RELEVANT NIH INSTITUTES:

NHGRI, NIAID, NIBIB, NLM, NIEHS, NIGMS, as well as all disease-focused Institutes including NCI, NHLBI, NIDDK, and NINDS

Data-Sharing Movement:

BD2K Centers Make Headway

BY KATHARINE MILLER

Science that isn't reproducible isn't science at all. And science that relies on big biomedical datasets will only be reliably reproducible if those datasets are **FAIR**—findable, accessible, interoperable and reusable.

Achieving the laudable goal of **FAIR** datasets requires a shift in scientific culture. Researchers accustomed to storing their data in silos at individual research institutions must become more mindful about how they handle, describe and store their data. In addition, there must be an infrastructure that makes data sharing possible.

When the National Institutes of Health (NIH) funded the twelve Big Data to Knowledge (BD2K) Centers of Excellence in October of 2014, **Philip Bourne, PhD**, the NIH associate director for data science at the time, understood that an emphasis on the **FAIR** standards within the BD2K Centers would seed this cultural change.

"We view this as a virtuous cycle," Bourne told this magazine. The Centers would generate **FAIR** data and data-sharing tools that others would use to do the same; and this ongoing cycle would serve to simplify and normalize the process. "Sharing the data and the software across the Centers and to other investigators and beyond is key," he said.

Fast forward two and a half years and the Centers are in full swing, propelling the data sharing revolution forward at every level of research and demonstrating that adherence to the **FAIR** principles is an achievable goal.

Metadata Entry by Humans: Achieving FAIRness Up Front

Biomedical researchers are generating datasets at unprecedented rates. To describe, store and share these datasets in ways that are **FAIR**, the researchers must create

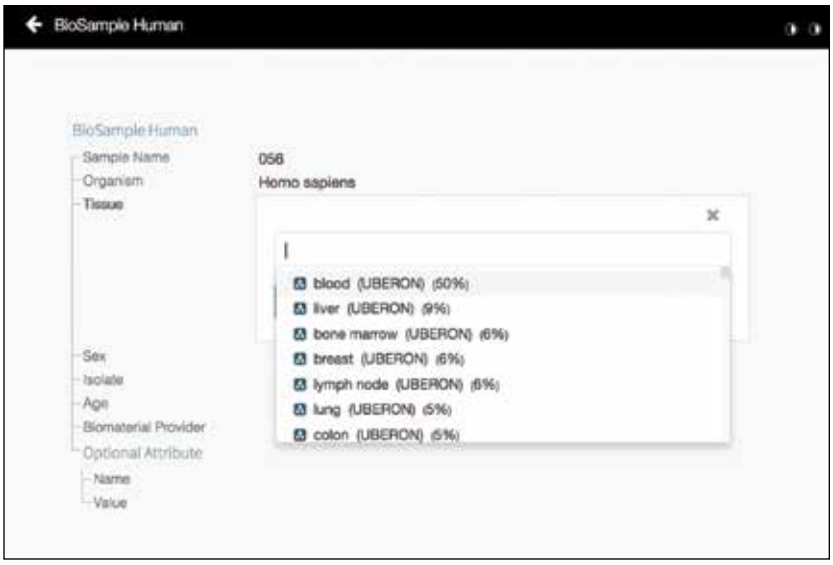
metadata—clear, accurate, computer-readable descriptions of the data. "A lot of metadata are produced by people who are forced to produce them under duress when they have other things they'd rather do," says **John Graybeal, PhD**, technical program manager for the BD2K Center known as **CEDAR (the Center for Expanded Data Annotation and Retrieval)**. "In the absence of good verification processes and helpful suggestions to such people about what information they need to provide, you get a lot of pretty useless metadata." And if the metadata are useless, the data itself will never be FAIR.

To address that problem, CEDAR has created the CEDAR Workbench, which researchers can use to access libraries of standard templates for defining metadata. The Workbench makes metadata entry easy by suggesting appropriate templates, enabling the use of appropriate terminology from various biomedical ontologies, and suggesting such terms in drop-down menus. In addition to making data findable and accessible, CEDAR's Workbench adds a lot of value for interoperability and reproducibility, Graybeal says, by ensuring that people are using the same terminology in consistent ways.

For example, investigators who use high-throughput B-Cell and T-Cell receptor repertoire sequencing (Rep-Seq) can deposit their data into four different repositories (BioProject, BioSample Sequence Read Archive [SRA] and GenBank) at the National Center for Biotechnology Information (NCBI). CEDAR is working with members of this community to provide a simple and standardized metadata entry process that can be integrated with the data submission process already in place. If these attempts are successful, researchers will be able to submit their data and metadata through the CEDAR Workbench, and it will flow to the appropriate NCBI repositories, considerably simplifying the submission process.

