BY KATHARINE MILLER

THE HARMONIZOME:

A Prototype for Integrated Datasets

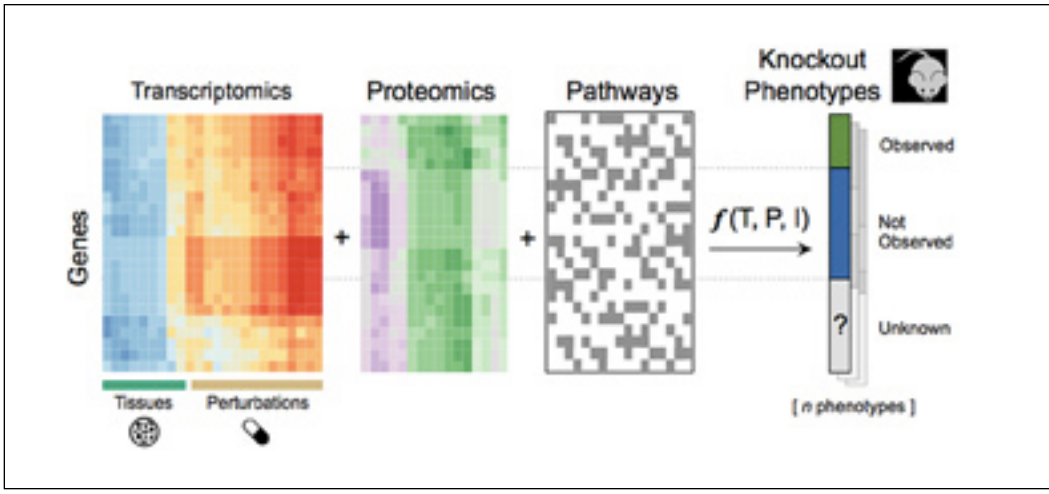
For years now, biocomputational scientists have been talking about the need for better data integration. "People talk a lot about how data are in silos and not connected," says **Avi Ma'ayan, PhD**, professor of pharmacological sciences at the Icahn School of Medicine at Mount Sinai and principal investigator of the BD2K-LINCS Data Coordination and Integration Center. After all that talk, Ma'ayan and his colleagues figured it was time to take action. So they created the Harmonizome, a collection of all the hottest and most exciting databases that everyone is using. "It allows you to find knowledge about genes and proteins that was buried in those silos but now is accessible."

To create the Harmonizome, Ma'ayan's team gathered together the major omics databases as well as databases on mouse and human phenotypes and processed them into a relatively simple format. That processing involved taking either raw data or formatted data from existing databases and mapping it onto common IDs for genes. They also processed the data into simplified formats such as relational tables, making it ready for machine learning. "It makes it very easy for someone to do predictions of functions for genes," Ma'ayan says.

That's what makes Ma'ayan most excited—the potential for using the Harmonizome to impute knowledge across data resources. His favorite example thus far, which was included among other examples in a paper about the Harmonizome published in *Database* in 2016, involves the prediction of mouse phenotypes. Using the Harmonizome, his team was able to create tables that describe functions and attributes of various genes and then use those to predict mouse phenotypes associated with specific knockouts. For example, from mouse knockout experiments, the researchers first flagged gene knockouts that increase mouse lifespan. Using the Harmonizome, Ma'ayan and his colleagues predicted the probability of genes, not yet knocked out in mice, for likelihood of increasing lifespan. "You can do this—predict other genes that should be relevant to aging—using machine learning," he says. "And those could be future drug targets for potentially increasing our lifespan or improving our healthspan."

Ma'ayan thinks of the Harmonizome as a prototype that is leading the way by showing what can be done. Some other data integration efforts allow search at the metadata level only. "The nice thing about the Harmonizome is that it enables search at the data level," he says. But, he acknowledges, making it scalable could be challenging.

Still, the Harmonizome has proven popular. During its first year, the site had 60,000 unique users visit and 250,000 page views. "We get about 400 users per day now," Ma'ayan says, with about 40 percent sticking around for a while because they are finding it useful. He'd like to learn more about how others are using the resource. "I'm sure people can think of creative ways to use it that we haven't thought of," Ma'ayan says. "That will be the coolest thing." □



By applying machine learning to data from diverse resources integrated to create the Harmonizome, Ma'ayan's team was able to predict likely knockout mouse phenotypes that have not yet been observed. Image Courtesy of Avi Ma'ayan.