BY KATHARINE MILLER

BIG DATA AND DRUGS: BD2K CENTERS SOLIDIFY EMERGING APPROACHES

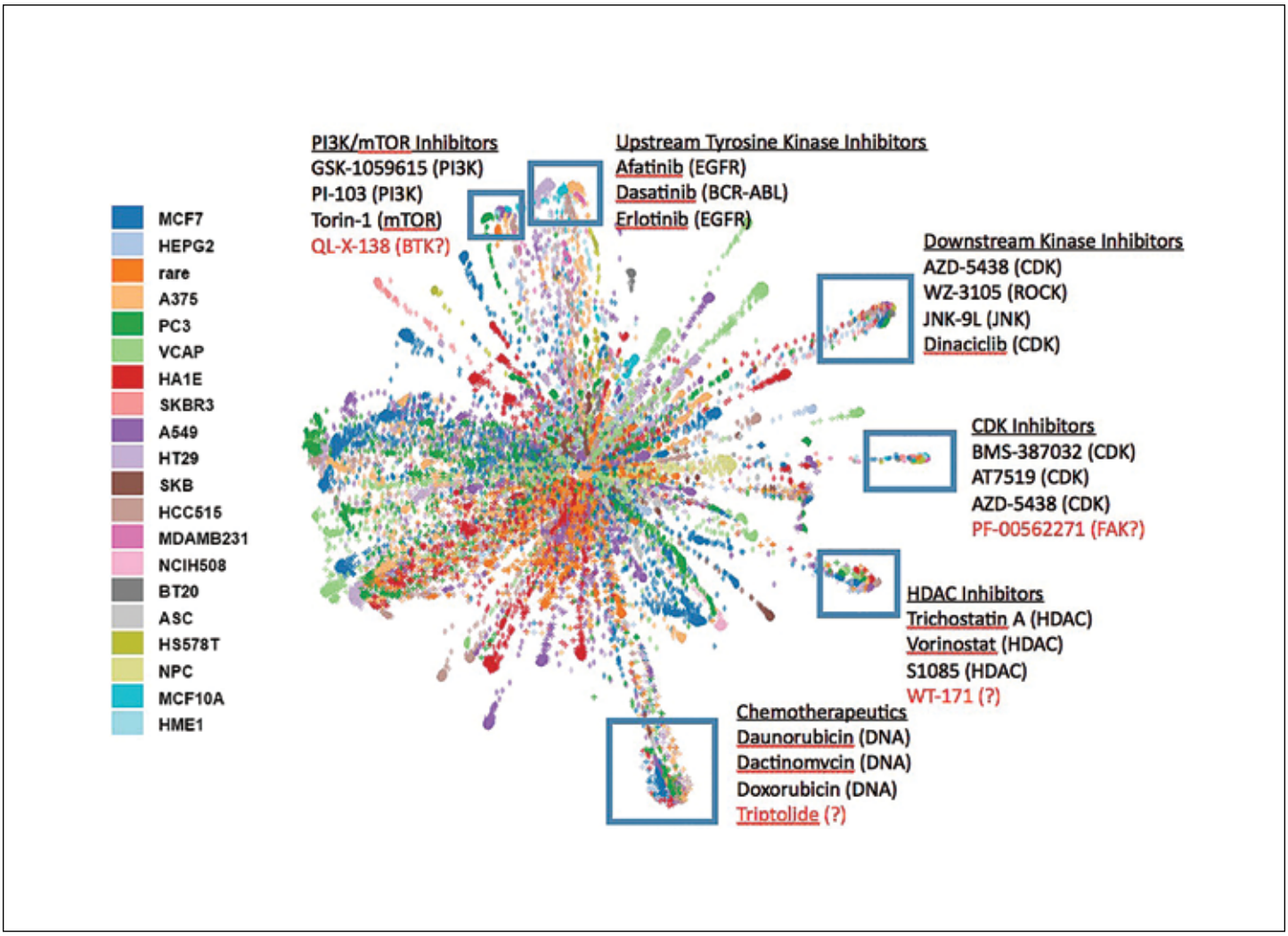
The Big Data era in biomedicine offers a grand promise: that by crunching vast quantities of multi-omics data through appropriate statistical analyses, researchers will gain a comprehensive understanding of health and disease that will lead to new, effective, and personalized treatment options.