CEDAR is also collaborating with the Library of Network-Based Cellular Signatures (LINCS) consortium and the **BD2K-LINCS Data Coordination and Integration Center** (**BD2K-LINCS-DCIC**). LINCS researchers, who use various methods to disrupt biological pathways and observe the altered phenotypes, have generated huge amounts of data. Their standardized metadata procedure involves several manual steps: They enter information into an Excel spreadsheet, then submit it for manual review by a person who checks it for completion and accuracy and then emails the submitter to request corrections. "Heavily manual processes like this don't scale well," Graybeal says. "And Excel spreadsheets are limited in the support they can offer metadata providers." With a supplemental BD2K grant, CEDAR is helping LINCS researchers develop an integrated workflow for managing metadata in real time, with the system giving users feedback to correct mistakes right away. "From the user's standpoint and LINCS' standpoint, that's a big change," Graybeal notes. "Curators will be able to review created metadata much faster." The system is now functioning in a prototype environment and is targeted for production over the summer.

CEDAR is also contemplating how to fix flawed metadata that's already stashed in data repositories. Looking at the Gene Expression Omnibus (GEO) data repository, for example, researchers on the CEDAR team have noted inconsistent entries for such basic information as age and gender. CEDAR could support workflows to automatically enhance these metadata, or even provide simplified ways for users to correct these issues in a wiki-fied environment. The metadata updates could be forwarded back to the



In the CEDAR Workbench, a user selects a metadata template and then fills in the template by selecting values from various dropdown menus. Here the user is selecting a value for the tissue field in the BioSample Human template from candidate tissue types from the UBERON ontology in BioPortal. Courtesy of Mark Musen and CEDAR.

Variants of 'age' metadata field in Gene Expression Omnibus (GEO) repository

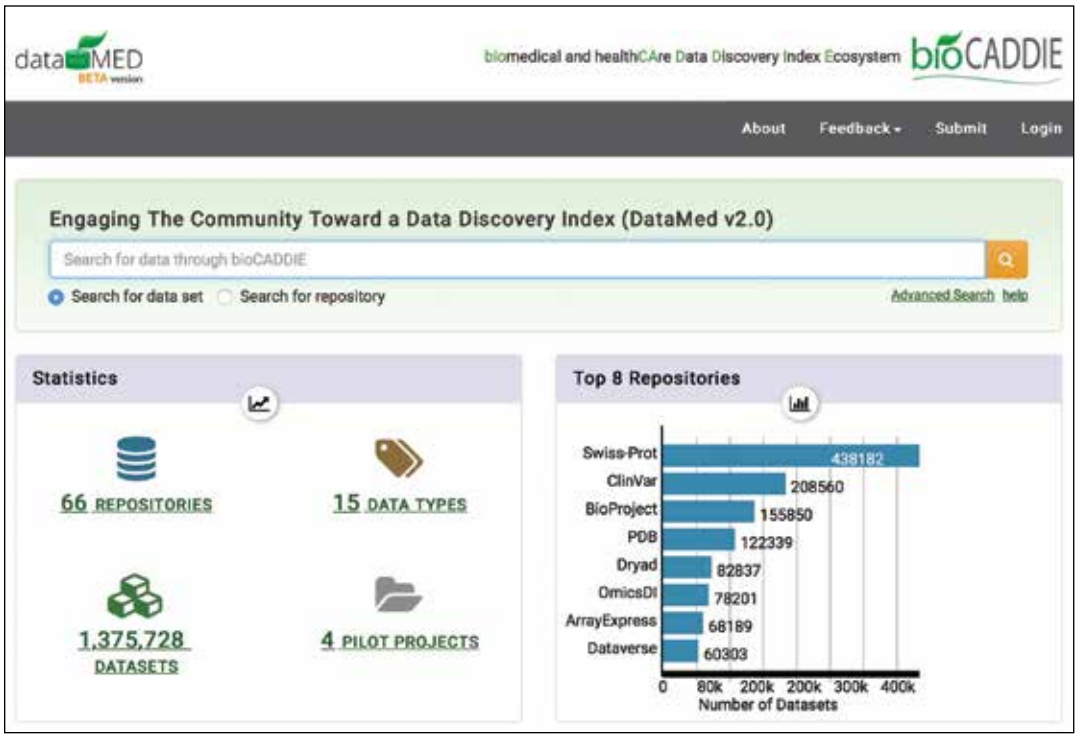
age	age [y]
Age	age [year]
AGE	age [years]
`Age	age in years
age (after birth)	age of patient
age (in years)	Age of patient
age (y)	age of subjects
age (year)	age(years)
age (years)	Age(years)
Age (years)	Age(yrs.)
Age (Years)	Age, year
age (yr)	age, years
age (yr-old)	age, yrs
age (yrs)	age.year
Age (yrs)	age_years

