but the raw halo score is linear over a wider range than SPOD. Bars represent standard errors calculated from 3 replicates. (B) halo scores predict ec_{50} in *Saccharomyces cerevisiae*. log-log plot of ec_{50} measurements obtained from liquid culture (*y*-axis) plotted against the halo score obtained in agar (*x*-axis) for 19 chemicals of varying toxicity against *S. cerevisiae*. The solid line represents the least squares regression line; dashed lines show the upper and lower bounds of a 95% confidence interval. Linear regression with 95% prediction intervals was performed using GraphPad Prism v5.0b software. The 2 estimates for a compound have an R^2 of 0.98 in log-log space, computed over ec_{50} potencies ranging from 14.4 nM to 1.32 mM.

In addition to producing more accurate inhibition values, the HALO384 method compensates for other inaccuracies, including edge growth effects (the tendency of yeast near the edge of the plate to grow better than yeast in the middle due to better access to oxygen and less competition for nutrients), tilt (the agar is poured in a liquid state and then solidifies as it cools, if the location where the plate is stored during cooling is not level the

agar will be thicker in some areas of the plate than others), and meniscus affect (which caused the agar to be to be thicker near the edge of the plate).

By examining the symmetry of the scanned data the HALO384 method identifies which pinning locations cannot be accurately read due to inhibition caused by nearby compounds and these locations are flagged as unavailable. HALO384 also identifies those pinned locations that have some but not all of the criteria for a halo and flags these as potential halos. The HALO384 method includes a web interface allowing a technician to examine both a line plot showing the optical density as read by the plate reader and the HALO384 analysis so that corrections can be made (figure 13).

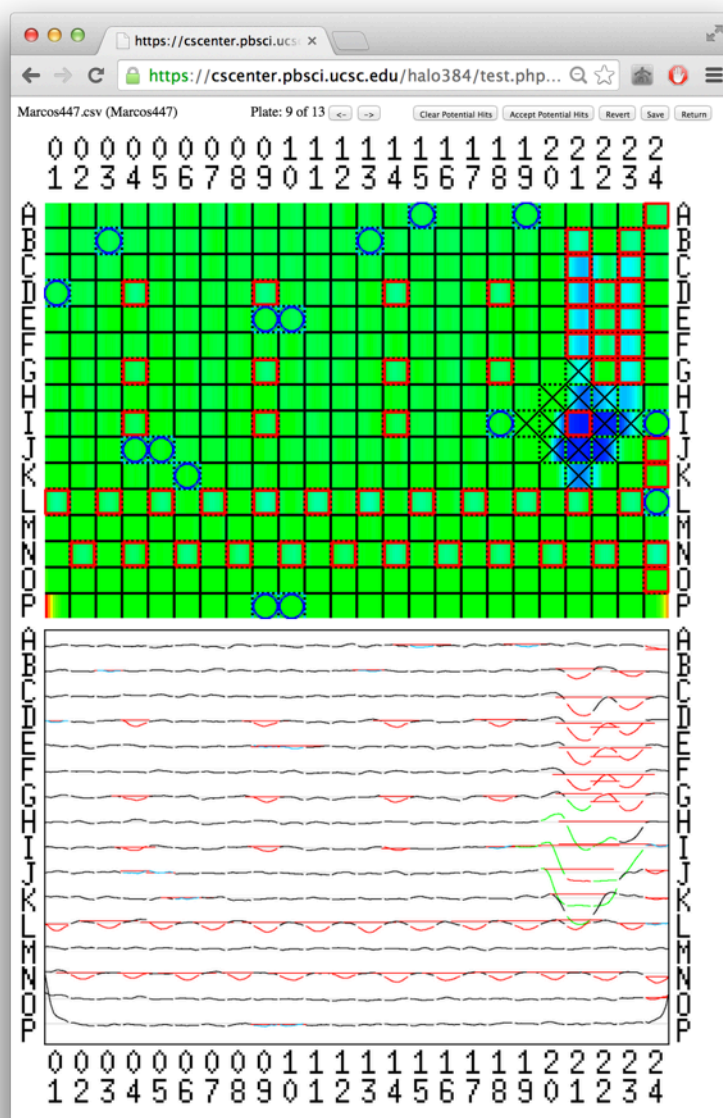


Figure 13 - Screen shot showing the HALO384 review capabilities on a test plate. Red outlines indicate pinned locations positively identified as having a halo, blue circles are a possible halo, and black X's are areas where a halo cannot be determined because a large halo is overlapping those locations (seen at position I21). Of the sites identified as possible halos only I18 was pinned with a test compound. The halos identified at A24, J24, K24, and O24 are false positives, even after correcting for edge effects reliably detecting allows in columns 1 and 24 and rows A and P is difficult.

The quality of a high throughput screen is measured by its Z-factor (Zhang, et al., 1999); a Z-factor equal to or greater than 0.50 is considered an excellent assay. The HALO384 system was evaluated at several different halo sizes and produced a high Z-factor score for even small halos (Figure 14).

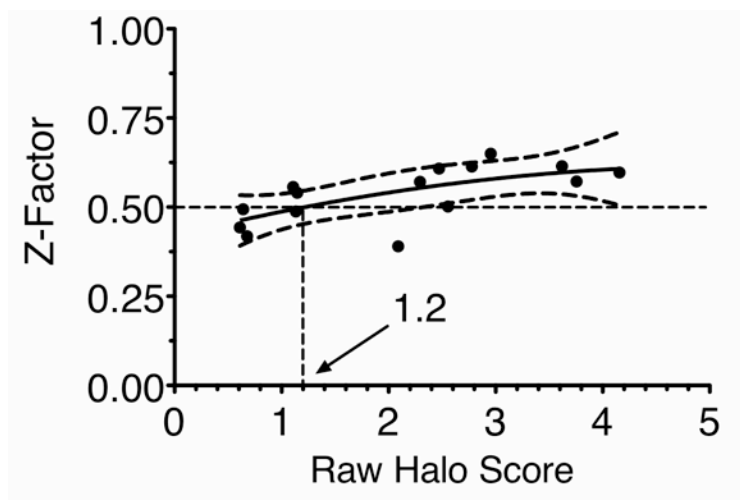


Figure 14 - Validation of HALO384 via Z-Factor analysis.

As part of the HALO384 work I also developed a web site which allows visual comparison of plates with the same compounds but grown under different conditions, with different organisms, or different knockouts, figure 15.

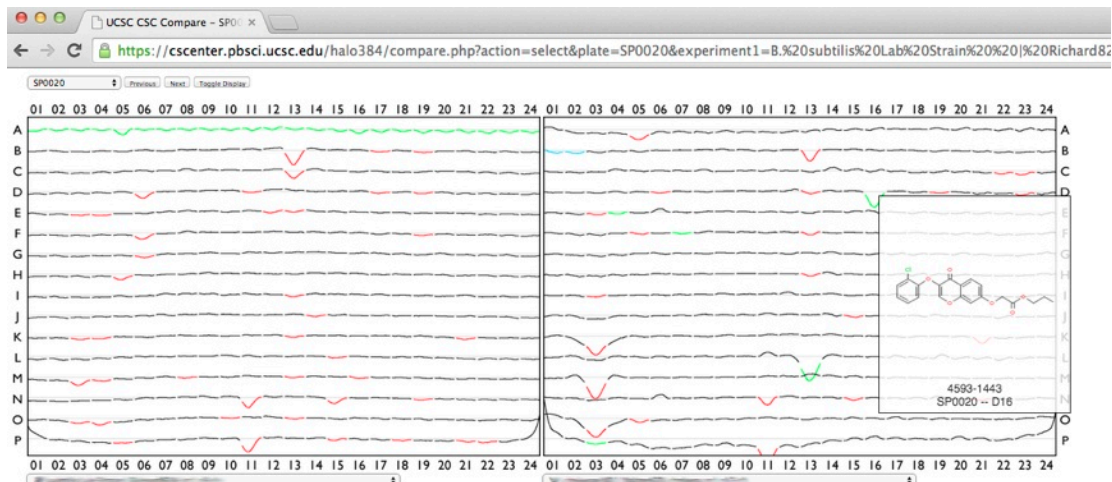


Figure 15 - Screenshot of the HALO384 Plate Compare web site.

In this example two different microorganisms are compared (*Bacillus subtilis* on the left and *Vibrio cholerae* on the right). The compound at location D16 shows a clear difference; mousing over that location brings up a popup showing the compound's structure.

After completing the HALO384 work I attempted to rerun the D-Map experiment to confirm that more accurate growth inhibition readings would improve the ability of the method to identify compound classes. Unfortunately I was not successful; many of the previously seen halos were no longer present, some plates had halos that covered large portion of the plates, and some of the technical replicates produced wildly varying results. I believe these problems were caused by a combination of some compounds precipitating out of solution and others losing potency due to numerous freeze/thaw cycles of the stock compounds.

3.2 BioSpace

Due to limitations in the D-Map method, a second method using the Boone SGA map was developed: BioSpace. Where D-Map fingerprints are generated by using a set of library compounds to find patterns of synthetic lethality, BioSpace fingerprints instead use a set of yeast knockouts. The library set of yeast knockouts is grown exposed to a compound being screened and a Bliss independence log ratio is calculated for each, the vector of ratios giving the BioSpace fingerprint (figure 16).

Because the yeast knockouts are of individual genes and can be chosen this method will allow more control over the synthetic lethal interactions used to build fingerprints. Additionally since the synthetic lethal interactions of the gene knockout set is known, direct prediction of a compound's target is possible (figure 17).

Method Summary

The BioSpace method exposes a set of yeast mutants, each with a different gene knocked out, to a set of compounds and the resulting growth inhibition fingerprints is then compared to fingerprints of known compounds or used to directly predict targets. To determine which genes are to be knocked out the the SGA map (Costanzo, et al., 2010) is used.

Most genes participate in a small number of synthetic lethal relationships, five being the most common number of interactions. Hub genes are those that are synthetically lethal with a large number of genes. Hub genes that have a large number of intra-pathway synthetic lethal relationships and few inter-pathway are known as pathway specific hub (PSH) genes. PSH genes are of interest as BioSpace screening genes since they will allow the largest sensitivity in fingerprint calculation.

The larger the set of PSH genes and the higher the ratio of intra-pathway to inter-pathway synthetic lethal connect, the higher confidence with which the target pathway can be predicted. Since hub genes tend to be synthetically lethal with functionally diverse genes (Costanzo, et al., 2011), finding which PSH genes to knock out requires special consideration.

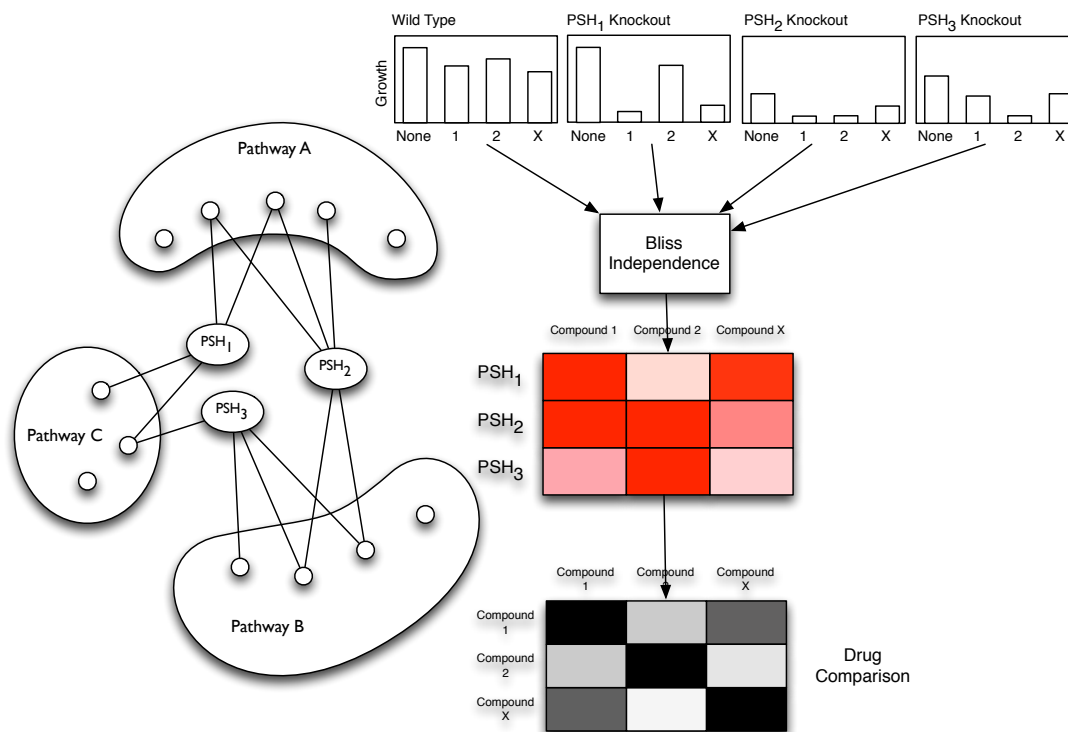


Figure 16 - Overview of BioSpace compound similarity prediction. Pathway specific hub genes (PSH₁₋₃) are chosen for the maximum number of synthetic lethal connections to pathways of interest and minimal connections to other pathways. After incubation the growth of known compounds 1 and 2 and unknown compound X in wild-type (WT) yeast and each of the hub gene knockouts is measured. The fingerprint for compound X is then compared to the fingerprints of the known compounds and a prediction is made. The results of a BioSpace assay can also be used in making direct predictions of a compound's target, figure 17.

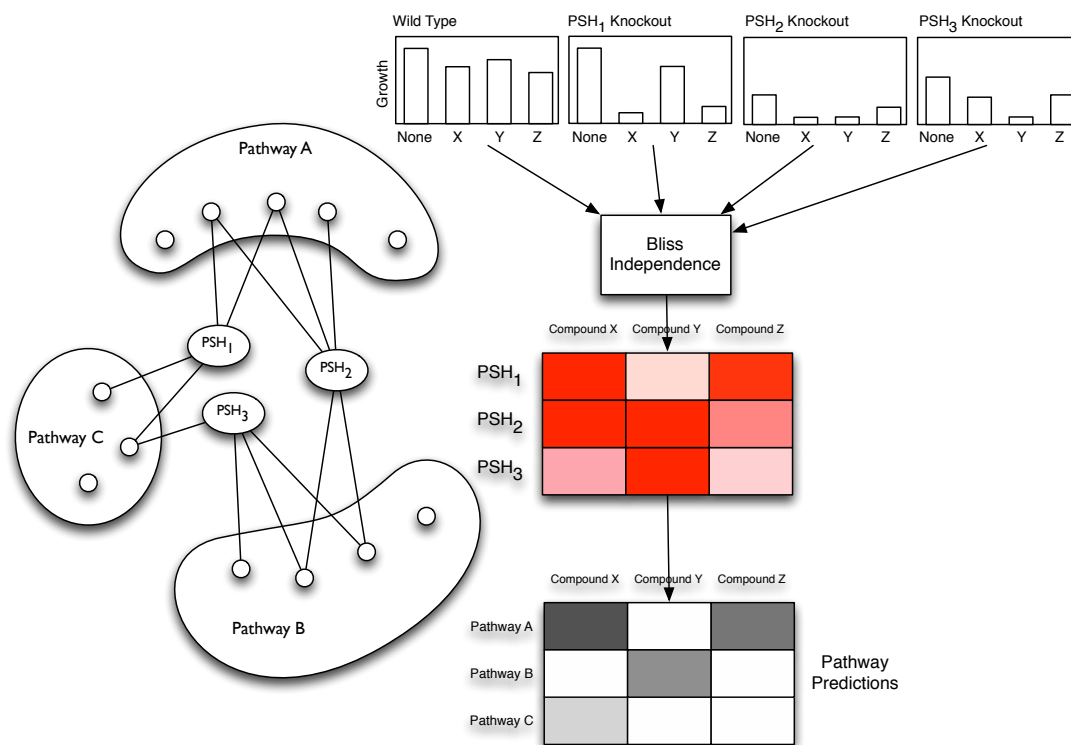


Figure 17 - Overview of BioSpace direct target pathway prediction. The same PSH genes as in figure 16 are used. However, no known compounds need be screened. After incubation the growth of unknown compounds X, Y, and Z in wild-type (WT) and each of the hub gene knockouts is measured. Predictions to the genes targeted by compounds X, Y, and Z are then made.