Current work in systems pharmacology by several BD2K Centers offers a glimpse at that potential. Researchers at the **BD2K-LINCS-DCIC** are mapping the global space of responses of human cells to many drugs and other small molecules as well as exploring the universe of drug-induced adverse effects. Those at **KnowEnG** are analyzing multi-omics measurements in hopes of understanding drug response in cancer patients. Meanwhile, the **PIC-SURE Center** is standardizing procedures and developing open-source tools for drug repositioning research.

Taken together, BD2K Centers' systems pharmacology work solidifies a set of high-quality approaches to the field, giving hope that one day the grand promise of big data will be realized.

Mapping the Drug Universe

By determining how various healthy and diseased cells respond to a wide range



This fireworks plot displays the universe of cellular responses to drugs. Each spot represents one of 17,041 significant drug-induced gene expression signatures for 3,713 drugs and other compounds applied to 63 cell lines in 3 time points and 51 dosages. Colors represent different cell types while the boxes indicate a few of the cellular states induced by specific types of drugs. This visualization enables assigning function and mechanism to new small molecules, suggesting their potential to serve as drugs. Courtesy of the Ma'ayan Lab and the BD2K-LINCS-DCIC.

RELEVANT NIH INSTITUTES:

NCI, NHLBI, NIDDK, NINDS and all other disease-focused Institutes

of perturbations (by drugs and other chemical compounds in varying doses; reagents that mutate, activate and deactivate genes; and changing the microenvironment), researchers could perhaps map the universe of cellular phenotypes and drug responses. That is one goal of the NIH's Library of Integrated Network-based Cellular Signatures (LINCS) program. Now in Phase II, the LINCS program has generated a vast quantity of gene expression, proteomic and epigenetic data, and the LINCS-DCIC (a BD2K Center) is building that map.

"We are interested in mapping the chemical space to the cellular phenotype space through the molecular signature space, and that will give us a global view of cells, all their states, how they respond to small molecules, and then how those match to cellular phenotypes and diseases and drugs," says **Avi Ma'ayan, PhD**, principal investigator of the BD2K-LINCS-DCIC.

At this point, Ma'ayan and his colleagues are building an interactive web page that will report—for many of the drugs studied by LINCS researchers—the pathways that a drug potentially targets, the genes that are up/down regulated, and other small molecules that are similar to that drug. "We're trying to visualize this space of drug perturbations," Ma'ayan says. The result is a plot of a network that reveals how small molecules in general affect gene expression and cluster into several responses associated with cellular phenotypes. In this global picture of what happens to cells when they are exposed to drugs, the space of responses is not infinite. "It's likely going to be about 100 states that can be well-defined and those states can be associated with disease states, and then you can use drugs to manipulate the system in the direction that you want," Ma'ayan says.

This kind of map is a global goal for biomedical research in general, Ma'ayan says. "By pushing cells in different directions, drugs make a perfect case study."

Predicting Drug Response

Mayo Clinic cancer researchers associated with the KnowEnG Center are also perturbing cells but with a different goal: They are most interested in which cells die in response to chemotherapy drugs. "We know that the same drug given to different patients elicits different responses," says **Saurabh Sinha, PhD**, principal investigator of the KnowEnG Center. "So this is just repeating that observation in a controlled setting in cell lines." The resistant cells are the problem: "You'd like to know why they are resistant," he says. So, for each individual cell line, the researchers also sequence the DNA and measure gene expression and methylation patterns before treatment. The goal: to determine whether these high-dimensional data (millions of gene variants and DNA methylation spots as well as tens of thousands of gene expression measurements) can accurately predict whether a particular drug would or would not work on a particular patient.

To tackle the computational and statistical challenges of relating multi-omics data to phenotype, KnowEnG researchers took several different lines of attack. One example: Even if they couldn't accurately predict the response of each individual, could they at least identify the most important genes whose variation from individual to individual are predictive of the phenotypic differences? "It might not be a 90 to 100 percent accurate final model," Sinha says, "but if we can identify the most significant genes related to the underlying biology, then we can follow up with more targeted biological studies."