Metadata entry is difficult and leads to inconsistencies that make data reuse challenging. CEDAR is addressing this problem by creating standardized metadata templates. Information courtesy of Mark Musen, MD, PhD, principal investigator for CEDAR.

repository, if it had a way to handle these sorts of changes and suggestions. "Ideally, you'd end up with well-reviewed and more accurate documentation," Graybeal says, which would be a step forward for the FAIR principles.

Finding Accessible Data

Given the huge quantity of biomedical data that has been generated by high-throughput experiments as well as the vast troves of clinical data residing in electronic health records, many researchers hope to address interesting research questions by finding and accessing existing data rather than generating more. The NIH recognized the potential for re-use of existing datasets when, as part of the BD2K program, it funded the **biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE)** under the leadership of **Lucila Ohno-Machado, MD, PhD**, professor of medicine at the University of California, San Diego.



Searching for datasets in dataMED is akin to searching for scientific literature in PubMed.

bioCADDIE set out to develop a prototype data discovery index to help people find relevant datasets they would otherwise have a hard time finding. The result is dataMED, a repository and search engine that does for data what PubMed does for biomedical literature: Rather than having to search individual repositories for relevant data, researchers can search dataMED to find what they are looking for. As of March 2017, dataMED had indexed 64 data repositories containing more than 1.3 million datasets, and it is still growing.

Just as scientific journals must meet certain requirements for inclusion in PubMed, repositories included in dataMED must meet certain standards of quality, sustainability and interoperability. And their metadata must be capable of being digested into dataMED's DATS (DAta Tag Suite) metadata system. The core DATS metadata includes information about a dataset's accessibility through an application programming interface (API) as well as whether access requires approvals or security clearances.

To test the effectiveness of dataMED, bioCADDIE ran a dataset retrieval challenge competition to see who could develop an algorithm that would identify the best set of data to address a well-defined research question. The top two winning algorithms are now being incorporated into dataMED.

"Achieving findability and accessibility is just the beginning of the journey," Ohno-Machado says. More work will be required to achieve interoperability and reusability. For now, she says, "it's critical to find the data in the first place."

Making Dataset Recommendations: Findability Goes Deeper

In a separate effort to make datasets more findable, the **HeartBD2K Center** created the Omics Discovery Index (OmicsDI). Whereas dataMED indexes a broad array of datasets (including OmicsDI), OmicsDI focuses solely on omics datasets (proteomics, transcriptomics, genomics etc.). It is also searchable at a deeper level than dataMED.

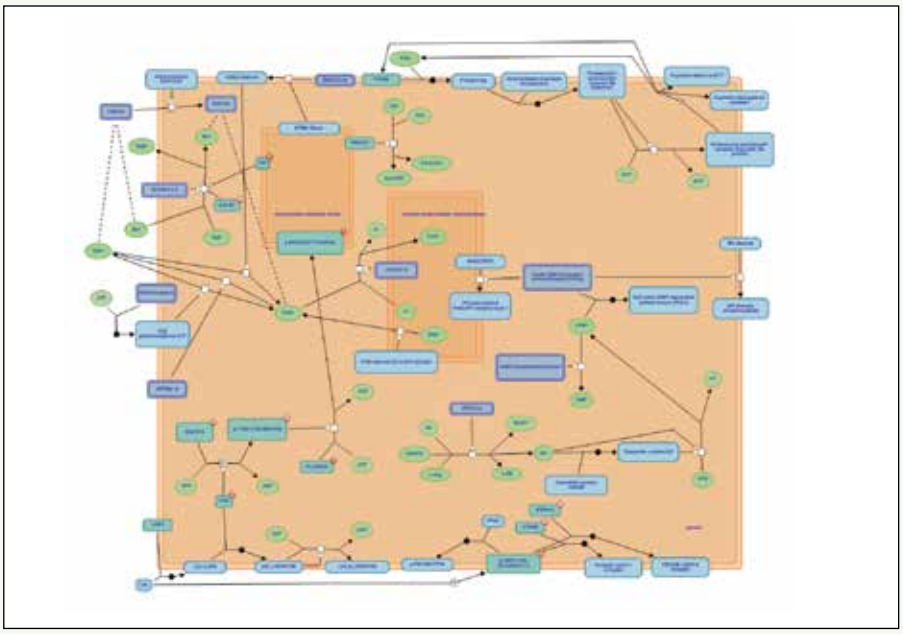
OmicsDI evolved from an even more specifically defined collaboration called the ProteomeXchange, a global network of four proteomics databases that coordinates how they accept data and then centralizes their metadata. The HeartBD2K Center has extended these concepts to multiple omics data types including genomics, transcriptomics, and metabolomics, says **Henning Hermjakob, MSc**, team leader for molecular networks services at the European Bioinformatics Institute (EBI) and co-director of HeartBD2K. "We got off the mark quite fast because we could build on existing experience and infrastructure."