Knockout Selection Method

To determine what pathways a gene is involved in the Gene Ontology (GO) annotations were used (Ashburner, et al., 2000). To limit the number of knockouts required to a reasonable number it was decided to focus on two GO biological process categories: cell cycle (GO:0007049) and chromosome organization (GO:00051276).

To find the optimal set of pathway specific hub genes I generated a matrix of ratios of genes synthetically lethal with the target gene to the total number of genes in the sub-pathway for each sub-pathways of the two pathways of interest versus the universe of genes analyzed in the SGA map. For example, the GO category “mitotic spindle elongation” (GO:0000022) is a subcategory of the cell cycle category that contains 20 yeast genes found in the SGA map. Of these 5 are synthetically lethal to YMR198W, resulting in an inside sub-pathway versus outside sub-pathway connection ratio of 0.4 for that sub-pathway/gene pair. After calculating a similar ratio for every subcategory of the cell cycle and chromosome organization categories and every gene principle component analysis (PCA) using R is then performed to remove redundancy and reduce the number of dimensions (there are 668 subcategories of the two categories of interest). Based on the Kaiser-Guttman rule (Guttman, 1954; Kaiser, 1960; Kaiser, 1970) of excluding components that contribute less than 1% of addition variance I cut the number of components to 20. This data was then whitened and hierarchically clustered (Eisen, et al., 1998), resulting in the heatmap shown in figure 18.

To reduce the number of knockouts to a reasonable number, which had been previously been determined to be approximately 20, I manually divided the heatmap into groups of similar signatures, indicated by the cyan horizontal lines in figure 18. This resulted in 21 groups, the last being a rather large group of genes that had very little variation.

For each group I found the genes that had at minimum 50 synthetic lethal connections to genes in the subcategories of interest. These were sorted by the ratio of synthetic lethal connections to genes in the subcategories of interest to genes not in those subcategories and the highest scoring gene was chosen as the knockout. In some cases this gene was known to produce a mutant that was very slow growing and therefore not suitable for use in a high-throughput screen, so the second highest scoring gene was chosen instead. See appendix 3, table 6 for the complete list of knockouts chosen.

Method

Wild-type *S. cerevisiae* yeast (strain BY4743) and mutants from the Open BioSystems knockout collection was cultured overnight in YPD. After dilution to an optical density of 0.1, 40 μ l of media per well was added via peristaltic pump (Matrix Wellmate) to a 384 liquid well plate. Compounds were pinned at 1:200 dilution from the stock plate. Plates were then incubated overnight at room temperature with shaking every 90 minutes before being read by an EnVision plate reader.

A library of 615 diverse synthetic compounds (ChemDiv) that were known to be active in yeast was screened against these 21 yeast knockout mutants. A fingerprint for each test compound was generated using the Bliss independence calculation from BioSpace (equation 3). The fingerprints were then pairwise compared using Pearson correlation.

Because the ChemDiv compounds do not have known modes of action I instead compared the BioSpace fingerprints to other similarity measures.

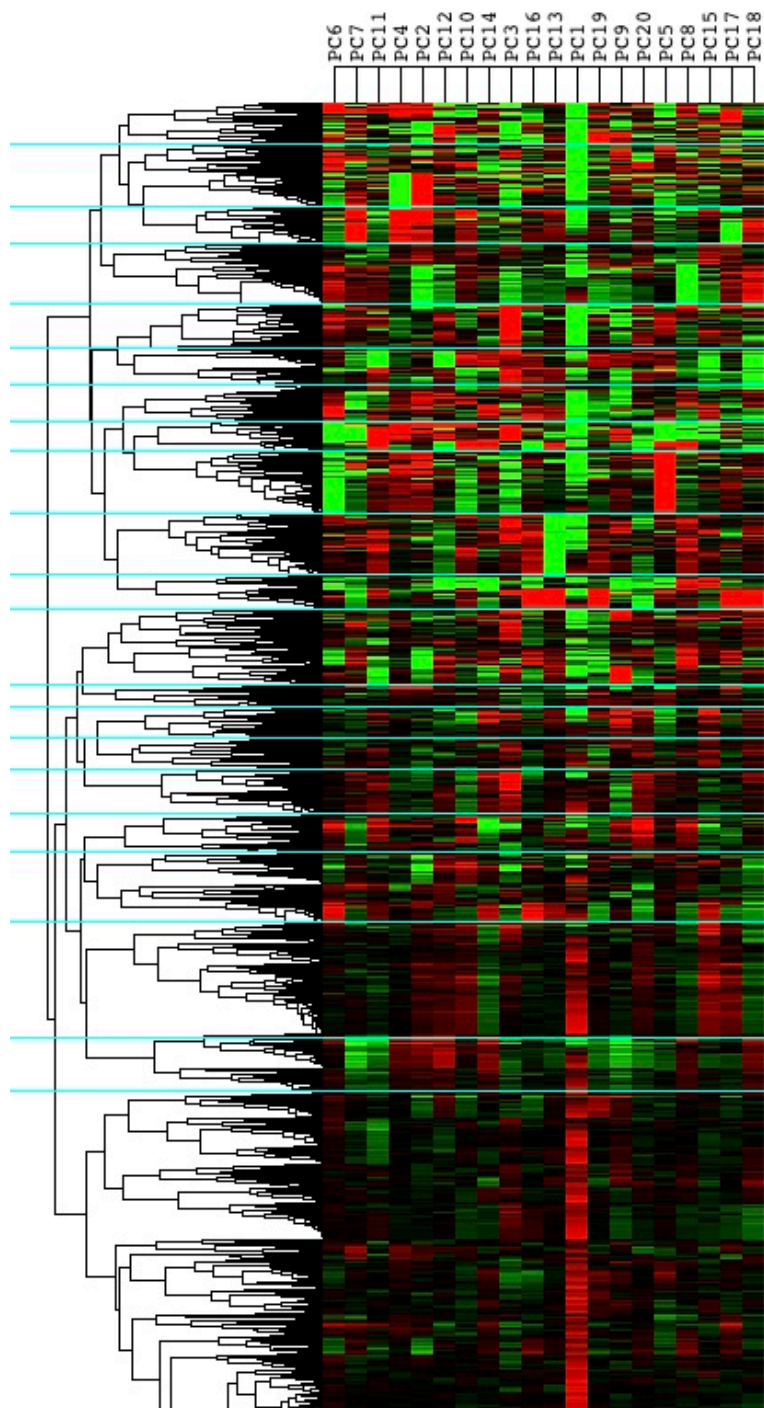


Figure 18 - Partial heatmap showing the clustering of the synthetic lethal connections classes, horizontal lines indicate cuts (see text). The heatmap continues with little variation.

Fingerprint Similarity

Comparing the BioSpace fingerprint pair correlations to MACCS Tanimoto pair coefficients (figure 19) shows little total correlation, however there are more pairs which are score highly in both (figure 19, upper right corner) than in the other extremes, indicating compound pairs that are structurally similar and have nearly the same BioSpace fingerprints.

More interesting regions are the upper left, compounds pairs that have high BioSpace fingerprint similarity but low structural similarity, suggesting that they are either targeting the same protein at two different locations or that even though the compounds have little structural similarity they share an external characteristic that allows them to dock at the same location. And the lower right region, compound pairs that have only a minor change in structure which results in differing protein interactions. External validation of the BioSpace results of these compound pairs in these regions would demonstrate the value of the BioSpace method to predict bioactive similarity.

To allow viewing of the results I built a web site that allows mousing over the compound correlations to bring up a popup allowing visual comparison of the structure and showing the BioSpace fingerprints: <<http://users.soe.ucsc.edu/~marcosw/chemdiv/>>.

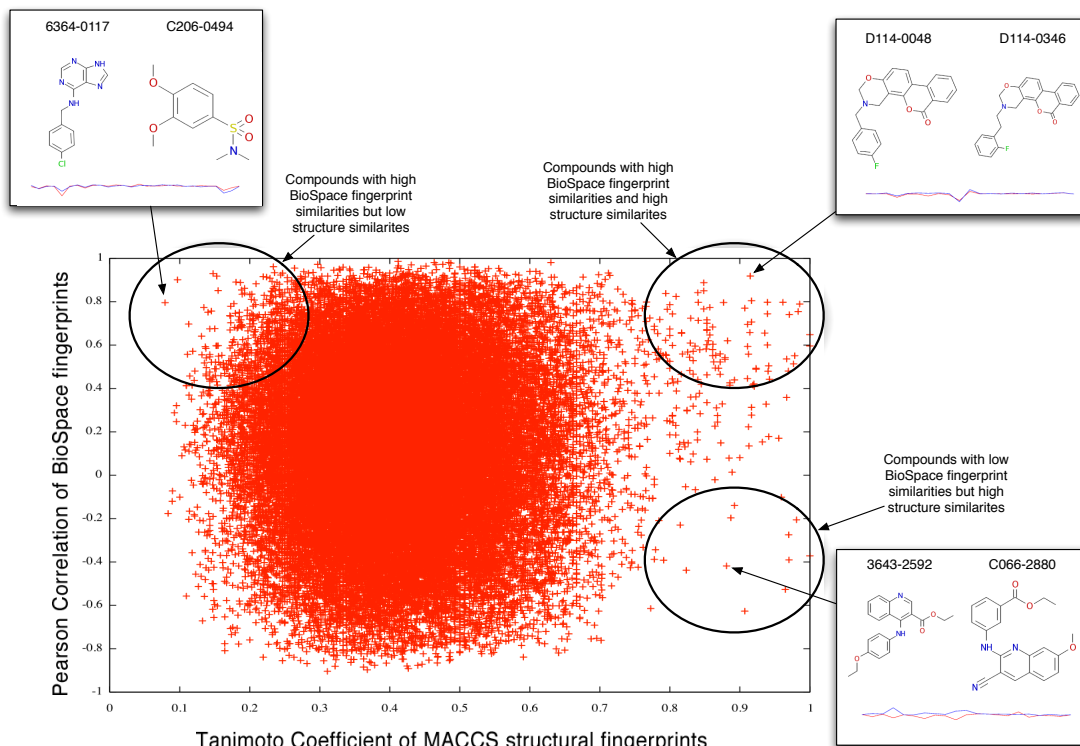


Figure 19 – Correlation of BioSpace fingerprints (vertical scale) against MACCS structure correlation (horizontal scale) for each pair of the 600+ ChemDiv compounds. Mousing over points shows the structure of both compounds (top of each inset figure) and the BioSpace fingerprints (bottom). Most compound pairs show little correlation in the structure space or BioSpace fingerprints, but for highly correlated compounds (>0.70) more have a positive BioSpace fingerprint correlation (upper right hand corner of the main plot). The compound pairs in the upper-left and lower-right regions exhibit differing MACCS and BioSpace correlations. Of particular interest are the compound pairs in the upper left corner that have low structural fingerprint but high BioSpace fingerprint similarity e.g. the ChemDiv compounds 6364-0117 and C206-0494.

To further validate the BioSpace results, I compared the BioSpace fingerprint correlations to cytological profiling fingerprints correlations, previously shown to give good correlation results for many drug classes.

This comparison was done by pairwise comparing the cytological profiling fingerprints using Pearson correlation and hierarchically clustering the results (figure 20 A). The BioSpace fingerprints were similarly pairwise compared and these comparisons were sorted using the same order as the cytological profiling fingerprints (figure 20 B). As a control Tanimoto fingerprints from MACCS keys for the same compounds were sorted in the same order (figure 20 C).

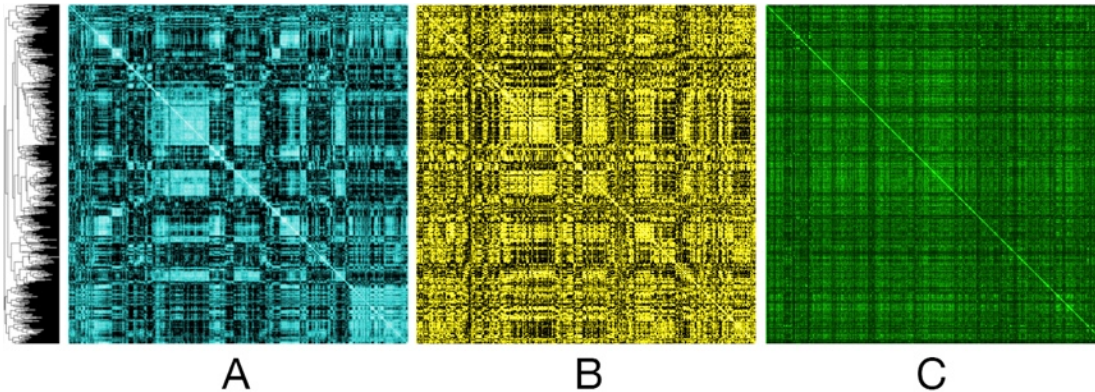


Figure 20 - Heatmaps of Cyto Profiling (A), BioSpace (B), MACCS (C) pairwise fingerprints similarities (Pearson correlation for A and B and Tanimoto Coefficient for C). The similarities are in a 615 x 615 matrix ordered by the clustering of the Cyto Profiling similarities. Some visually similar areas can be observed in A and B.

As can be seen there are several areas that are visually similar between the cytological profiling fingerprint and the BioSpace fingerprint matrices (Figure 20 A and B). To quantify if similarities visible in the correlation matrices are statistically significant I calculated the Spearman correlation coefficient (Spearman, 1904) using the cytological profiling and BioSpace

matrices (figure 20 A and B). The correlation is 0.115, presumably low because similarity is only seen in small clusters. Nevertheless, the P-value is 5.5×10^{-15} . An analogous calculation using the cytological profiling and MACCS matrices (figure 20 A and C) resulted in a correlation of 0.00206 with a P-value of 0.185.