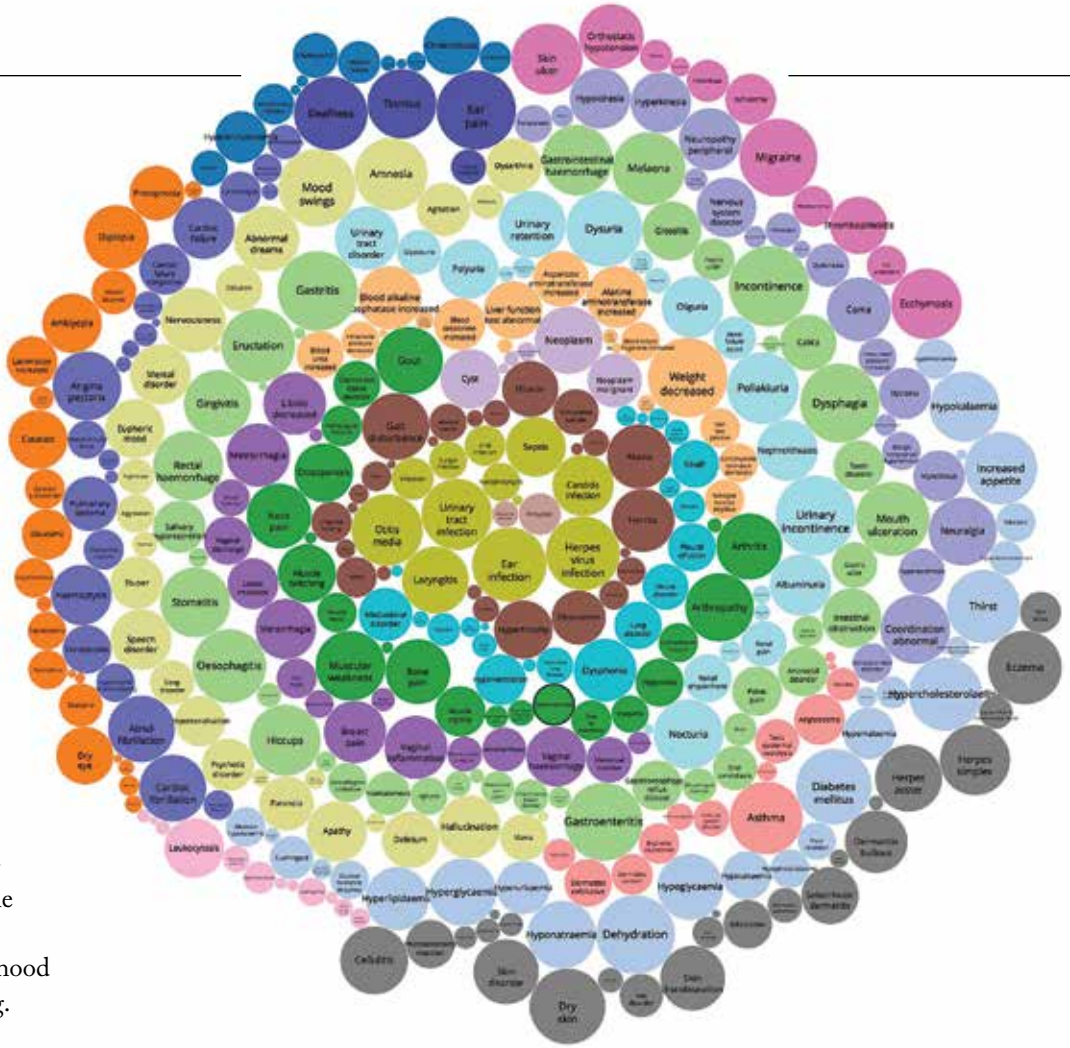
As another example, they set out to identify pathways (rather than individual genes) implicated in drug response. Genes tend to work together as part of complicated pathways of interaction. "Are there pathways triggered or not triggered leading to differences in the phenotype?" Sinha says. If so, then follow up studies can confirm findings and perhaps design drugs to target those pathways.

A third strategy traced gene expression back to the transcription factor (TF) responsible for controlling that gene expression. "If we find that a whole bunch of genes are changing their expression levels in a particular individual, then it's reasonable to hypothesize that these changes were regulated by some transcription factor," Sinha says. Instead of predicting individual genes as key players, this approach predicts that one

"We are interested in mapping the chemical space to the cellular phenotype space through the molecular signature space, and that will give us a global view of cells, all their states, how they respond to small molecules, and then how those match to cellular phenotypes and diseases and drugs," says Avi Ma'ayan.

transcription factor is an important regulator of those key players. "This has the possibility of statistically reducing the noise," Sinha says. And in fact they found that to be the case. "We were able to identify a small number of TFs for each drug that might play a role in drug response variation," he says. And they experimentally validated their results for several drugs by knocking down TFs and seeing the expected drug response changes. These results could also help in designing appropriate ways to overcome chemotherapy resistance.

The team is also working on the original problem of building a predictor of drug response levels using all the multi-omics data with the intent of outputting a single number: the likelihood that the patient will respond to a drug.