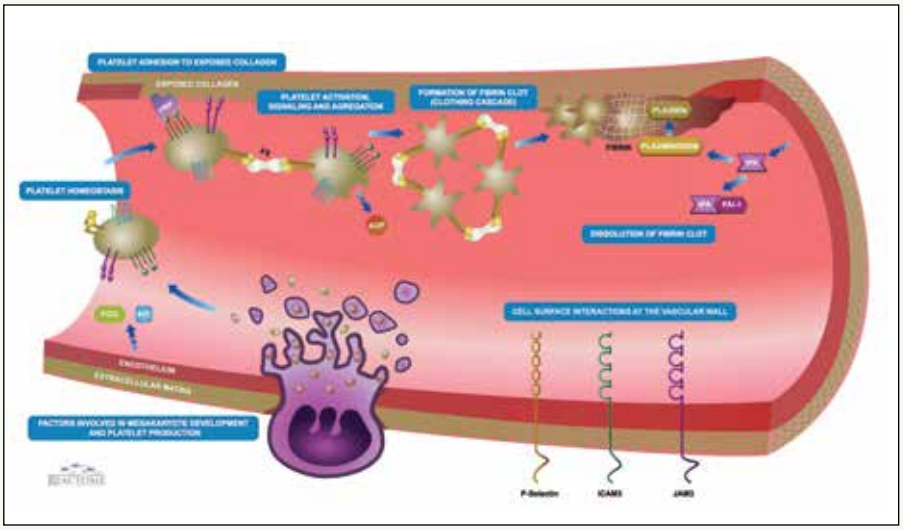
Like bioCADDIE's dataMED, OmicsDI can be easily searched to find datasets of interest. "But what we offer beyond pure metadata indexing is where it gets interesting," Hermjakob says. In addition to indexing the metadata, OmicsDI indexes part of the data content. For example, it might index the proteins observed in a proteomics

Extracting Needles from Haystacks: The Reactome

Reactome.org is an open-source curated database of human biological pathways. As of early 2017, it comprised 10,391 human reactions organized into 2,080 pathways involving 10,624 proteins encoded by 10,381 different human genes, and 1,735 small molecules. And under BD2K, the web interface has been re-designed to create a user-friendly tool. Researchers gathering gene expression data about normal and diseased tissue or cells perturbed with a certain drug want to know what the observed expression changes mean. Now they can upload their datasets to Reactome.org and it will show the pathways that are over- or under-represented in their specific gene set. "It's very helpful to the biologist in reducing large changes in large datasets to something that is much more understandable biologically," Hermjakob says. It doesn't necessarily give answers, he says, but it points researchers in the right direction. "It provides a magnet for extracting a needle from a haystack."