That only pockets of similarity are observed between cytological profiling and BioSpace is possibly due to the limited gene space covered by the BioSpace screening knockouts. Recall that genes were chosen based on their synthetically lethal connections to the cell cycle and chromosome organization categories. BioSpace fingerprints of compounds that target genes outside of these categories would not be expected to be similar.

Somewhat better results are obtained if ECFP₄ fingerprints are used instead of MACCS fingerprints (figure 21). The Spearman correlation coefficient improves to 0.0180.

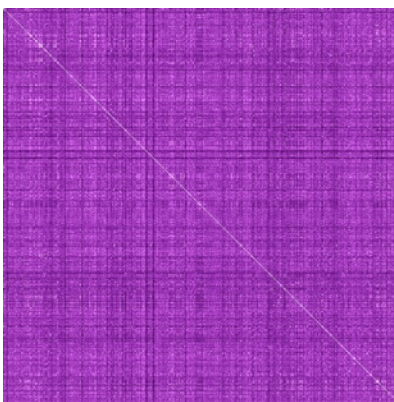


Figure 21 – Heatmap of Tanimoto Coefficient of ECFP₄ fingerprints ordered by Cyto profiling similarity clustering.

Direct Target Prediction

Direct target prediction can be done from the BioSpace fingerprint data using the SGA map via the method summarized in figure 17. Starting with the previously calculated BioSpace fingerprint I applied a cutoff to the values, such that Bliss scores less than a specific value are considered synthetically lethal. I then used the hypergeometric distribution to calculate the overlap of the BioSpace fingerprint to each gene in the SGA map. The larger the number of synthetically lethal combinations in common the higher the likelihood of that gene being the target of the compound under test.

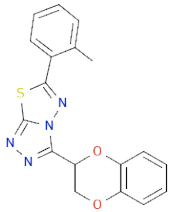
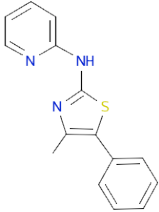
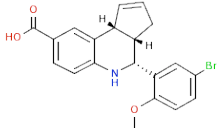
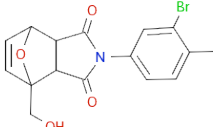
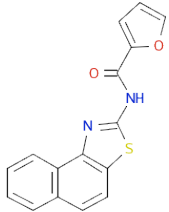
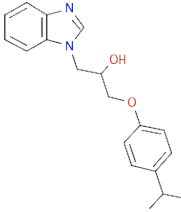
Rather than pick a specific cutoff to determine synthetic lethality I swept through the range of cutoffs from -1.0 to -0.1 by 0.1. The value of -1.0 indicating a large degree of Bliss synergy and therefore a high confidence of a synthetic lethal interaction, and the value of -0.1, a low confidence of synthetic lethality. Using a hypergeometric confidence of 99.9% the list of compounds and target predictions that appeared most frequently is shown in table 2.

The two genes most often included as targets, SWI4 and CTF18, and POL32 are all involved in DNA damage checkpoint or repair, so I predict that compounds D041-0029, 5408-0742, 1935-0139, D054-0047, 5901-0031, and 7756-0709 are either DNA damaging agents or target a protein in the DNA damage repair pathway. The only other target predicted by any of the

compounds is CTF8, is a subunit of Ctf18p that is not known to be involved in pathways related to DNA damage.

Two limitations of BioSpace are the sparse nature of the SGA Map, currently only 30% complete, and the small number of knockout mutants that were used as a screening set. While SGA Map coverage cannot be easily improved, limiting the search to a smaller subset of pathways and/or increasing the number of mutants would increase the accuracy of predictions.

Including positive genetic interactions (synthetic rescue, a double gene silencing resulting in a healthier organism than silencing either individual gene) would also likely result in improvements of the results. Finally, it may be advantageous in treating both the SGA data set and the BioSpace fingerprints as continuous Bliss Independence values rather than applying cutoffs to determine synthetic lethality.

ID	Compound	Best P-value	Predicted Target	SGD Description
D064-0269		9.41E-05	CHS5	involved in export of selected proteins, such as chitin synthase Chs3p, from the Golgi to the plasma membrane;
D041-0029		1.41E-04	CTF18	is required for sister chromatid cohesion; may have overlapping functions with Rad24p in the DNA damage replication checkpoint
5408-0742	Chiral 	2.08E-04		
1935-0139		8.10E-04		
3447-0080		4.94E-04	MUM2	Protein essential for meiotic DNA replication and sporulation
2582-0036		5.65E-04	PMR1	required for Ca ²⁺ and Mn ²⁺ transport into Golgi; involved in Ca ²⁺ dependent protein sorting and processing

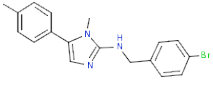
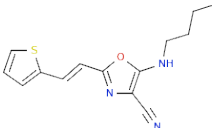
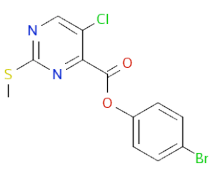
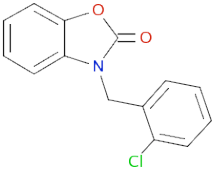
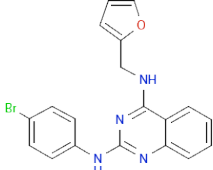
ID	Compound	Best P-value	Predicted Target	SGD Description
7756-0709		3.47E-04	POL32	involved in chromosomal DNA replication; required for error-prone DNA synthesis in the presence of DNA damage and processivity
D151-0850		3.47E-04	SIN3	involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity
D054-0047		4.94E-04	SWI4	DNA binding component of the SBF complex (Swi4p-Swi6p); a transcriptional activator that in concert with MBF (Mbp1-Swi6p) regulates late G1-specific transcription of targets including cyclins and genes required for DNA synthesis and repair
5901-0031		4.94E-04		
4408-0539		4.94E-04	THG1	can also catalyze reverse (3'-5') polymerization with certain substrates in a template-dependent reaction; couples nuclear division and migration to cell budding and cytokinesis

Table 2 - Most frequently predicted BioSpace gene targets from ChemDiv library compounds.

4 Discussion

I have demonstrated three different screening methods for finding the mode of action for unknown compounds, each with advantages and disadvantages. Cytological Profiling uses HeLa or other mammalian cells, therefore its predictions are the most applicable to humans, however, it is difficult, expensive, and time-consuming to do correctly. D-Map, using drug-drug synthetic lethality to build fingerprints, is in principle the easiest to perform, since it uses easy to grow wild-type yeast, however, careful tuning of compound dilutions is necessary and in practice the results were not reliably reproducible. If these problems can be overcome the HALO384 method I developed will make producing accurate fingerprints of D-Map possible. Finally, BioSpace, which uses gene-drug synthetic lethality to build fingerprints, has the advantage that direct prediction of drug targets is

possible. While the target predictions I have made are yet to be validated, I believe this method will be valuable.

Finally, prediction results from Cyto Profiling, D-Map, and BioSpace could also be combined to increase accuracy, similar to the ClueGene method (Ng and Woehrmann, 2007; Ng, et al., 2007) that I co-developed which combined diverse microarray and protein-protein interaction data to predict gene pathway membership.

The problem of drug resistance and new diseases is increasing. We may already have passed the golden age of antibiotics (Davies, 2006) and increases in population density and ease of travel make the emergence of diseases a serious concern. Fortunately the arrival of synthetic chemistry and the searching of the oceans for natural products with therapeutic value combined with high-throughput screening gives hope that new drugs can be found. The methods I have developed can assist with this effort.

Appendices

Appendix 1 - D-Map supplementary data

Well	Compound	Concentration
C03	Sodium Orthovanadate	50mM
C05	benomyl	10mM
C07	phenanthroline monohydrate	50mM
C09	nocodazole	10mM
C11	chlorpromazine hydrochloride	50mM
C13	clotrimazole	5mM
C15	LY 294002	50mM
C17	arabinofuranosylcytosine	40mM
C19	hydroxyurea	130mM
C21	sodium butyrate	90mM
D04	Nystatin	3mM
D06	5-fluorouracil	150mM
D08	rapamycin	8uM
D10	amphotericin B	0.5mM
D12	camptothecin	30mM
D14	calcofluor white	10mM
D16	phleomycin	2.5mM
D18	Z-Leu-Leu-Leu-al	50mM
D20	brefeldin A	50mM
D22	echinomycin	100mM
E03	actinomycin D	
E05	MG132	1mM
E07	motuporamine C	10mM
E09	motuporamine E	10mM
E11	cisplatin	30mM
E13	C239-0032	10mg/ml
E15	griseofulvin	100mM
E17	ketoconazole	10mM
E19	fluconazole	30mM
E21	itraconazole	30mM
F04	terbinafine	30mM
F06	disulfiram	30mM
F08	neomycin	40mM
F10	Anisomycin	1.7mM
F12	Ellipticine	1.3mM
F14	Methyl Methane Sulfonate (M	10%
F16	Tunicamycin	.74mM
F18	thujaplicin	100mM
F20	dyclonine hydrochloride	60mM

Well	Compound	Concentration
F22	novobiocin sodium salt	50mM
G03	ciclopiroxolamine	75mM
G05	fenpropimorph (pestanal)	30mM
G07	2-pyridine-carboxaldehyde	100mM
G09	streptozocin	50mM
G11	aminopterin dihydrate	100mM
G13	rotenone	100mM
G15	haloperidol	100mM
G17	methotrexate	100mM
G19	hygromycin	20mM
G21	NSC-208734 (aclacinomycin)	10mM
H04	NSC-78502	50mM
H06	NSC-65669	50mM
H08	NSC-42559	50mM
H10	NSC-39147	50mM
H12	NSC-185 (cycloheximide)	.5mM
H14	NSC-259968 (bouvardin)	50mM
H16	NSC-8806 (melphalan)	50mM
H18	NSC-7527	50mM
H20	NSC-7212	50mM
H22	NSC-3927	20mM
I03	NSC-3364	50mM
I05	NSC-103645	50mM
I07	NSC-95397	50mM
I09	NSC-82699	50mM
I11	NSC-82150	50mM
I13	NSC-138320	50mM
I15	NSC-123538	50mM
I17	NSC-106997	10mM
I19	NSC-106193	50mM
I21	NSC-32982	50mM
J04	NSC-9219	50mM
J06	NSC-8625	50mM
J08	NSC-7571	50mM
J10	NSC-264713	50mM
J12	NSC-253272	50mM
J14	NSC-221019	20mM
J16	NSC-157035 (NPB04)	10mM
J18	NSC-408120	10mM
J20	NSC-403818	10mM
J22	NSC-349447	10mM
K03	NSC-267694	5mM

Well	Compound	Concentration
K05	NSC-670224	10mM
K07	NSC-180973 (tamoxifen)	10mM
K09	NSC-10777	10mM
K11	NSC-17383	10mM
K13	NSC-638432	10mM
K15	NSC-150117	10mM
K17	NSC-122657	10mM
K19	NSC-207895	10mM
K21	NSC-64875	10mM
L04	NSC-306864	10mM
L06	NSC-371777	10mM
L08	NSC-65238	10mM
L10	NSC-47932	10mM
L12	NSC-301460	5mM
L14	NSC-312033	20mM
L16	NSC-322661	10mM
L18	NSC-48160	10mM
L20	NSC-35446	10mM
L22	NSC-4623	10mM
M03	Alverine Citrate Salt	20mM
M05	Thiabendazole	50mM
M07	Cantharidin	100mM
M09	Thimerosal	25mM
M11	Tyrphostin 1	50mM
M13	Berberine Chloride	25mM
M15	Thio-TEPA	50mM
M17	Psoralen	100mM
M19	Splitomicin	100mM

Table 3 - D-Map screening set compounds

Class Number	Description	Compound Count
1	Cell Wall Distruption	7
2A	DNA Disruption - Damage	5
2B	DNA Disruption - Synthesis	7
3	Microtubule Binders	2
4	Phosphatase inhibitors	1
5	Ribosome Disruption	2
6	Mammalian Neurological Activity	3
7	Kinase Inhibitors	1
8	Protein Transport	1

Table 4 - D-Map classes

Class	Compound	Concentration	Batch
	2-pyridine-carboxaldehyde	100 uM	10
2B	5-fluorouracil	40 uM	1
6	alverine citrate	500 uM	6
2B	aminopterin dihydrate	25 uM	10
1	amphotericin b	300 nM	1
1	amphotericin b	750 nM	7
3	benomyl	10 ug ml	1
3	benomyl	25 ug ml	7
	berberine chloride	125 uM	6
	berberine chloride	250 uM	7
	c239	25 uM	2
2B	camptothecin	10 uM	11
4	cantharidin	100 uM	6
6	chlorpromazine hydrochloride	25 uM	13
10	ciclopiroxolamine	5 uM	11
2A	cisplatin	250 uM	1
2A	cisplatin	500 uM	8
1	clotrimazole	2 uM	2
	disulfiram	30 uM	10
2B	ellipticine	25 uM	9
1	fluconazole	10 ug ml	1
1	fluconazole	15 ug ml	8
	griseofulvin	100 uM	10
6	haloperidol	250 uM	12
2B	hydroxyurea	100 mM	1
5	hygromycin	5 uM	2
5	hygromycin	15 uM	9
1	itraconazole	20 ug ml	9
1	ketoconazole	10 ug ml	12
2B	methotrexate	25 ug ml	2
2A	mms	0.005	1
5	neomycin	80 uM	10
3	nocodazole	10 ug ml	1
3	nocodazole	2.5 ug ml	13
3	nocodazole	20 ug ml	8
2B	novobiocin sodium salt	150 uM	11
	nsc-10777	35 uM	3
	nsc-122657	5 uM	4
	nsc-17383	10 uM	3
	nsc-180973	10 uM	3
	nsc-207895	6 uM	4
	nsc-208734	10 uM	3

Class	Compound	Concentration	Batch
	nsc-221019	5 uM	11
	nsc-301460	2.5 uM	5
	nsc-306864	800 nM	5
	nsc-312033	35 uM	5
	nsc-322661	40 uM	5
	nsc-3364	25 uM	11
	nsc-35446	40 uM	5
	nsc-371777	6 uM	4
	nsc-4623	40 uM	5
	nsc-47932	10 uM	5
	nsc-48160	40 uM	5
	nsc-638432	25 uM	3
	nsc-64875	7 uM	4
	nsc-65238	8 uM	5
	nsc-670224	7 uM	3
1	nystatin	1 uM	2
2A	phenanthroline monohydrate	50 uM	11
	psoralen	25 uM	6
	psoralen	50 uM	7
10	rotenone	500 uM	9
10	rotenone	1 mM	12
10	splitomicin	25 uM	6
10	splitomicin	50 uM	7
2A	streptozocin	50 uM	10
1	terbinafine	.5 ug/ml	9
3	thiabendazole	250 uM	6
	thimerosal	125 nM	6
	thimerosal	150 nM	10
2A	thio-tepa	1 mM	6
2A	thio-tepa	2 mM	7
	thujaplicin	10 uM	11
8	tunicamycin	500 nM	9
7	tyrphostin 1	50 uM	11