LINCS-DCIC combined drug structural information with gene expression profiles to predict adverse drug reactions for the 20,412 drugs and small-molecule compounds profiled by the LINCS L1000 project. These bubble plots show distinctly different sorting pattern for the side effects when sorted by the system and organ affected (above) versus by drug similarity (opposite). The team also created a freely available web portal at http://maayanlab.net/SEP-L1000/ where each drug and adverse drug reaction has a dedicated page with a list of the relevant predictions and external links to relevant sites. Courtesy of the Ma'ayan Lab and the BD2K-LINCS-DCIC.

Predicting Adverse Drug Reactions

When some friends at the FDA approached Ma'ayan to see if the LINCS' gene expression data could predict adverse drug reactions for a specified group of drugs, he gave a surprising response: "We can do it for all drugs."

Other researchers have tried to predict side effects from drug structure alone. Ma'ayan's group integrated that structural information with LINCS gene expression signatures for 20,000 compounds (including the subset of FDA-approved drugs) and showed that combining these two types of information improved adverse drug event predictions. "This can be helpful to the FDA, which could use computational methods to assess potential toxicity of new compounds," Ma'ayan says. LINCS-DCIC also developed a web portal for browsing and searching connections between small molecules and adverse drug reactions.

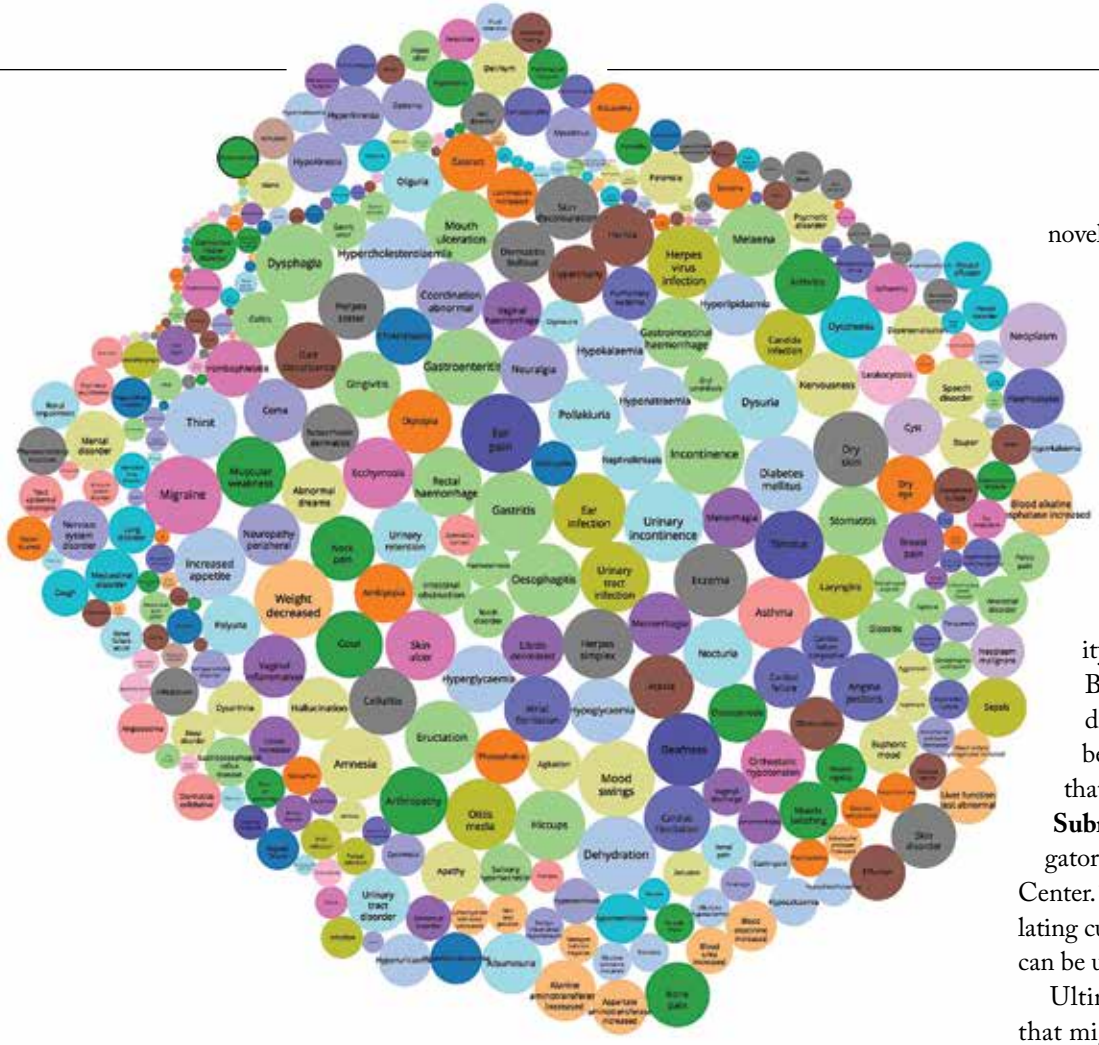
Right now, Ma'ayan says, "This is ready as a suggestive tool, not as a primary approach." With time, these kinds of computational approaches will become mainstream, he says.