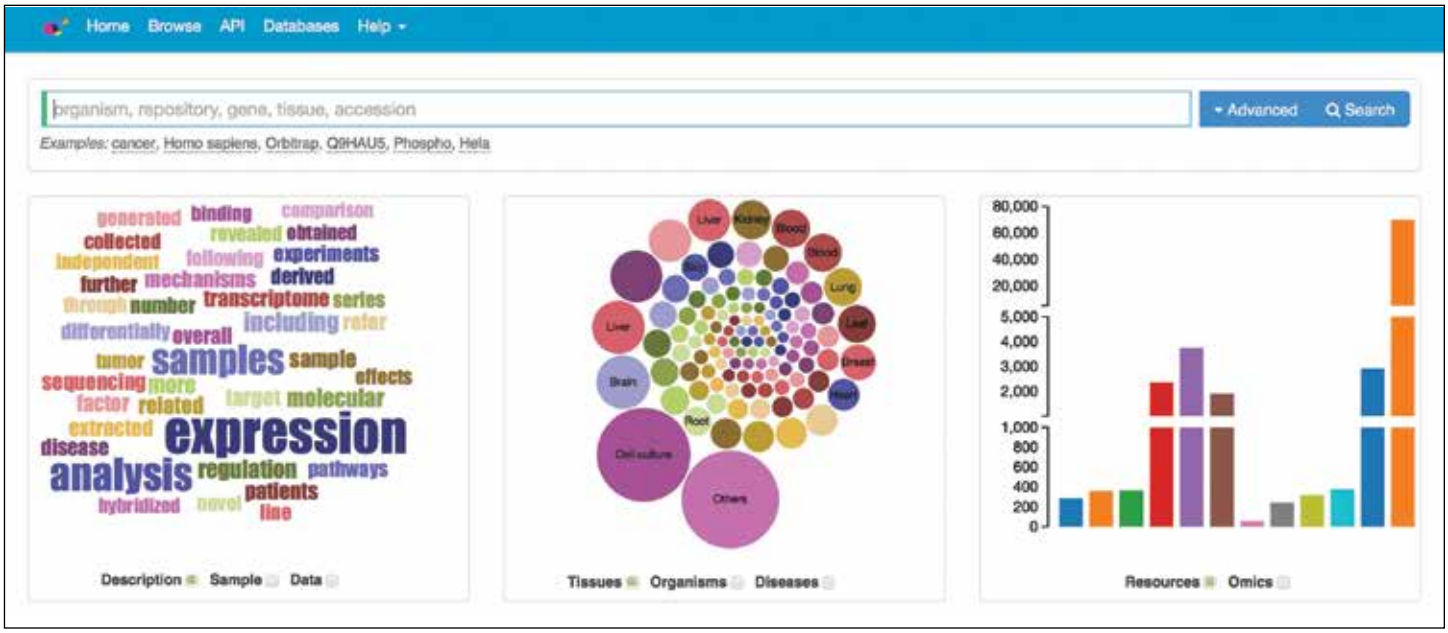
The updates to Reactome make it not only more accessible but also more interoperable. "We've developed computational components allowing Reactome functionality in other websites with very little effort," Hermjakob says. LINCS-DCIC, for example, provides high-quality new data on systematic perturbations of different cellular systems and Reactome aims to provide analysis capability for exactly these kinds of data output. "So the two fit together quite nicely," Hermjakob says. Reactome functionality has also been incorporated into several non-BD2K projects including the Human Protein Atlas and the Open Targets project.





In the very latest release of Reactome.org,Hermjakob and his colleagues have significantly enhanced the visual presentation of biological pathways. Schematic diagrams (such as the one at top showing platelet homeostasis) have been replaced with textbook style pathway diagrams (bottom, showing homeostasis more generally) drawn by a professional biomedical illustrator. Portions of the diagrams are clickable to dig deeper into the various pathways. "These make Reactome easier to navigate and the diagrams are also released in an editable form that can be used for publications and slides," Hermjakob says. Courtesy of Reactome.org.

experiment; or the differentially expressed genes in a transcriptomics experiment. OmicsDI then calculates similarity metrics between all the experiments in a domain. Using this capability, a user can get recommendations for datasets of interest. "The whole functionality is similar to recommendations in Amazon.com: 'If you're interested in this dataset you might also be interested in this other one,'" Hermjakob says. He's eager to see if this system leads to more datasets being reused—and OmicsDI is tracking that as well.



When searching OmicsDI for relevant datasets, the search box offers a dropdown menu of options. When the search is complete, researchers may further refine it by tissue, disease, or organism, and search results can be sorted by relevance—a measure of how closely related the datasets are to the specific query.

Findability at the Deepest (Data) Level

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** (**BD2K-LINCS-DCIC**), decided to take findability to another level. Their creation, the Harmonizome, offers 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 data silos but now is accessible."

To create the Harmonizome, Ma'ayan's team gathered together 66 major online omics resources and processed them into more than 70 million associations between nearly 300,000 attributes and all human and mouse genes and proteins. 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. The data are now served online through a user-friendly interface. "It makes it very easy for someone to do predictions of functions for genes," Ma'ayan says.