Table 5 - D-Map plate compounds

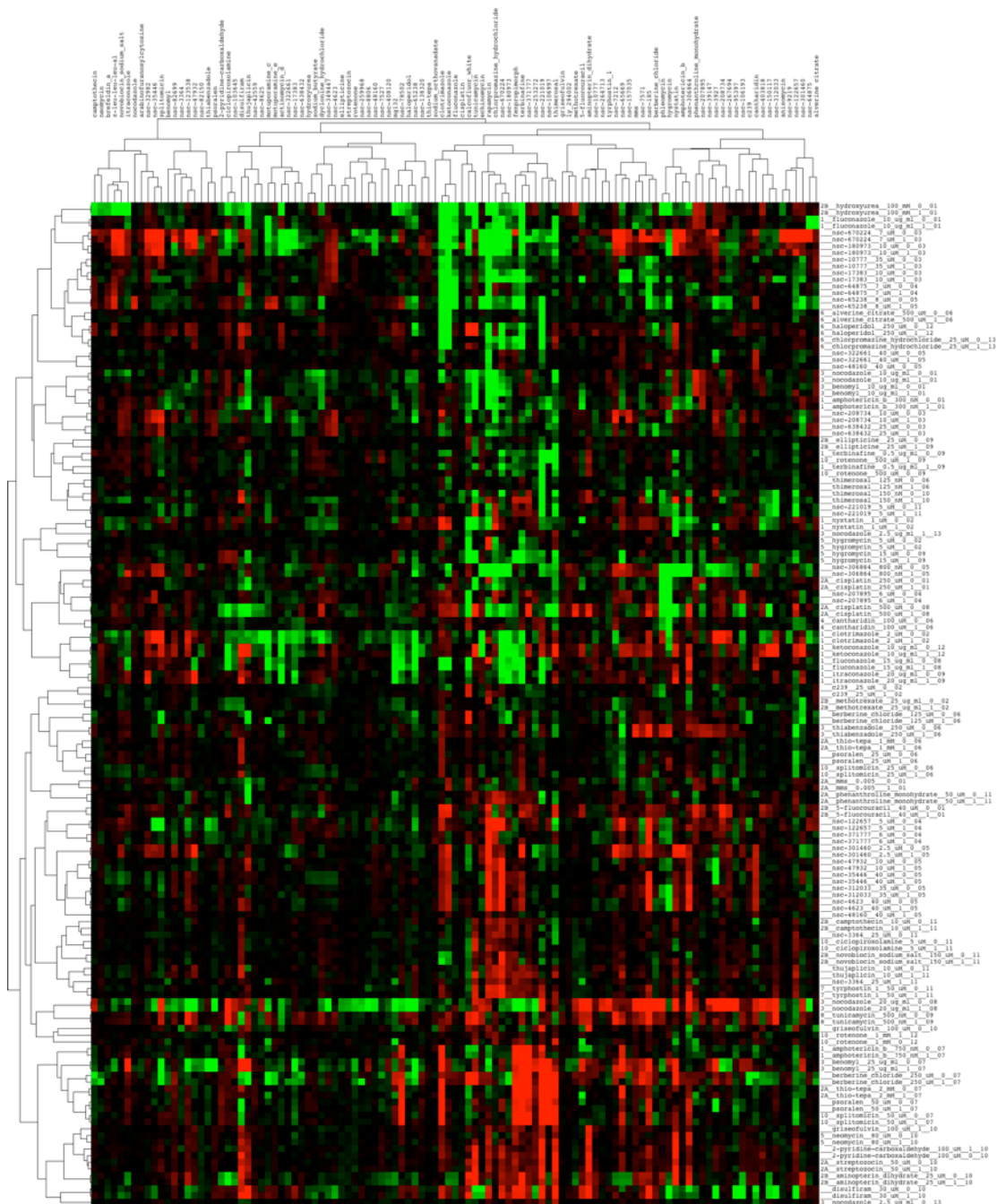


Figure 22 - Heatmap of all D-Map compounds. Test compound names (right column) consist of class (see Table 4, compounds with no known class are shown as ‘_’), compound name, concentration, technical replicate, and batch number. The top dendrogram is labeled with the D-Map screening compounds (Table 3).

Appendix 2 - HALO384

Working with the UCSC Chemical Screening Center I developed a method to accurately determine the growth inhibition of yeast and other microorganisms grown on an agar or similar medium.

With the exception of the Structure-Activity Relationship (SAR) analysis and the wet lab work the contributions were entirely mine.

HALO384: A Halo-Based Potency Prediction Algorithm for High-Throughput Detection of Antimicrobial Agents

MARCOS H. WOEHMANN,¹ NADINE C. GASSNER,^{2,3} WALTER M. BRAY,^{2,3}
JOSHUA M. STUART,¹ and SCOTT LOKEY^{2,3}

A high-throughput (HT) agar-based halo assay is described, which allows for rapid screening of chemical libraries for bioactivity in microorganisms such as yeast and bacteria. A pattern recognition algorithm was developed to identify halo-like shapes in plate reader optical density (OD) measurements. The authors find that the total growth inhibition within a detected halo provides an accurate estimate of a compound's potency measured in terms of its EC₅₀. The new halo recognition method performs significantly better than an earlier method based on single-point OD readings. An assay based on the halo algorithm was used to screen a 21,120-member library of drug-like compounds in *Saccharomyces cerevisiae*, leading to the identification of novel bioactive scaffolds containing derivatives of varying potencies. The authors also show that the HT halo assay can be performed with the pathogenic bacterium *Vibrio cholerae* and that liquid culture EC₅₀ values and halo scores show a good correlation in this organism. These results suggest that the HT halo assay provides a rapid and inexpensive way to screen for bioactivity in multiple microorganisms. (*Journal of Biomolecular Screening* XXXX:xx-xx)

Key words: halo assay, high throughput, optical density, chemical library, agar

INTRODUCTION

THE CLASSIC DISK DIFFUSION, OR HALO, ASSAY is commonly used to evaluate the antimicrobial activity of small molecules and natural product extracts. An agar-plated lawn of microorganism is exposed to a filter disk soaked in a test solution. Lethal or growth-inhibitory compounds cause visible halos, representing a concentration-dependent decrease in growth surrounding the disk. This assay has the advantage of simplicity, and halos provide unmistakable visual confirmation of bioactivity. In addition, because halo size is correlated with potency, the halo assay can be used as a simple and effective way to compare activities among groups of compounds. To accurately determine the inhibition of a compound, we calculate a halo score by measuring the optical density at multiple points across the diameter of the halo and integrating across the area of inhibition. This gives a much more reliable indication of a compound's effect than using either the optical density at the center or the diameter of the zone of death. Indeed, we show that the halo score is accurate enough to estimate a compound's EC₅₀. We recently developed a

high-throughput version of the classic halo assay, in which compounds are delivered robotically to agar plates seeded with a microorganism using a 384-pin tool.¹ Here we describe a computational algorithm to score and quantify potency. We used the algorithm to screen 21,120 compounds in the yeast *Saccharomyces cerevisiae* and identified 590 bioactive compounds from 30 structural classes. The method generalizes across species; we show EC₅₀ can also be predicted in a pathogenic bacterium *Vibrio cholerae*.

MATERIALS AND METHODS

Quantification of drug toxicity from soft-agar pinning using a "halo score"

In the high-throughput (HT) halo assay described here, trays are filled with agar seeded with microorganism, and compound stock solutions are deposited from library plates into the agar using a robotically driven 384-pin array (**Fig. 1A**). Active (i.e., lethal or growth-inhibitory) compounds generate halos, or zones of growth inhibition, in which the effect decays as a function of the distance from deposition.

OD measurements are then taken with a plate reader, scanning 9 points around each well (4 on each side and 1 centered on the site of compound transfer; **Fig. 1B**). When viewed in cross section, a halo gives a characteristic bowl-shape pattern of optical density (OD) that reaches a minimum at the site of compound addition. To aid visual inspection of the results, we produce an "OD line plot" to summarize all of the readings on

¹Department of Biomolecular Engineering and ²Department of Chemistry and Biochemistry, UC Santa Cruz, Santa Cruz, California.

³UCSC Chemical Screening Center, Santa Cruz, California.

Received Aug 17, 2009, and in revised form Oct 2, 2009. Accepted for publication Oct 6, 2009.

Journal of Biomolecular Screening XX(X); XXXX
DOI: 10.1177/1087057109355060

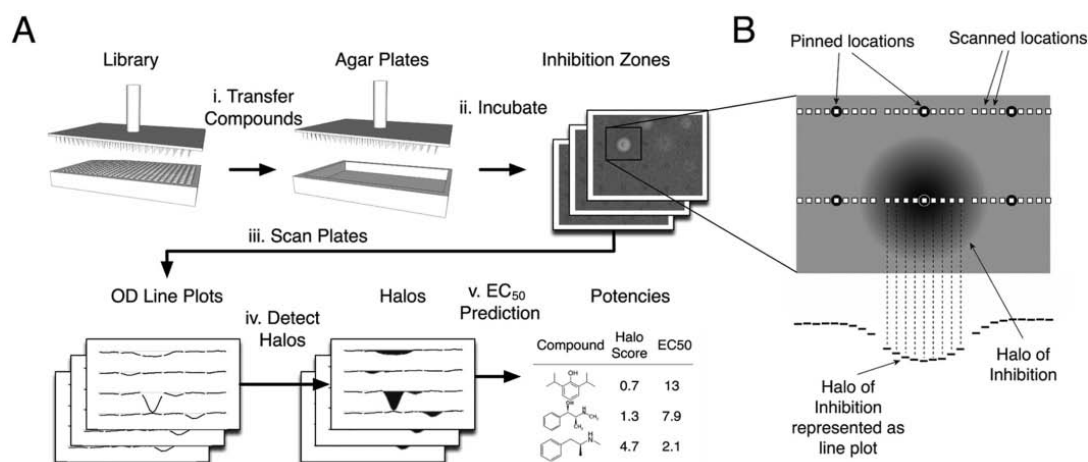


FIG. 1. (A) High-throughput pipeline for drug potency prediction. A library of compounds is transferred from standard 384-well plates into agar using a pinning robot (step i). Following incubation at room temperature (step ii), optical density measurements are scanned from the resulting plates, producing optical density (OD) line plots (step iii) from which halos are detected and quantified (step iv). EC₅₀ potencies are predicted for all compounds producing halos and recorded (step v). (B) Overview of agar-based pinning, reading, and halo detection strategy. Compounds are pinned into soft agar in a grid pattern (small black circles). A plate reader records 9 optical density readings across each pinning location (white squares) that can be viewed in cross section to visualize the pattern of growth inhibition as a function of the distance from pinning (OD line plot; bottom). Toxic compounds show a zone of clearing centered on the location where a compound has been pinned into soft agar (large shaded circle).

a plate in cross section. A raw “halo score” is then calculated for any detected halo-like patterns in the agar by estimating the volume of growth inhibition originating from a single compound (described below). Compound potencies, in the form of EC₅₀s, are then estimated and recorded.

Strains and chemicals

Yeast strain BY4741 was obtained from Open Biosystems (Huntsville, AL). Smooth FY_Vc_1, *V. cholerae* O1 El Tor A1552 was a gift from Fitnat Yildiz.² Growth media reagents were purchased from Sigma (St. Louis, MO). OmniTrays were purchased from Fisher (12565450; Fisher Scientific, Hampton, NH). Library compounds were obtained from the National Cancer Institute’s Developmental Therapeutics Program and ChemDiv, Inc. (San Diego, CA).³

High throughput yeast and cholera halo assay

Media were prepared as previously described.¹ The warm media were inoculated with overnight culture diluted to give a final concentration of $A_{600} = 6 \times 10^{-2}$ and poured into an OmniTray. The tray was set on a flat surface to cool for 15 min and dried in a biological safety cabinet for 15 min.

Compounds were pin-transferred from DMSO stocks plated in 384-well polypropylene trays (Fisher AB1056) into the cooled agar with a pin-tool robot (JANUSMPD; PerkinElmer, Waltham,

MA) using notched pins that deliver 200 nL ($\pm 8\%$) each (VP 384FP3S100; V&P Scientific, San Diego, CA). Before and between applications, pins were cleaned by submersion in 70% ethanol (3 \times), 50% DMSO sonication bath (3 \times), and finally a 95% ethanol circulating 384-channel bath (3 \times). Between each wash step, the pins were applied to blotting paper (V&P Scientific VP540D-100) to absorb excess solvent. At the end of the cleaning cycle, pins were dried in an air drier manifold. The soft agar plate was incubated at 24°C for yeast and 37°C for cholera for 14 h, and then the A600 was read in an EnVision plate reader (PerkinElmer 2104-0010). Each of the 384-pin array points was scanned in a 9-by-1 horizontal line. The data were saved in CSV file format for use as input for the halo detection software.

Data preprocessing

The density of cells in agar, the amount of agar, and other effects can vary across a plate. To mitigate the influence of these local fluctuations of cell density, we normalized the optical density readings by subtracting out 3 main location effects due to the solid agar assay. These effects included the orientation of a reading relative to the site of pinning, the tilt of a plate that may cause systematic differences in cell density across a plate, and whether a reading was taken near a plate edge where cell density can increase because of adherence of the media onto the plastic. These final normalized quantities were then used to detect the presence of bioactive compounds on a plate.

Halo detection and quantification

Using the normalized optical density readings, the presence of bioactive compounds producing characteristic halos of inhibited cell growth was detected and quantified. The halo results from the diffusion of compounds into the agar, which can be used to identify compounds over a wide range of toxicity. Compounds found to produce a zone of inhibition consistent with the shape of a halo were subsequently quantified using a score that reflects the overall amount of inhibition produced.

Intuitively, the algorithm detects hits in agar in an analogous fashion to the way humans identify hits—instead of only inspecting the level of microorganism at the point where the compound was pinned, it searches for circles of reduced growth in a neighborhood around the site. Potent compounds can inhibit cell growth spanning multiple wells in the plate. Therefore, the algorithm first detects multiwell halos, flags any wells that are included in any of these large halos, and then searches for single-well halos within the remaining wells. Because it is unlikely that a circular pattern would be produced by chance and because integrating multiple readings can mitigate the noise present in any single reading, this shape-based approach has the potential to be much more accurate than a single-reading-based approach.

Once a halo is detected, the amount of total inhibition is quantified for wells detected to be centered on the halo. A *raw halo score* is computed for all multiple- and single-well halos. The raw halo score sums up all of the readings to the left, right, and center of a detected halo. The final *halo score* is computed by dividing the raw halo score by the concentration of the compound that was pinned onto the plate.

Growth inhibition measurement in liquid culture

Yeast (wild-type haploid strain BY4741, $A_{600} = 6 \times 10^{-2}$) or cholera (smooth FY_Vc_1, *V. cholerae* O1 El Tor A1552, $A_{600} = 1 \times 10^{-4}$) were incubated with 6 two-fold dilutions of each compound in 200- μ L cultures in 96-well plates, in addition to DMSO controls. ODs were read every 45 min using an EnVision plate reader, and the plate was agitated just prior to reading to suspend the cells. Yeast or bacteria doubling times at each concentration were calculated and compared to the doubling time in DMSO.