Drug Repositioning Tools

Several BD2K Centers are involved in the computational effort to discover new uses for existing FDA-approved drugs. This makes a lot of sense: Big data will likely prove useful in this effort, and computational drug repositioning can save a lot of money while benefiting many patients. At PIC-SURE, researchers in **Chirag Patel**'s lab have developed tools that will make it easier for anyone to do drug repositioning research.

Frustrated that existing drug repositioning tools required specific data sources or formats, they created a tool called ksRepo that allows researchers to greatly expand the datasets usable to generate predictions about potential drug repositioning candidates.

In addition, concerned that computational researchers were each using a different database to validate their drug repositioning methods, they developed repoDB, a set of standardized drug successes and failures drawn from DrugCentral and ClinicalTrials.gov. "It's important to have a consistent benchmark set that everyone uses so you can say, 'my method outperforms this method using this same benchmark,'" says **Adam Brown**, a graduate student in biomedical informatics at Harvard Medical School and member of the PIC-SURE team. "Without that consistency, you just cherry-pick the dataset that fits your story." Many researchers were also calculating sensitivity and specificity without true negatives (failed drug candidates). RepoDB addressed that problem as



well. "To our knowledge, it's the only database that includes both approved and failed drugs," Brown says. "This is something the field has really been missing."

RepoDB will be particularly useful in studies where researchers are trying to predict associations between all diseases and all drugs, Brown says. "Hopefully people will use it."

In another effort to make drug repositioning research more reproducible, the **Broad Institute's LINCS Center for Transcriptomics and Toxicology** (**LINCS-Transcriptomics**) is creating a novel comprehensive screening library called the Broad Drug Repurposing Hub. First they identified and created a physical collection of 5,000 compounds, including more than 3,000 drugs of interest, and they curated them as a means of quality control. They then distributed them to anyone interested in screening them in their assays (gene expression, cytotoxicity, proteomics and morphology). But there is a hitch: "It's a hub to distribute reagents, with the price being contributing the data back so that others can use it," says **Aravind Subramanian, PhD**, principal investigator for the LINCS-Transcriptomics Center. The hub has already begun accumulating curated, quality-controlled data that can be used for drug-repositioning research.

Ultimately, identifying existing drugs that might cure or alleviate symptoms of rare diseases could give patients hope of a treatment. "That's something I'm pretty passionate about," Brown says. "It's important to get good drug/disease pairs into the hands of clinicians."

BD2K Systems Pharmacology in Context

Plenty of systems pharmacology research happens beyond the BD2K context, Sinha says. But the BD2K Centers have brought a big picture view to the field as well as a sense of gravitas: Doing this research well and reproducibly requires reliable data such as that generated by the LINCS program; well-designed and validated analytical tools such as those BD2K-LINCS-DCIC and KnowENG are building; and quality controls, incentives for data-sharing, and standardized benchmarking and validation procedures such as those being modeled and made publicly available by PIC-SURE and the LINCS-Transcriptomics Center. □

DETAILS

BD2K Drug Repositioning Tools

PIC-SURE:

ksRepo: a generalized tool that expands the datasets usable to generate predictions about potential drug repositioning candidates (freely available for download at https://github.com/adam-sam-brown/ksRepo)

RepoDB: a standard set of drug repositioning successes and failures that can be used to fairly and reproducibly benchmark computational repositioning methods. (freely available for download at http://apps.chiragjpgroup.org/repoDB/)

MeSHDD: uses MeSH-term enrichment to discover literature-based similarities between FDA approved drugs (interactive online app at http://apps.chiragjpgroup.org/MeSHDD/)

Broad-Transcriptomics

Broad Drug Repurposing Hub: a best-in-class drug screening collection with more than 3,000 clinical drugs (https://clue.io/repurposing)