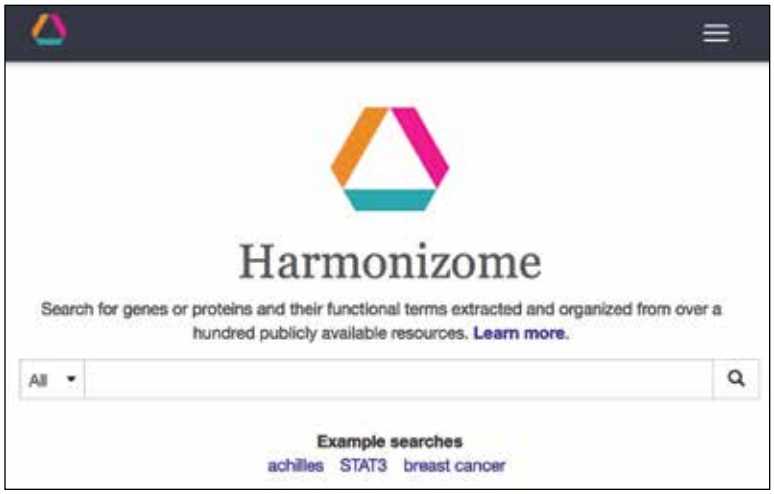
Ma'ayan thinks of the Harmonizome as a prototype that shows what can be done. "The nice thing about the Harmonizome is that it enables search at the data level," he says. He acknowledges that making it scalable could be challenging. Still, the Harmonizome has proven popular. Since it became public in 2015, the site has had more than 100,000 unique user visits and 300,000 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."

Accessing Data Where It Lives

For genomics research to achieve its promise of improving human health, vast data resources must be brought to bear. "We absolutely have to share the data so that we can compare cases," says **David Haussler, PhD**, principal investigator for the **Center for Big Data in Translational Genomics (BDTG)**, a BD2K Center. But access to genomic datasets is often restricted because of privacy concerns and confidentiality laws established by the countries and institutions where the data reside. This makes it impossible to create a central, unified genomic database. "The only path forward is to create a common API and a common language for containerized workflows so that you can literally ship analysis software off to different countries and medical institutions and let it run on their platforms," Haussler says.

And that has been one focus of the BDTG Center. They've created an ecosystem of genomics tools and a standard interface for interacting with genomic data—the Global Alliance API. It allows genomics researchers to work together on a global scale, sharing software to achieve reproducible results.

As a test case, the Center developed Toil, a portable, open-source workflow, and demonstrated its use in a cloud environment by reprocessing more than 20,000 RNA-seq samples from four major studies. The effort reduced the time and cost of such processing by 30-fold; created a new, publicly available dataset free of batch effects (the statistical problems created by data processed in different ways at different research institutions); and set a precedent for the use of portable software in contemporary cloud workflows as a path toward allowing research groups to reuse data and reproduce one another's results.



Sharing and Integrating Clinical Data

Often, researchers want to study different types of data for the same patient population but the data—much of it privacy protected—are distributed across multiple databases and institutions. In a pilot study called Count Everything, four BD2K Centers of Excellence worked together to create a prototype for integrating various types of data without compromising privacy. Ohno-Machado's Center, bioCADDIE, played the aggregator role using APIs (for genomics, electronic health records, and mobile health data) developed by hree other centers: BDTG, **PIC-SURE** (**Patient-centered Information Commons: Standardized Unification of Research Elements**), **MD2K** (**Mobile Sensor Data-to-Knowledge**). The result: a system that can make simple distributed queries across simulated data from these Centers in a secure and anonymized way. Queries could ask, for example,

To create the Harmonizome, researchers with LINCS-DCIC distilled information from original datasets into attribute tables that define significant associations between genes and attributes, where attributes could be genes, proteins, cell lines, tissues, experimental perturbations, diseases, phenotypes, or drugs, depending on the dataset. These attribute tables can be searched and integrated to perform many types of computational analyses for knowledge discovery and hypothesis generation.

Integrated Transcriptomics: Clue.io

In addition to LINCS-DCIC, the NIH funds six LINCS (Library of Integrated Network-Based Cellular Systems) Data and Signature Generating Centers. They all gather data on perturbed cells, but each center has a different focus (such as transcriptomics, proteomics, the cellular microenvironment, disease, drug toxicity, or the brain).

The LINCS Transcriptomics Center, located at the Broad Institute in Cambridge, Massachusetts, also receives BD2K funding. The aim: to integrate LINCS-generated transcriptomics data (approximately 2 million gene expression profiles) with all the other transcriptomics data in the world (for example, the approximately 1 million profiles that reside in the Gene Expression Omnibus). "That was the premise of our proposal—to create a unified system across all these transcriptomic sources," says **Aravind Subramanian, PhD**, principal investigator for the **Broad Institute LINCS Center for Transcriptomics and Toxicology**.