Determination of EC_{50} from liquid culture

To estimate EC_{50} from liquid culture, we fit a dose-response curve to the liquid culture optical density readings for a single compound using the GraphPad software (GraphPad, San Diego, CA). We then used the concentration (independent coordinate) at which the fitted curve passed through the midpoint of the optical density readings (dependent coordinate) as the most likely EC_{50} . For compounds with steep Hill slopes, as was the case for many of the NSC compounds, GraphPad either produced no confidence interval or output an excessively wide

range. Therefore, rather than use a confidence interval estimated by GraphPad, we determined a range of possibility (ROP) for the EC_{50} , consisting of the minimum and maximum concentrations that must bracket the most likely EC_{50} value based on the data. First, we identified the 2 points that straddled the 50% inhibition point lying nearest to the fitted sigmoid. Then, the concentrations of these 2 points were used as the minimum and maximum of the ROP. For each compound, we report the ROP along with the most likely EC_{50} estimate.

Molecular similarity analysis

The molecular similarity component of Pipeline Pilot (Accelrys Software, San Diego, CA) was used to calculate the similarity between molecules in the ChemDiv collection using SciTegic's molecular fingerprints (FCFP = 4). Similarity was calculated using Tanimoto coefficients to generate up to 2000 compound clusters with an average of 50 members.

RESULTS

The halo algorithm detects a broad range of compound toxicity

To determine the halo algorithm's utility for predicting a compound's potency, we evaluated the correlation between the raw halo score and the stock solution concentration for a series of known drugs that span a wide range of potencies: rapamycin, disulfiram, and ciclopiroxolamine (EC_{50} s: 14 nM, 94 μ M, and 39 μ M, respectively; Fig. 2). We used a constrained linear regression in which fitted lines were forced to pass through the origin so that compound concentrations of zero were matched with halo scores equal to zero. Raw halo scores and compound concentrations were strongly correlated for rapamycin ($R^2 = 0.93$, $p < 1.2 \times 10^{-4}$) and disulfiram ($R^2 = 0.79$, $p < 0.0175$) and not significantly correlated for ciclopiroxolamine ($R^2 = 0.10$, $p < 0.45$) where the raw halo scores were very small and influenced by excess noise. These results are conservative as the R^2 values are underestimated during the constrained linear regression.

The HT halo assay reported previously¹ used a single-point OD (SPOD) reading per compound, rather than the set of 9 readings used to calculate the raw halo score used here. To quantify whether the new halo score approach improves the detection range compared to the previous method, we plotted the SPOD readings against an increasing concentration of rapamycin and compared it to the results obtained for the raw halo score (Fig. 3A). For compounds that are less potent, the raw halo score and SPOD readings are both able to discriminate between halos of different diameter. However, above a critical concentration of pinned stock solution (for rapamycin around 15 μ M), the SPOD readings flatten out while the raw halo scores continue to increase linearly. The current method takes advantage of the spatial pattern created by compound deposition, expanding the upper limit of potencies predicted for toxic compounds.

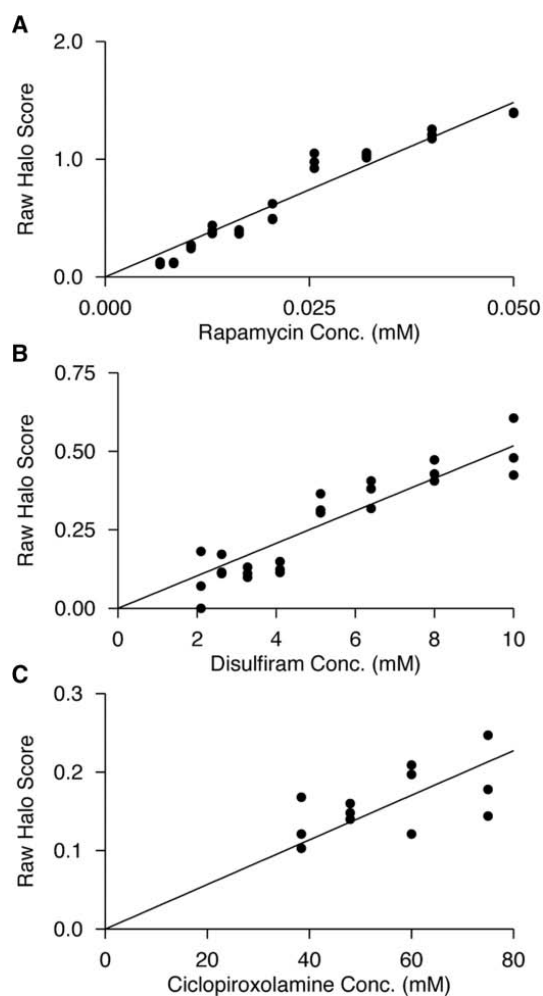


FIG. 2. Linear correlation of halo score with compound concentration. (A-C) Scatter plots of stock concentration (x-axis) against raw halo score (y-axis) for different compounds. Points represent compounds pinned into a different well on a plate. Plates were run in triplicate for each of 3 compounds, including (A) rapamycin, (B) disulfiram, and (C) ciclopiroxolamine.

Because the raw halo score increases linearly with the pinned stock concentration of a compound, we calculated a final halo score (H) by dividing the raw halo score by the compound's stock concentration. The final halo score allows for direct comparison of measurements from different compounds or from the same compound run on different plates and, as we show next, can be used to predict a compound's EC_{50} .

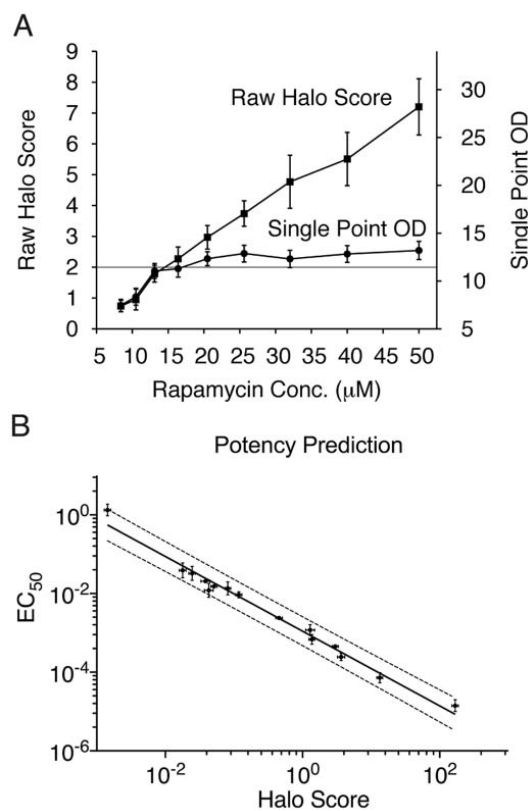


FIG. 3. (A) The dynamic range of the halo score is larger than single-point optical density (SPOD). Raw (prenormalized) halo scores (left y-axis) and a well's SPOD reading (right y-axis) plotted against the pinned concentration of rapamycin (x-axis). Both the halo score and SPOD increase linearly with rapamycin concentration, but the raw halo score is linear over a wider range than SPOD. Bars represent standard errors calculated from 3 replicates. (B) Halo scores predict EC_{50} in *Saccharomyces cerevisiae*. Log-log plot of EC_{50} measurements obtained from liquid culture (y-axis) plotted against the halo score obtained in agar (x-axis) for 19 chemicals of varying toxicity against *S. cerevisiae*. The solid line represents the least squares regression line; dashed lines show the upper and lower bounds of a 95% confidence interval. Linear regression with 95% prediction intervals was performed using GraphPad Prism v5.0b software. The 2 estimates for a compound have an R^2 of 0.98 in log-log space, computed over EC_{50} potencies ranging from 14.4 nM to 1.32 mM.

An intuitive and widely used measure of drug toxicity is the EC_{50} or the *effective concentration* that causes a reduction of 50% in cell population. In microorganisms, EC_{50} s are generally determined in liquid culture using a low-throughput measurement of

Table 1. Potency Prediction of Uncharacterized Compounds^a

Compound	Liquid Culture		Agar	
	Dose Response Curve	EC ₅₀ (range) μ M	Halo Profile (Raw Halo Score)	Pred. EC ₅₀ (range) μ M
NSC 371777		29.6 (28.6-50.0)		5.27 (2.27-12.3)
NSC 17383		24.6 (20.0-47.4)		33.40 (14.2-78.9)
NSC 638432		11.7 (11.2-20.0)		12.96 (5.55-30.3)
NSC 65238		9.79 (8.00-18.5)		7.93 (3.41-18.4)
NSC 207895		2.88 (2.12-3.91)		3.31 (1.42-7.67)
NSC 301460		0.66 (0.57-1.30)		2.24 (0.96-5.20)

^aSix compounds from the National Cancer Institute (NCI) mechanistic, diversity, and natural product (MDNP) libraries were compared with the EC₅₀ values predicted by the halo score method. The columns show the compound identity (column 1), the EC₅₀ determined by liquid culture (column 2; gray demarks the 95% confidence interval estimated by GraphPad), the EC₅₀ value (column 3), the optical density (OD) line plot (column 4), the predicted EC₅₀ value, and associated range of possibility (column 5; see Materials and Methods).

growth rate as a function of compound concentration across a dilution series. We investigated whether the halo score, generated from a single concentration of pinned stock solution, could be used directly to predict EC₅₀ values. If successful, this would enable drug potency characterization in a high-throughput setting. To do this, we determined EC₅₀s in liquid culture for 19 chemicals of varying toxicity (from 1000 μ M to 0.1 μ M) in *S. cerevisiae*. We asked whether H could predict EC₅₀ determined from liquid culture. We first plotted the EC₅₀s against the halo scores in log-log space (Fig. 3B). Regression analysis revealed that the logarithm of the EC₅₀ was linearly correlated with log H ($R^2 = 0.98$; $p < 1.4$

$\times 10^{-4}$). These results suggest that an EC₅₀ estimate (E) can be calculated from a compound's halo score using the equation $E = \alpha H^\beta$, where α and β were estimated from the intercept and slope of the linear regression. In particular, for *S. cerevisiae*, $\alpha = 10^{-3.0}$ and $\beta = -1.0$ so that $E = 10^{-3}H^{-1}$. Using a sliding window across the log H values, the standard deviation of the log EC₅₀ values was calculated from which 95% prediction intervals were derived.

We next evaluated the halo score's ability to predict the potencies of unknown compounds. We selected 6 test compounds from the 3081-member National Cancer Institute (NCI) mechanistic, diversity, and natural product (MDNP) libraries

that were known to have a range of potencies in *S. cerevisiae*. EC_{50} s were predicted for each compound using the EC_{50} s estimated from the regression in log-log space. The predicted EC_{50} s ranged from 0.61 μ M for NSC-301460 up to 16.2 μ M for NSC-371777. We determined EC_{50} s using standard methods in liquid culture and compared these values to the EC_{50} -based predictions (Table 1). The EC_{50} s determined in liquid culture showed good agreement with those predicted based on *H*. Five of the 6 compounds had EC_{50} s within the ROP calculated from the data (see Materials and Methods). One compound fell outside the range, NSC-371777. Repeated attempts to assess the EC_{50} in liquid confirmed the difference in the liquid- versus the agar-based EC_{50} estimation (data not shown). In this case, the disagreement could be due to the compound's differential effects on cells in liquid versus agar. Apart from NSC-371777, 4 of the remaining estimated EC_{50} s differed by no more than 50% from the EC_{50} s determined in liquid, and all 5 had ranges of possibilities that overlapped with the 95% confidence interval for EC_{50} s from the halo score. Thus, the agar-based halo score produces highly comparable EC_{50} s to the more laborious liquid-based approach in 80% of the cases and is applicable to a broad range of hit potencies.

Accuracy of halo scores for high-throughput screening

For the purpose of screening large chemical libraries, a trade-off exists between recall, the sensitivity to detect even moderately toxic compounds, and precision, the proportion of true positives among the detected hits. To measure the utility of the halo score in a screening setting, we compared its ability to detect hits to the previously deployed SPOD method.

We conducted 2 tests to measure the accuracy of the halo and SPOD methods. First, a "bioactives" test was performed in which true positives were defined as those wells pinned with one of the potent compounds listed in Figure 2. Various concentrations of these compounds were used as true positives. To measure the sensitivity of the methods at low yet biologically relevant concentrations, we set the minimum pinned concentration to 20-fold higher than the EC_{50} determined in liquid culture. Second, a "screening" test was performed on the NCI MDNP library. In this test, a human expert using visual inspection aided by OD line plots defined the true positives halos. The true positives in this case were defined based on symmetry and alignment to the site of pinning.

For both tests, we plotted the precision as a function of recall by sweeping through a set of cutoff values for both the halo score and SPOD methods (Fig. 4A). In both the bioactives and screening tests, the halo score method had a higher precision than the SPOD method for recall levels in the range of practical application. For the purpose of library screening, high recall rates are desirable, even if a few false positives are allowed because these can be discarded in follow-up screens. Therefore, the precision at a recall of 90% and higher is of particular interest. At the 90% recall rate for both tests, the

precision of the halo score method was significantly higher than SPOD. In the bioactives test, SPOD had a 40-fold higher false-positive rate than the halo method at the 100% recall rate. Thus, the halo score is predicted to drastically reduce the number of potential secondary screens should the primary screen be conducted at a desired maximum sensitivity. Upon inspection, the false positives called by the SPOD method at the 100% recall rate, which were correctly excluded by the halo score, were caused by edge effects or overlap with neighboring hits. Interestingly, in the screening test, the halo score method did not achieve 100% recall. This is due to the fact that many of the hits called by the human expert did not meet the shape criteria imposed by the halo score method. We checked all 204 of these expert-defined hits and found that 4 of them had no pinned compound. Thus, a recall of approximately 98% is optimal for the bioactive test, and the best recall achieved by the halo score is 94% after excluding the erroneous positive calls made by the human expert. The results of these 2 tests confirm the utility of the halo score as an automated method that is sensitive across a range of known activities and exceeds the accuracy of expert visual scoring.

To further assess the quality of the halo score-based high-throughput assay, we used Z factor analysis.⁴ The Z factor measures the degree of separation in the reported scores between the positive and negative calls relative to standard deviations of the scores. Higher Z factors indicate that fewer false positives and false negatives can be expected. A general rule of thumb for commercial applications is that a screen has a minimum Z factor of 0.50, which corresponds to a separation between signal and background of 3 standard deviations.

To measure the halo score's applicability for use in an HT setting, we estimated Z factors for 2 drugs, rapamycin and disulfiram, pinned at stock concentrations. To obtain a conservative estimate of the assay's Z factor, we pinned stock concentrations reflecting potencies at the low end of detection (5-20 μ M rapamycin, 2.5-5 mM disulfiram) at 11 sites on the plate. We repeated this analysis 3 times on 3 different plates and plotted the Z factors (Fig. 4B). The separation and corresponding Z factors were reproducible across plates. As expected, rapamycin at 20 μ M, corresponding to the most potent positive control, obtained the highest Z factor of 0.68, well within the accepted range of a high-quality HT assay. On the other hand, the weakest controls had more borderline Z factors, and 2 were outside the suggested range of HT. Nonlinear regression revealed that a raw halo score of 1.2 corresponded to a Z factor 0.5 (Fig. 4C), corresponding to a raw halo score cutoff predicted to provide a robust measure of potency when performing a primary screen. However, because the Z factor analysis is overly conservative, we have found that screening with raw halo scores around 0.30 provides sufficiently accurate calls. For standard commercial libraries with stock concentrations of 10 mM, this halo score allows detection of compounds with EC_{50} values of 30 μ M or lower, and we expect the assay to be applicable to compounds with much higher EC_{50} values (e.g., in the 200- μ M range).

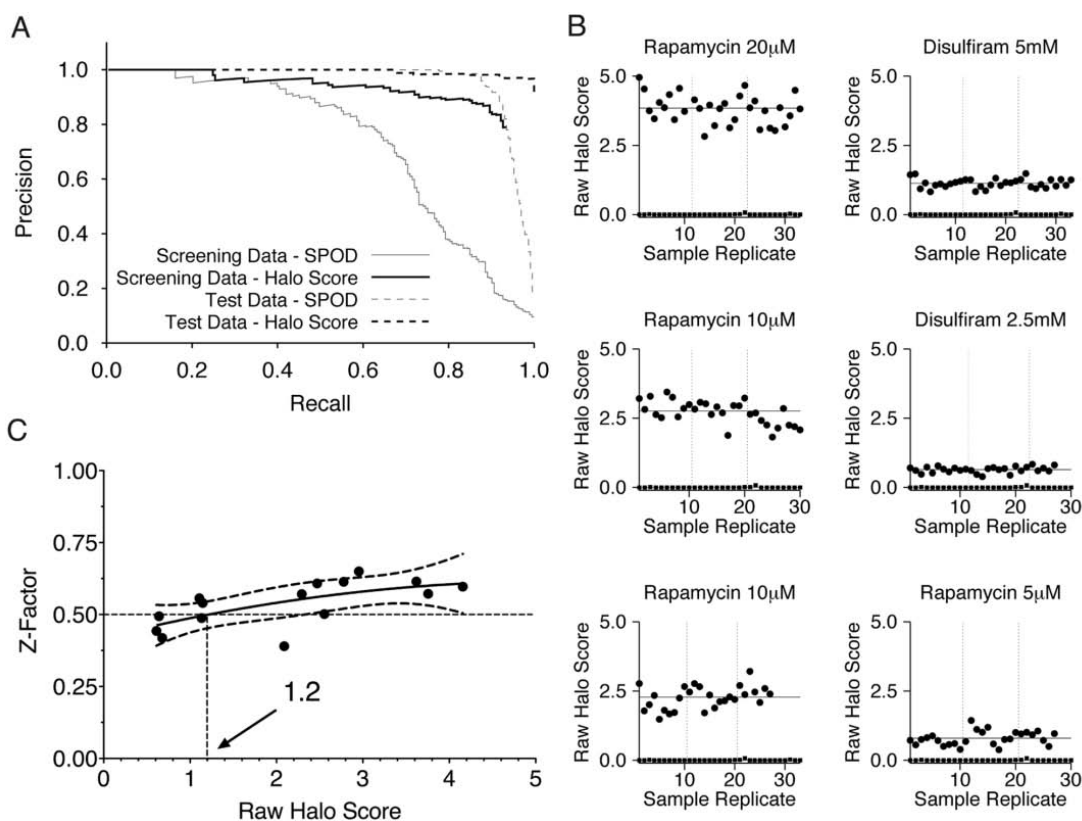


FIG. 4. (A) Precision recall plot comparing the halo score and single-point optical density (SPOD) methods. The performance of the halo score (black lines) and SPOD methods (gray lines) was measured in terms of precision (y -axis) and recall (x -axis) in a small screen in which halo score calls were compared to a human expert (Screening Data; solid lines) and in a second setting in which calls were made for a set of compounds with known activity (Test Data; dashed lines). (B) Separation of positive controls from background. Raw halo scores of positive controls for 30 or 33 replicates (circles) were plotted alongside raw halo scores estimated for background wells containing no pinned compound (triangles). Two separate rounds were performed for rapamycin at 10 μ M. (C) Assessment of halo score performance as a screening method. The Z factor for potency prediction (y -axis; see Materials and Methods) is plotted for several compounds with different halo scores.

Application of the pipeline to *S. cerevisiae* and *V. cholerae*

We used the algorithm to screen 21,120 compounds from a commercial library of drug-like compounds (ChemDiv, San Diego, CA).³ A screen of this library against *S. cerevisiae* resulted in the identification of 590 bioactive compounds comprising 30 distinct structural scaffolds, out of 1056 scaffolds in the library. Activity and structural trends for a cluster of 2,4-diaminoquinazolines were determined (Fig. 5). EC_{50} values were predicted from the halo scores, and 95% confidence intervals were determined (Table 2). Repeated pinning of the same compound showed a standard deviation of 10% in EC_{50} prediction from day to day.

DISCUSSION

We have developed an automated method for identifying antimicrobial agents that is rapid, sensitive, and accurate. The key component of this algorithm is a “halo score” that uses multiple OD readings at different intervals from the site of pinning. In this way, the method makes use of the symmetric decrease in OD as a function of distance from its point of deposition. Correlations with liquid culture EC_{50} measurements allow for an estimation of potencies over a broad range.

This method significantly increases the number of hits that can be detected relative to visual or single-point OD

Table 2. Structure-Activity Relationships of 2,4-Diaminoquinazolines in *Saccharomyces cerevisiae*

Compound	R ¹	R ²	Halo ^a (raw halo score)	EC ₅₀ (μM)
4408-0546				4.09 (1.76 - 9.49)
4408-0539				5.94 (2.56 - 13.8)
4408-0537				16.26 (6.95 - 38.0)
4408-0549				42.2 (17.8 - 100)
5940-0056				57.94 (24.3 - 138)
4408-0144				ND ^b
7765-0010				ND ^b
C301-5029				ND ^b
4408-0520				ND ^b

^a Each line in the Halo image represents a plot of the OD600 across the site of pinning.
^b Compound showed no detectable activity.

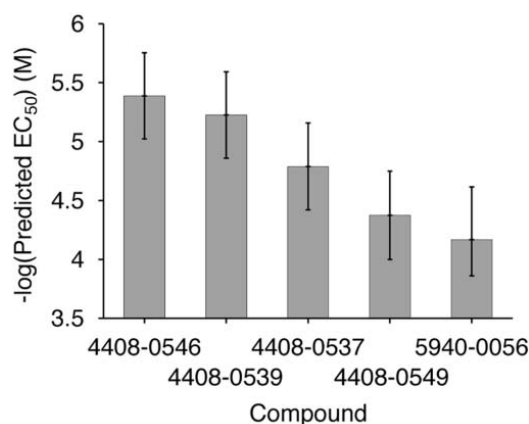


FIG. 5. Potency predictions for 2,4-diaminoquinazoline scaffold in *Saccharomyces cerevisiae*. The halo score-based EC_{50} predictions (y-axis) are shown for 5 modifications of the 2,4-diaminoquinazoline scaffold (ChemDiv compound code; x-axis). The uncertainty in the EC_{50} prediction is depicted by 95% prediction intervals for each compound (black error bars).

methods. A potential reason for this increase in hit rate is that the halo score is able to pick up compounds with weak effects that still produce characteristic halos. In addition, the current method picks up many hits that are obscured by edge and neighboring compound effects, indicating that the local background correction built into the halo score method helps deconvolute the signal from the noise for these cases.

A 21,120-member commercial library was screened, resulting in the identification of 590 active compounds in *S. cerevisiae*. Among the most active hits in yeast, several have known activities from previously reported screens in other organisms. The algorithm allowed quantification of structure-activity relationships (SAR), and trends were found for a cluster of 2,4-diaminoquinazolines (Fig. 5 and Table 2), a structural class with no previously reported antifungal activity.

The 4 most potent compounds in the 2,4-diaminoquinazoline structure/activity series (4408-0546, 4408-0539, 4408-0537, and 4408-0549) have been shown to modulate hepatocyte growth factor activity, suggesting that they may be useful in the treatment of cancer.⁵ The second-most bioactive (4408-0539) has been identified as having antimicrobial activity against *Escherichia coli* and *Pseudomonas aeruginosa*.⁶ In addition, 4408-0539 was identified in a cell-based assay as an inhibitor of the tyrosine kinase DYRK1A, a protein encoded on the critical region of chromosome 21 thought to be involved in learning and memory deficits associated with Down syndrome.⁷

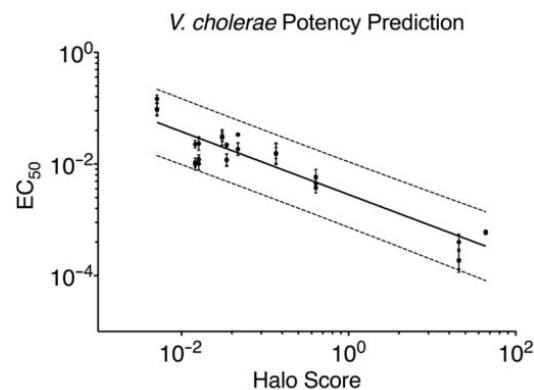


FIG. 6. Halo scores predict EC_{50} in *Vibrio cholerae*. Log-log plot of EC_{50} measurements obtained from liquid culture (y-axis) plotted against the halo score obtained in agar (x-axis) for 19 chemicals of varying toxicity against *V. cholerae*. The solid line represents the least squares regression line; dashed lines show the upper and lower bounds of a 95% prediction interval. Linear regression with 95% prediction intervals was performed using GraphPad Prism v5.0b software. The 2 estimates for a compound have an R^2 of 0.88 in log-log space, computed over EC_{50} potencies ranging from 186 nM to 150 μ M.

The least potent compound in the series (5940-0056) is reported to be toxic to mycobacteria.⁸

SAR analysis provides insights that may be useful in the synthesis of more potent derivatives and/or affinity reagents for future efforts aimed at target identification. In the series of compounds active in yeast, replacement of the 2-furanylmethyl group at the R1 position of the most potent 2,4-diaminoquinazoline, 4408-0546, results in a complete loss of activity for 4408-0144 (tertiary amine) and C301-5029 (primary amine), suggesting that a secondary aromatic amine in this position is important for bioactivity against yeast. There is a 2-fold reduction in potency when the 4-methoxy group of the phenyl in the R2 position (4408-0546) is replaced with bromine (4408-0539) and a 4-fold decrease in potency when the *p*-methoxybenzyl group is replaced with an *o*-methylbenzyl group (4408-0537). Changing the *p*-methoxy group of 4408-0546 to an *o*-methoxy substituent (4408-0549) results in a 10-fold decrease in activity, suggesting that para substitution in R2 is critical for activity. When both the 2-furanylmethyl of R1 and *p*-methoxybenzyl of R2 are replaced with benzyl groups (5940-0056), activity is reduced 14-fold. The SAR trends identified with the help of the halo score algorithm allow rapid determination of synthetic directions to take with hits against yeast. We are employing genetic and genomic approaches to identify molecular targets.

Identification of the target(s) of these compounds in yeast may lead to investigation of homologous targets in higher eukaryotes. In addition, their newly discovered antifungal

activity may be due to a novel biological mechanism. The algorithm was equally successful at identifying compound toxicity in microorganisms other than yeast. A calibration in wild-type *V. cholerae* resulted in accurate EC₅₀ estimation (see Fig. 6). The algorithm can be used to identify compounds with antibacterial activity with novel scaffolds. In addition, the screen can be used to predict the potency of natural products now that several new technologies have emerged for the expansion of libraries containing both crude and purified extracts.⁹ Thus, we expect the halo score-based method to generalize to many diverse organisms, as we have shown that it is useful for both a eukaryote and prokaryote.

ACKNOWLEDGMENTS

We thank F. Yildiz for providing *V. cholerae* stocks, K. Bilecen for help with *V. cholerae* agar plate preparation, L. Rocha for EC₅₀ determinations, and T. Wipke and J. Davis for assistance with chemical library clustering. RSL, NCG, and WCB were supported by the US Civilian Research and Development Foundation Grant Assistance Program (GTR-G7-044). MHW and JMS were supported by a fellowship from the Alfred P. Sloan Foundation.

REFERENCES

1. Gassner NC, Tamble CM, Bock JE, Cotton N, White KN, Tenney K, et al: Accelerating the discovery of biologically active small molecules using a high-throughput yeast halo assay. *J Nat Prod* 2007;70:383-390.
2. Yildiz FH, Schoolnik GK: *Vibrio cholerae* O1 El Tor: identification of a gene cluster required for the rugose colony type, exopolysaccharide

production, chlorine resistance, and biofilm formation. *Proc Natl Acad Sci USA* 1999;96:4028-4033.

3. Balakin KV, Kozintsev AV, Kiselyov AS, Savchuk NP: Rational design approaches to chemical libraries for hit identification. *Curr Drug Discov Technol* 2006;3:49-65.
4. Zhang JH, Chung TD, Oldenburg KR: A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen* 1999;4:67-73.
5. Zembower J, Mishra R: 2,4-Diaminoquinazoline compound modulators of hepatocyte growth factor/c-met activity, and use in the treatment of cancer and other dysproliferative diseases. US Patent Application Publication, 2006.
6. De La Fuente R, Sonawane ND, Arumainayagam D, Verkman AS: Small molecules with antimicrobial activity against *E. coli* and *P. aeruginosa* identified by high-throughput screening. *Br J Pharmacol* 2006;149:551-559.
7. Kim ND, Yoon J, Kim JH, Lee JT, Chon YS, Hwang M-K, et al: Putative therapeutic agents for the learning and memory deficits of people with Down syndrome. *Bioorg Med Chem Lett* 2006;16:3772-3776.
8. Wynne O, Johnson PD, Vickers R: Preparation of 2,4-diaminoquinazolines and analogs, and their use for the treatment of mycobacterial infections, especially tuberculosis. PCT International Application, 2008.
9. Koehn FE: High impact technologies for natural products screening. *Prog Drug Res* 2008;65:175, 177-210.

Address correspondence to:

Joshua M. Stuart
 Department of Biomolecular Engineering, University of California
 1156 High St., Santa Cruz, CA 95064
 E-mail: jstuart@soe.ucsc.edu

Scott Lokey
 Department of Chemistry and Biochemistry, University of California
 1156 High St., Santa Cruz, CA 95064
 E-mail: lokey@chemistry.ucsc.edu

SUPPLEMENTAL MATERIAL

SUPPLEMENTAL METHODS

Data Preprocessing

The density of cells in agar, the amount of agar, and other effects can vary across a plate. To mitigate the influence of these local fluctuations of cell density, we normalize the optical density readings by subtracting out three main location effects due to the solid agar assay. The raw, un-normalized OD measurement detected in the k^{th} reading of a well at row i , column j is a function of the amount of inhibition of a compound deposited in the well plus other effects due to the location of the reading on the plate and random noise. We can express these intuitions in the following linear model:

$$D_{ijk} = I_{ijk} + R_k + T_{ijk} + E_{ijk} + \eta_{ijk} \quad (1),$$

where the raw optical density readings on the plate for reading k at well (i,j) are D_{ijk} . Here, i is the row index ranging from 1 to n , j is the column index ranging from 1 to m , and k is the optical density reading across one pinning location, ranging from 1 to r . We use the triple (i,j,k) as a shorthand to refer to the k^{th} reading for a well located at row i and column j on a plate.