The project created clue.io, a website with an API that provides a uniform programmatic interface to all the transcriptomic datasets. Clue.io also offers web applications that can be used to find relationships between genes, compounds and diseases. "BD2K funding allowed us to build all these tools," Subramanian says. "We're hoping that these APIs we are creating to expose the data will be taken up by other BD2K centers and integrated with the other datasets and methodologies they've been developing."

the number of individuals in these datasets who share a clinical phenotype, genomic variant and activity profile. And they could achieve this interoperability without any Center seeing another Center's data. According to **Benedict Paten, PhD**, associated research scientist at BDTG, "This is an example of what can happen when big centers with expertise in different areas coordinate and come together."

Interoperability of Genomics Datasets and Tools

In the good old days, a researcher would upload data to a web server where stand-alone tools would do the analysis. Today, large datasets often reside in specialized cloud environments. Tools must be brought to the data, rather than the other way around.

Researchers at the BD2K Center **KnowEnG**, a collaboration between the University of Illinois at Urbana-Champaign and the Mayo Clinic, realized that the analytical tools they develop have to be more than just accessible and downloadable. "Our tools have to be able to talk to other data repositories and other code bases and analysis systems for the user to have a reasonable experience in their analysis pipeline," says **Saurabh Sinha, PhD**, professor of computer science at the University of Illinois at Urbana-Champaign.

For example, Sinha says, The Cancer Genome Atlas (TCGA) is a big data repository that is part of the Stanford Genomics Cloud. "If you want to analyze those data using our kinds of tools, there needs to be a convenient and formal mechanism for these different systems to talk to each other, rather than the researcher making it happen by brute force." The solutions KnowEnG researchers are developing should be available within a year. "We're working on a way by which researchers can invoke our tools from their cloud and analyze the TCGA data right away on that cloud," he says.

Ideally, Sinha says, many such interactions will be possible in the future. "Researchers will be able to say 'get me this slice of the LINCS data, and analyze it with this pipeline in KnowEnG,'" he says. "If this kind of goal can be achieved, it will be fantastic."

Accessing Data for Reuse

Some large health datasets created over many years remain locked in formats that aren't truly accessible. For example, in the 1960s the Centers for Disease Control (CDC) began been conducting surveys and interviews to better understand the health and nutrition status of the American people. Since 1999, this survey, called the National Health and Nutrition Examination Survey (NHANES) has been continuous, covering about 5,000 people each year. The data gathered covers a broad range of topics and, until recently, was stored in about 250 Excel spreadsheets at the CDC. Anyone hoping to analyze the data could do so only in these discrete subsets.

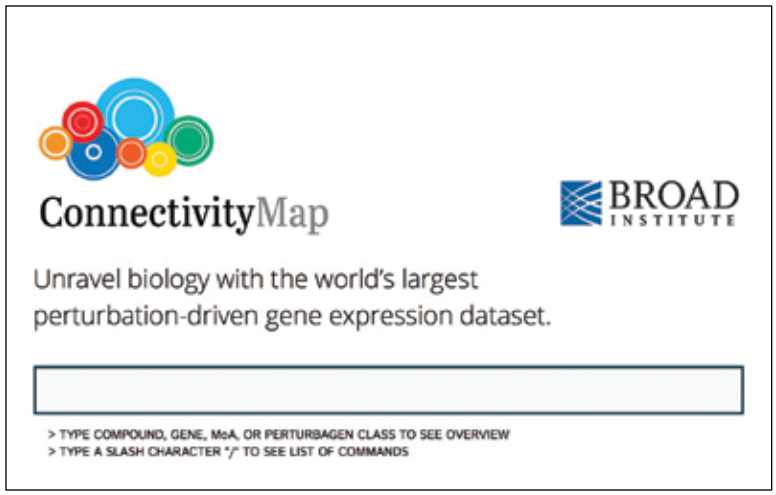
Researchers at the **PIC-SURE** BD2K Center set out to correct this problem and establish a prototype user-interface that would simplify analysis of NHANES data. First, they integrated the NHANES data, combining thousands of different variables—clinical, environmental, self-reported, and genomic—into one set of data structures. They then loaded the data into a software system called i2b2/tranSMART, which makes the data accessible to researchers in a web

Real World Data Sharing

Genetic testing sometimes reveals that patients have a heightened risk of breast cancer. Certain variants in the BRCA1 and BRCA2 genes, for example, raise the lifetime risk from about 12 percent to as much as 65 percent. But if a woman has a different variant of these genes (and there are literally thousands of possible variants), doctors don't necessarily know their significance. As a result, clinical tests often reveal variants of uncertain significance, as much as 20 percent of the time, and occasionally different clinical testing companies give women different information about their breast cancer risk. To reduce that uncertainty and inconsistency, BDTG, ENIGMA and many collaborators have created the BRCA Exchange, a public resource that aggregates and unifies data on BRCA variants. The Exchange is now the world's largest aggregation of genetic variance, Paten says.