Because the raw D_{ijk} reading is convoluted with various systematic effects, we are interested in estimating the quantity I_{ijk} , which reflects the amount of inhibition due to the presence of compound(s) through the well. To estimate I_{ijk} , the above effects are subtracted from D_{ijk} : a read-to-pin effect R_k , a tilt effect T_{ijk} , and an edge effect E_{ijk} . In addition to these effects, each read is assumed to also reflect a small amount of zero-

centered Gaussian noise, η_{ijk} due to random fluctuations in local cell density (i.e. $\eta_{ijk} \sim N(0, \sigma)$ where σ is small relative to the other effects). The preprocessing described here could be used for various plate and read geometries; in the applications described in this manuscript, the plates had 16 rows ($n=16$), 24 columns ($m=24$), and each well had 9 readings per scanned well ($r=9$) that were spread symmetrically across a well in the horizontal direction. Effects are removed sequentially in the order that they appear in the right hand side of Eq. (2). Each normalization step is described below. All parameters estimated from a plate are represented with Greek symbols.

Read-to-pin normalization. The nine readings within a single well are taken at five distinct distances relative to the point where the compound was pinned into the agar. A slightly different amount of local cells are present as a function of a reading's distance relative to the point of pinning. Readings from the center of the well, nearest the site of pinning, have an elevated OD due to possible local compression of agar in the site surrounding pin penetration.

Let μ be the global mean optical density reading obtained for the plate, averaged over all readings from internal wells; i.e.:

$$\mu = \frac{1}{(n-2)(m-2)r} \sum_{i=2}^{n-1} \sum_{j=2}^{m-1} \sum_{k=1}^r D_{ijk} \quad (2),$$

where the average is taken only over internal wells that do not reside on any of the edges (i.e. not on the left-most or right-most column, and not on the top-most or bottom-most row). Using only internal wells helps avoid any influences due to edge effects (dealt with subsequently) during this normalization step. Using this global mean, we estimate the read-to-pin effect for reading k as the average deviation from this global mean over

the internal wells:

$$\hat{R}_k = \rho_k = \frac{1}{(n-2)(m-2)} \sum_{i=2}^{n-1} \sum_{j=2}^{m-1} (D_{ijk} - \mu) \quad (3).$$

We then use $D'_{ijk} = D_{ijk} - \hat{R}_k$ as the read-to-pin corrected density for subsequent normalization steps.

Tilt correction. The bench surface can be tilted in the x - and/or y -direction when agar is poured into a plate. During cooling, this can introduce a bias in the amount of cells deposited into an area of the plate as a function of their horizontal and vertical position. We correct for horizontal and vertical effects sequentially assuming we can treat the tilt in each direction independently: i.e. $T_{ijk} = T_{jk}^h + T_i^v$, where T_{jk}^h and T_i^v are tilt effects only in the horizontal and vertical directions respectively. Note that, since the nine reads in a well's scan are spread out in the horizontal direction in our application, the tilt effect is a function of read position in the horizontal, but not the vertical, dimension.

To estimate effects due to tilting, we make use of a set of background readings apparently unaffected by any compound(s). Because bioactive compounds are sparsely distributed and infrequent, most readings quantify areas where no compound has diffused across. Therefore, to estimate the background set of readings we identify internal readings (i.e. those not from a well on an edge) that have optical density levels very similar to the median optical density level on the plate. We label a reading as background if its optical density is within 10% of the median of all internal readings, λ , where

$$\lambda = \text{median}_{i=2, \dots, (n-1), j=2, \dots, (m-1), k=1, \dots, 9} \{D'_{ijk}\}. \text{ Reading } (i,j,k) \text{ is labeled as a background reading if } |D'_{ijk} - \lambda| <$$

0.1. The b^{th} background reading is then stored in a 3-column matrix, B , where $B_{b1}=i$, $B_{b2}=j$, and $B_{b3}=k$. Using readings with values near the median helps avoid those

containing inhibiting compounds as well as those registering infrequent optical anomalies.

Using the set of background readings, we first estimate the amount of tilt in the horizontal direction. To do this, we fit a regression line to all of the background readings using their x -coordinate on the plate as the predictor variable and their D'_{ijk} values as the response variable. The x -coordinates are in units of distance c , the distance between consecutive reads within a well. In our scans, the distance between the last reading in a well at column j and the first reading of the next well in column $j+1$ is approximately $2c$. With these coordinates, we use linear regression to solve for an overall slope in the horizontal direction, α that minimizes the following equation:

$$\alpha = \arg \min_s \left\{ \left\| \bar{x} \cdot s - \bar{d} \right\|_2 \right\} \quad (4),$$

where $\|\bar{v}\|_2$ is the Euclidean norm of a vector \bar{v} , and \bar{x} and \bar{d} are vectors containing the positions and optical densities of the background readings respectively:

$$\bar{x} : x_b = (r+1)(B_{b,2} - 1) + B_{b,3} \quad (5),$$

and

$$\bar{d} : d_b = D'_{B_{b,1}, B_{b,2}, B_{b,3}} \quad (6).$$

The solution to equation (4) gives an estimate of the slope in the horizontal direction so that the quantity $\hat{T}_{jk}^h = \alpha(r+1)j+k$ is used as an estimate for the horizontal tilt effect and the next normalized estimate for reading (i,j,k) is set to $D''_{ijk} = D'_{ijk} - \hat{T}_{jk}^h$ for all readings.

Tilts in the vertical direction are also corrected using an analogous approach. Because the plates are shorter and have fewer unique reads in the vertical direction, we use a more straightforward approach to estimate a linear tilt. Rather than use a regression, we

estimate the slope of the tilt using the median difference between consecutive rows of background readings. More formally, let δ_i be the median background level of row i after read-to-pin and horizontal tilt correction, i.e. $\delta_i = \underset{b:B_{b,1}=i}{\text{median}} \{D''_{i,B_{b,2},B_{b,3}}\}$. The vertical tilt effect is then estimated as the median difference between consecutive internal rows:

$$\hat{T}_i^v = \text{median}\{\delta_{i+1} - \delta_i\}_{i=2}^{n-2} \quad (7),$$

and set $D''_{ijk} = D''_{ijk} - \hat{T}_i^v$, for all readings (i,j,k) .

Edge correction. Due to the agar adhering to the edge of a plate, wells in the first and last row, or in the first and last column, tend to have higher scan values than other wells. We assume horizontal and vertical edge effects act independently for all readings within a well; i.e. $E_{ijk} = E_{jk}^h + E_i^v$ for all readings k at well (i,j) . Because more readings are available in the horizontal direction, its edge effect is quantified as a function of the readings, k , along the horizontal direction, while vertical edge effects are not a function of the read index. Each component of the edge effect is a simple partition function that detects if the well is on the perimeter of the plate. The edge effect in the horizontal direction is estimated from the difference between the columns on the edge and the internal columns next to the edge using each column's median level for the k^{th} reading:

$$\hat{E}_{jk}^h = \begin{cases} \text{median}\{D''_{ilk}\}_{i=2..n-1} - \text{median}\{D''_{ilr}\}_{i=2..n-1}; & \text{if } j = 1 \\ \text{median}\{D''_{im1}\}_{i=2..n-1} - \text{median}\{D''_{imk}\}_{i=2..n-1}; & \text{if } j = m \\ 0; & \text{otherwise} \end{cases} \quad (8).$$

The edge effect in the vertical direction is estimated from the median difference between all of the readings in the rows on the edge and the internal rows next to the edge:

$$\hat{E}_i^v = \begin{cases} \text{median}\left\{\left\{D_{1,jk}'' - D_{2,jk}''\right\}_{\substack{j=2..m-1 \\ k=1..r}}\right\}; & \text{if } i = 1 \\ \text{median}\left\{\left\{D_{n,jk}'' - D_{n-1,jk}''\right\}_{\substack{j=2..m-1 \\ k=1..r}}\right\}; & \text{if } i = n \\ 0; & \text{otherwise} \end{cases} \quad (9).$$

Note that all wells in the interior are assumed to have zero edge effect and corners have effects from both the horizontal and vertical directions. The final normalized optical density for reading (i,j,k) used in subsequent halo detection and quantification is obtained by subtracting off the edge effect: $\hat{P}_{ijk} = D_{ijk}''' - \hat{E}_j^h - \hat{E}_i^v$. These final normalized quantities are then used to detect the presence of bioactive compounds in a plate.

Halo detection and quantification

Using the normalized optical density readings, the presence of bioactive compounds producing characteristic halos of inhibited cell growth are detected and quantified. The halo results from the diffusion of compounds into the agar, which can be used to identify compounds over a wide range of toxicity. Compounds found to produce a zone of inhibition consistent with the shape of a halo are subsequently quantified using a score that reflects the overall amount of inhibition produced.

Potent compounds can inhibit cell growth spanning multiple wells in the plate. Therefore, the algorithm first detects multi-well halos, flags any wells that are included in any of these large halos, and then searches for single-well halos within the remaining wells. Thus, the overall procedure for detecting and scoring halos is to: 1) calculate single well features used for halo detection; 2) detect multi-well halos; 3) detect single-well halos; and then 4) quantify the level of overall inhibition within any multi- or single-well

halo. Each of these steps is described in more detail below.

Single well features. The pattern of normalized optical density across a single well is summarized by several discrete features. The features reflect the direction (but not the magnitude) of change in the horizontal and vertical directions across a well and whether the readings are lower than readings for the background lawn of cells. As features we include an indicator for whether the center reading is lower than the readings in the local background lawn of cells around the well (*UNDER*), an indicator for whether the center reading is lower than the well's end readings (*VSHAPED*), and whether the density is flat, rises, or falls in the horizontal direction across the well (*CHANGE*). We used a set of calibration plates to determine cutoffs for determining discrete values for these features (see Calibration section below).

Multi-well halo detection. We identify halos covering multiple wells by searching first for large halos and then smaller halos. This procedure allows us to exclude any wells found to be associated with large halos from consideration when searching for smaller halos. A halo is called if it matches a feature template matrix (FTM). The FTM detects whether a symmetric change in density is consistent for the readings surrounding a central well. If the patterns match the FTM, the well in the center is marked as the middle of a halo. For example, for the smallest multi-well halo search, we use the following 3-by-3 FTM centered on well (i,j) :

$$\begin{bmatrix} = - & + = & = + \\ + - & + = & + + \\ = - & + = & = + \end{bmatrix},$$

where each entry in the matrix represents the feature pair (*UNDER*, *CHANGE*) in which *UNDER* is set to '+' if the well reading's center must be under its local

background or '=' if it's irrelevant to the position, and *CHANGE* is set to '-' if the change in density from left to right must be falling, '+' if it must be rising, and '=' if the change is irrelevant.

Patterns of 9-by-9 grids are searched, then 5-by-5, and finally 3-by-3 patterns. Template areas occluded or off the edge are not considered in the template. Effectively, smaller templates are used in these cases so that halos on edges and corners can still be detected. We screen left to right, and top to bottom. The detection accuracy could be increased by considering additional traverses since the traversing leads to an order-dependent detection method. However, in practice, this does not affect many calls.

Single-well halo detection. Halos occupying a single well are more straightforward to detect. Any wells not flagged as associated with a multi-well halo are considered for the presence of a single-well halo. A single-well halo is detected if either 1) the well's *VSHAPED* feature is true or 2) the well's left neighbor has a falling *CHANGE* and its right neighbor has a rising *CHANGE*.

Raw halo score calculation. The amount of total inhibition is quantified for wells detected to be centered on a halo. We calculate a raw halo score for all halos whether they span multiple or single wells. Given a well centered on a halo, we find those readings representing the left and right boundaries of the halo, which may include readings from neighboring wells. The left and the right boundary are identified as the readings where the change in normalized optical density is close to zero (i.e. within a tolerance threshold determined from calibration plates; see below). Because the background lawn of cells can fluctuate somewhat across a plate, the baseline level used for quantification is the minimum density of the two readings at the left and right

boundaries. The raw halo score of well (i,j) reflects the amount of inhibition beyond this baseline level and is computed using all reads within the boundaries:

$$R(i, j) = \sum_{k=lb(i,j)}^{rb(i,j)} (D_{ijk}''' - B(i, j)) \quad (10),$$

where $lb(i,j)$ and $rb(i,j)$ are the left and right boundary of the well and $B(i,j)$ is the baseline level for the well; i.e. $B(i, j) = \min\{D_{i,j,lb(i,j)}''', D_{i,j,rb(i,j)}'''\}$. Intuitively, the raw halo score sums up all of the readings to the left, right, and center of a halo. Finally, the halo score, H , is computed by dividing the raw halo score R by the concentration of the compound that was pinned onto the plate. The halo score is used to predict compound potency.

Calibration to determine thresholds for halo detection

We used a set of calibration plates to determine cutoffs for classifying wells as halo shaped including a horizontal slope threshold t_h , and a V -shape threshold t_v used to set the *UNDER* and *VSHAPED* flags respectively. The thresholds were empirically determined by iteratively inspecting detected halos and tuning the parameters until no false negatives were found. For a given set of thresholds, the halo detection method was run. Any wells found in disagreement with a human expert were examined to determine which portions of the algorithm were at fault and the appropriate threshold was adjusted.

Compounds discarded from analysis

A minority of compounds were discarded for several reasons due to occasional technical complications of the experimental procedure. Compounds that neighbored large halos

such that the well they were deposited into were obscured. Some compounds stuck to pin heads and were carried over from a previous plate which affected the reading of a second compound. In a limited number of cases, pins damaged the agar making OD assessment unreliable. Finally, for EC50 to halo score correlation analysis, compounds that were also found to be problematic in the liquid culture were also discarded.

Table S3. List of compounds used to correlate halo score versus liquid culture-based EC50s for *V. cholerae*.

Compound	Halo Score	Liquid EC50		
		Repeat 1	Repeat 2	Repeat 3
Alverine Citrate	0.0051	147.00	93.42	97.00
Chlorpromaxine	0.047	33.68	18.57	18.48
Ciclopiroxolamine	0.13	15.72	15.39	15.19
Clotrimazole	0.015	22.73	10.63	9.99
Disulfiram	0.016	23.29	12.05	10.16
Novobiocin	44	0.61	0.59	0.58
Phenanthroline	0.031	31.34	30.33	29.93
Streptozocin	0.41	3.79	4.53	5.88
Thimerosal	21	0.19	0.19	0.40
Thujaplicin	0.035	22.16	11.97	11.70

SUPPLEMENTAL TABLES

Table S1. Compounds used in EC50 vs. Halo Score correlation calculations.

Compounds are listed in column 1, those that are used in the *S. cerevisiae* EC50 to halo score correlation have a * in column 2, those used in *V. cholera* have a * in column 3.