To curate information for the BRCA Exchange, experts review information about BRCA1 and BRCA2 variants to develop consensus about what they signify. "We started with 1,000 expert curations and now have 3,500," Paten says. Ultimately, he says, all the variations on the site will be curated. The various clinical testing companies will be able to cross-reference the BRCA Exchange and deliver better information to patients. "It's a real-world project that is trying to deal with data sharing right now," Paten says. He also sees it as a model for other diseases for which there are genes of interest and communities of people who want to understand their potential risks.

browser as well as adding a layer of analytical tools. In the user interface, researchers can easily drag and drop different patient characteristics into boxes for comparison using statistical tools that are part of i2b2/tranSMART. Furthermore, to



Clue.io is a website with an API that provides a uniform programmatic interface to all the transcriptomic datasets.

allow large-scale analysis across the entire dataset, the BD2K PIC-SURE team implemented a RESTful API called the PIC-SURE RESTful API "This is a programmatic way of accessing data for large scale computing," says **Paul Avillach, MD, PhD**, assistant professor of bioinformatics at Harvard Medical School and member of the PIC-SURE team.

In addition to the NHANES dataset, the API can be used to analyze i2b2 patient electronic health records (with different levels of privacy access for different types of users), data from the Exome Aggregation Consortium (ExAC) browser at the Broad Institute, or any other dataset a researcher would like to import into the system.

Reusability and Reproducibility Using Jupyter

Often, computational research is done on a postdoc's laptop. Eventually that person moves on to a different lab or project and leaves an insufficient record of the steps taken or even the location of the computer script. The result: the work is not reproducible. But using Jupyter Notebooks, an open-source web application, researchers can detail all the steps of a computational project from input to output using any combination of 40 different computer languages. When researchers publish a paper, the Notebooks can be published alongside the data. "Jupyter Notebooks are a very nice way of doing reproducible science for real," Avillach says. "They allow you to share how you managed to process the data so someone else can reproduce the exact same results."

At the PIC-SURE Center, Avillach and his colleagues established a system for using Jupyter Notebooks to track their work. They also added Docker technology (that they then contributed back to the Jupyter open-source hub) to create a protected research environment for each researcher. "We have one Docker container per investigator," Avillach explains, "so no one can crash the container of another investigator." It's a prototype for computational research that Avillach hopes becomes the norm.

Piloting the Commons Cloud Credits Model

In October 2014, Bourne announced plans to create the "NIH Commons" to catalyze the sharing, use, reuse, interoperability and discoverability of shared digital research objects, including data and software. The Commons is portrayed as a layered system consisting of three primary tiers: high-performance and cloud computing (at the bottom); data, including both reference datasets and user-defined data (in the middle); and (at the top) services and tools, including APIs, containers, and indexes, as well as scientific analysis tools and workflows and—eventually—an app store and interface designed for users who are not bioinformaticians.

To be eligible for use in the Commons, data and software must meet the FAIR principles. For example, the products of all the BD2K centers will be part of the Commons ecosystem, including dataMED from bioCADDIE and the CEDAR Workbench. And to incentivize participation in the Commons, the NIH is piloting a plan to offer cloud computing credit vouchers that researchers can use with a provider of their choice, so long as the provider complies with the FAIR principles. Several BD2K Centers are participating in the pilot, including KnowEnG, PIC-SURE, and BDTG.

As part of their pilot, PIC-SURE took a HIPAA-compliant research environment (for using privacy-protected patient data) that they had developed for use in the Amazon cloud and added Docker containers to make it cloud-vendor agnostic. "We realized that we didn't want to be limited to one cloud vendor," Avillach says. It is now useable across multiple cloud vendors including Amazon, Google and IBM cloud layers.

Haussler strongly supports the Commons effort. "It's very important that we make it easy for NIH researchers to access data and compute on the cloud and in so doing share data," he says. From a financial point of view, having NIH principal investigators each build their own computer facilities for these big data comparisons will end up costing billions more than if researchers can work together in a common computing environment with competitive pricing, he says. But the important thing is the science. "You can't make progress unless you can share data," he says. "The technology exists to do it. It's just the will and the organization. I think we're at a critical point. We're very enthusiastic about continuing to work on it." □