Compound	<i>S. cerevisiae</i>	<i>V. cholera</i>
Alverine Citrate		*
Chlorpromazine Hydrochloride	*	*
Ciclopiroxolamine	*	*
Clotrimazole	*	*
Disulfiram		*
Fenpropimorph	*	
Hygromycin	*	
Ketoconazole	*	
MMS	*	
Novobiocin		*
Nystatin	*	
Phenanthroline Monohydrate	*	*
Rapamycin	*	
Streptozocin		*
Tamoxifen	*	
Terbinafine	*	
Thimerosal	*	*
Thujaplicin	*	*
Tunicamycin	*	

Table S2. List of compounds used to correlate halo score versus liquid culture-based EC50s for *S. cerevisiae*.

Compound	Halo Score	Liquid EC50
Tunicamycin	13	0.072
Fenpropimorph	1.4	0.681
Nystatin	0.5	2.388
Tamoxifen	0.043	12.034
Phenanthroline_Monohyate	0.081	13.485
Terbinafine	0.12	9.271
Rapamycin	168	0.014
MMS	0.0014	1325.214
Clotrimazole	1.3	1.177
Thimersol	3.6	0.243
Ketoconazole	3.0	0.451
Chlorpromazine_Hydrochloride	0.051	15.16
Ciclopiroxolamine	0.018	38.819
Hygromycin	0.038	20.535
Thujaplicin	0.025	32.49

Appendix 3 - BioSpace supplementary data

ORF	Gene	SGD ID
YJR043C	POL32	S000003804
YER116C	SLX8	S000000918
YAL024C	LTE1	S000000022
YDL117W	CYK3	S000002275
YDR334W	SWR1	S000002742
YGR270W	YTA7	S000003502
YPR135W	CTF4	S000006339
YDR439W	LRS4	S000002847
YCR065W	HCM1	S000000661
YOL012C	HTZ1	S000005372
YNL298W	CLA4	S000005242
YGL058W	RAD6	S000003026
YOR195W	SLK19	S000005721
YMR070W	MOT3	S000004674
YJL164C	TPK1	S000003700
YGL016W	KAP122	S000002984
YJR135C	MCM22	S000003896
YNL072W	RNH201	S000005016
YBR023C	CHS3	S000000227
YDR260C	SWM1	S000002668
YPL115C	BEM3	S000006036
YHR167W	THP2	S000001210
YPL051W	ARL3	S000005972
YOR275C	RIM20	S000005801
YNL041C	COG6	S000004986

Table 6 - List of genes chosen as BioSpace screening knockouts.

Bibliography

- ABPI (2009) ABPI Resources for School. ASSOCIATION OF THE BRITISH PHARMACEUTICAL INDUSTRY, pp. Agar plate showing Halo Asscay.
- Ashburner, M., *et al.* (2000) Gene Ontology: tool for the unification of biology, *Nature genetics*, 25, 25-29.
- Barberis, A., *et al.* (2005) Yeast as a screening tool, *Drug Discovery Today: Technologies*, 2, 187-192.
- Bender, A. (2005) Studies on Molecular Similarity. University of Cambridge, pp. 182.
- Bender, A. and Glen, R.C. (2005) Molecular similarity: a key technique in molecular informatics., *Organic Biomolecular Chemistry*, 2, 15.
- Bergman, L.W. (2001) Growth and maintenance of yeast. In, *Two-Hybrid Systems*. Springer, pp. 9-39.
- Bliss, C.I. (1939) The Toxicity of Poisons Applied Jointly, *Annals of Applied Biology*, 26, 585-615.
- Boone, C. (2014) Yeast Systems Biology: Our Best Shot at Modeling a Cell, *Genetics*, 198, 435-437.
- Brown, I., *et al.* (1997) Antigenic and genetic analyses of H1N1 influenza A viruses from European pigs, *Journal of General Virology*, 78, 553-562.
- Cai, C., *et al.* (2011) RETRACTED: Nef from SIVmac239 decreases proliferation and migration of adenoid-cystic carcinoma cells and inhibits angiogenesis, *Oral Oncology*, 47, 847-854.

- Chain, E., *et al.* (1940) Penicillin as a Chemotherapeutic Agent, *Lancet*, ii, 13.
- Cohen, M.L. (1992) Epidemiology of drug resistance: implications for a post-antimicrobial era, *Science*, 257, 1050-1055.
- Costanzo, M., *et al.* (2010) The Genetic Landscape of a Cell, *Science*, 327, 425-431.
- Costanzo, M., *et al.* (2011) Charting the genetic interaction map of a cell, *Current opinion in biotechnology*, 22, 66-74.
- Darvas, F., *et al.* (2002) In silico and ex silico ADME approaches for drug discovery, *Current topics in medicinal chemistry*, 2, 1287-1304.
- Davies, J. (2006) Where have all the antibiotics gone?, *The Canadian Journal of Infectious Diseases & Medical Microbiology*, 17, 287.
- Daylight Chemical Informations Systems (2011) *Daylight Theory Manual v4.9*.
- DeLong, E.F. (2007) Modern microbial seascapes. Forward, *Nat Rev Microbiol*, 5, 755-757.
- DeWitt, S.H. and Czarnik, A.W. (1995) Automated synthesis and combinatorial chemistry, *Current opinion in biotechnology*, 6, 640-645.
- Deza, M.M. and Deza, E. (2009) *Encyclopedia of distances*. Springer.
- Dunstan, H.M., *et al.* (2002) Cell-based assays for identification of novel double-strand break-inducing agents, *Journal of the National Cancer Institute*, 94, 88-94.
- Durant, J.L., *et al.* (2002) Reoptimization of MDL keys for use in drug discovery, *J Chem Inf Comput Sci*, 42, 1273-1280.
- Eckert, H. and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches, *Drug Discov Today*, 12, 225-233.
- Eisen, M.B., *et al.* (1998) Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences*, 95, 14863-14868.
- Ericson, E., *et al.* (2008) Off-target effects of psychoactive drugs revealed by genome-wide assays in yeast, *PLoS genetics*, 4, e1000151.
- Fabricant, D.S. and Farnsworth, N.R. (2001) The value of plants used in traditional medicine for drug discovery, *Environmental health perspectives*, 109, 69.
- Fauci, A.S. (2005) Emerging and reemerging infectious diseases: the perpetual challenge, *Academic Medicine*, 80, 1079-1085.

- Feng, Y., *et al.* (2009) Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds, *Nature Reviews Drug Discovery*, 8, 567-578.
- Fleming, A. (1929) On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. influenzae, *British Journal of Experimental Pathology*, 10, 11.
- Fraser, T.R. (1872) The antagonism between the actions of active substances, *The British Medical Journal*, 2, 457-459.
- Gao, F., *et al.* (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes, *Nature*, 397, 436-441.
- Gassner, N.C., *et al.* (2007) Accelerating the Discovery of Biologically Active Small Molecules Using a High-Throughput Yeast Halo Assay \perp , *J Nat Prod*, 70, 383-390.
- Giaever, G., *et al.* (2002) Functional profiling of the Saccharomyces cerevisiae genome, *Nature*, 418, 387-391.
- Guttman, L. (1954) Some necessary conditions for common-factor analysis, *Psychometrika*, 19, 149-161.
- Hartman, J.L., Garvik, B. and Hartwell, L. (2001) Principles for the buffering of genetic variation, *Science*, 291, 1001-1004.
- Holliday, J.D., Hu, C. and Willett, P. (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings, *Comb Chem High T Scr*, 5, 155-166.
- Hu, Y., Lounkine, E. and Bajorath, J. (2009) Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density-Dependent Similarity Function, *ChemMedChem*, 4, 540-548.
- Hughes, T.R. (2002) Yeast and drug discovery, *Functional & integrative genomics*, 2, 199-211.
- Johnson, M.A. and Maggiora, G.M. (1990) *Concepts and applications of molecular similarity*. John Wiley & Sons, New York.
- Kaiser, H.F. (1960) The application of electronic computers to factor analysis, *Educational and psychological measurement*.
- Kaiser, H.F. (1970) A second generation little jiffy, *Psychometrika*, 35, 401-415.

- Khurana, V. and Lindquist, S. (2010) Modelling neurodegeneration in *Saccharomyces cerevisiae*: why cook with baker's yeast?, *Nature Reviews Neuroscience*, 11, 436-449.
- Koopmans, M., Martens, D. and Wijffels, R.H. (2009) Towards commercial production of sponge medicines, *Marine drugs*, 7, 787-802.
- Kuzmin, E., *et al.* (2014) Synthetic Genetic Array Analysis for Global Mapping of Genetic Networks in Yeast. In Smith, J.S. and Burke, D.J. (eds), *Yeast Genomics*. Humana Press, New York, pp. 143-168.
- Levy, S.B. (2001) Antibacterial household products: cause for concern, *Emerg Infect Dis*, 7, 512-515.
- Lewis, D.B. (2006) Avian flu to human influenza, *Annu. Rev. Med.*, 57, 139-154.
- Lucey, B.P., Nelson-Rees, W.A. and Hutchins, G.M. (2009) Henrietta Lacks, HeLa cells, and cell culture contamination, *Archives of pathology & laboratory medicine*, 133, 1463-1467.
- Mager, W.H. and Winderickx, J. (2005) Yeast as a model for medical and medicinal research, *Trends in Pharmacological Sciences*, 26, 265-273.
- Martin, Y.C., Kofron, J.L. and Traphagen, L.M. (2002) Do Structurally Similar Molecules Have Similar Biological Activity, *J Med Chem*, 45, 4350-4358.
- Massey Jr, F.J. (1951) The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American statistical Association*, 46, 68-78.
- Miller, C.M. and McCarthy, F.O. (2012) Isolation, biological activity and synthesis of the natural product ellipticine and related pyridocarbazoles, *RSC Advances*, 2, 8883-8918.
- Monga, M. and Sausville, E. (2002) Developmental therapeutics program at the NCI: molecular target and drug discovery process, *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K.*, 16, 520-526.
- Morse, S.S. (1995) Factors in the Emergence of Infectious Diseases, *Emerging Infectious Diseases*, 1, 7-16.
- Newman, D.J. and Cragg, G.M. (2007) Natural Products as Sources of New Drugs over the Last 25 Years, *J Nat Prod*, 70, 461-477.
- Ng, D.M. and Woehrmann, M.H. (2007) ClueGene v2.0 – Using a Compendium of Clusters for Pathway Gene Recommendation. ISM245, UCSC.

- Ng, D.M., Woehrmann, M.H. and Stuart, J.M. (2007) Recommending pathway genes using a compendium of clustering solutions. *Pacific Symposium on Biocomputing*. pp. 379-390.
- Norcliffe, J., *et al.* (2014) The utility of yeast as a tool for cell-based, target-directed high-throughput screening, *Parasitology*, 141, 8-16.
- Parsons, A.B., *et al.* (2003) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways, *Nature Biotechnology*, 22, 62-69.
- Pearson, K. (1895) Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London*, 58, 240-242.
- Peplow, M. (2014) Organic synthesis: The robo-chemist, *Nature*, 512, 3.
- Perlman, Z.E., *et al.* (2004) Multidimensional drug profiling by automated microscopy, *Science*, 306, 1194-1198.
- Pierce, S.E., *et al.* (2006) A unique and universal molecular barcode array, *Nature Methods*, 3, 601-603.
- Ramos, F.J., *et al.* (2012) Rapamycin reverses elevated mTORC1 signaling in lamin A/C-deficient mice, rescues cardiac and skeletal muscle function, and extends survival, *Science translational medicine*, 4, 144ra103.
- Rincon, P. (2006) 'Faster emergence' for diseases. BBC, newvote.bbc.co.uk.
- Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints, *J Chem Inf Model*, 50, 742-754.
- Rogers, D.J. and Tanimoto, T.T. (1960) A Computer Program for Classifying Plants, *Science*, 132, 1115-1118.
- Saunders, R.N., Metcalfe, M.S. and Nicholson, M.L. (2001) Rapamycin in transplantation: a review of the evidence, *Kidney Int*, 59, 3-16.
- Seto, B. (2012) Rapamycin and mTOR: a serendipitous discovery and implications for breast cancer, *Clinical and translational medicine*, 1, 29-35.
- Simon, J.A. and Bedalov, A. (2004) Yeast as a model system for anticancer drug discovery, *Nature Reviews Cancer*, 4, 1-8.
- Smith, M.G. and Snyder, M. (2006) Yeast as a Model for Human Disease. In, *Current Protocols in Human Genetics*. John Wiley & Sons, Inc., pp. 15.16.11-15.16.18.
- Spearman, C. (1904) The proof and measurement of association between two things, *The American journal of psychology*, 15, 72-101.

- Spilman, P., *et al.* (2010) Inhibition of mTOR by rapamycin abolishes cognitive deficits and reduces amyloid-beta levels in a mouse model of Alzheimer's disease, *PloS one*, 5, e9979.
- Steinmetz, L.M., *et al.* (2002) Systematic screen for human disease genes in yeast, *Nature Genetics*.
- Suffness, M. and Douros, J. (1982) Current status of the NCI plant and animal product program, *J Nat Prod*, 45, 1-14.
- Sussex Drug Discovery (2013) Predicting cancer targets modulated by Ayurvedic medicines.
- Tamble, C.M. (2008) Small molecule screening and characterization in *Saccharomyces cerevisiae* and mammalian cells. *Chemistry*. UC Santa Cruz, pp. 135.
- Tanimoto, T.T. (1957) IBM Internal Report. IBM Corp, Armonk, New York.
- Tong, A.H.Y., *et al.* (2004) Global mapping of the yeast genetic interaction network, *science*, 303, 808-813.
- Uetz, P., *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403, 623-627.
- Vezina, C., Kudelski, A. and Sehgal, S.N. (1975) Rapamycin (AY-22,989), a new antifungal antibiotic. I. Taxonomy of the producing streptomycete and isolation of the active principle, *The Journal of antibiotics*, 28, 721-726.
- Wang, Y., *et al.* (2013) PubChem BioAssay: 2014 update, *Nucleic acids research*, gkt978.
- Warner, L.M., Adams, L. and Sehgal, S. (1994) Rapamycin Prolongs Survival and Arrests Pathophysiologic Changes in Murine Systemic Lupus Erythematosus, *Arthritis & Rheumatism*, 37, 9.
- Woehrmann, M.H., *et al.* (2013) Large-scale cytological profiling for functional analysis of bioactive compounds, *Molecular BioSystems*, 9, 2604-2617.
- Woehrmann, M.H., *et al.* (2010) HALO384: A Halo-Based Potency Prediction Algorithm for High-Throughput Detection of Antimicrobial Agents, *Journal of Biomolecular Screening*, 15, 196-205.
- Woolhouse, M.E. and Gowtage-Sequeria, S. (2005) Host Range and Emerging and Reemerging Pathogens, *Emerging Infectious Diseases*, 11, 1842-1847.

Young, D.W., *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action, *Nature chemical biology*, 4, 59-68.

Zhang, J.H., Chung, T.D. and Oldenburg, K.R. (1999) A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays, *J Biomol Screen*, 4, 67-73.