Open-Source Science Makes Headway

By Margie Patlak

Open-source science is gaining ground as scientists explore the benefits of working together without copyright and patent constraints in virtual forums. Like Wikipedia, these "wiki-style" forums enable them to reuse, build on, and extend the work and resources of fellow researchers. In the past 5–10 years, more than a dozen websites have emerged that are dedicated to creating open access to many of the necessary ingredients for drug discovery, including data, information technology and analytic tools, biospecimens, and disease models (*see* sidebar).

Many advocates for open-source science claim that it can counter investigative redundancy and the squandering of biomedical research resources. They also say it will create a productive synergy that will speed up our understanding and treatment of complex diseases such as cancer.

"The availability of clinical data. . .allows us to concentrate on validating interesting features we have found in our analysis, rather than on generating data."

But both advocates and critics point out challenges to this approach, such as countering the cultural and economic norms that limit sharing, creating compatible datasets, ensuring high-quality data, and finding sustained financial backing.

Despite those challenges, open-source science is making substantial headway in the biomedical arena. The amount and quality of information available in the public domain has grown dramatically, according to Thinh Nguyen, counsel for Science Commons, a nonprofit organization whose mission is to make sharing scientific data and materials among scientists easier. "These changes mean that what is available in the commons is starting to be almost as good as what companies can develop themselves internally—it's starting to get a place at the table when you are doing serious drug development research," Nguyen said at a presentation at an Institute of Medicine (IOM) conference on precompetitive collaboration last February.

The participation of big-name players—such as the National Cancer Institute, Merck, Pfizer, Eli Lilly, several universities and foundations, and patient advocacy groups—is boosting the reputation of open-source science. So are some websites' impressive gains. The Open-Source Drug Discovery (OSDD), which is dedicated to discovering treatments for diseases that plague the developing world, surprised many when more than 400 of its volunteer researchers reannotated the tuberculosis bacterium genome, wiki style, in just 4 months—record time for such an endeavor, according to Eli Lilly's Bernard Munos.

The Cancer Genome Atlas

One of the most stunning examples of open-source science is the Cancer Genome Atlas (TCGA) project, which the National Institutes of Health funds. Started in 2005, it aims to make public the highly characterized genomes of 20 tumor types—500 cases each—and matched normal tissues.

A major aim of TCGA is to characterize the DNA and RNA extracted from tumor samples. This endeavor will yield not only DNA sequences for the tumors but also abundant gene expression, copy number, DNA methylation, and microRNA data that can be linked to clinical information. TCGA's Web-based portal currently includes extensive datasets for ovarian cancer and glioblastomas, as well as fewer data on lung cancers. The data have led to more than 20 scientific publications. Nearly three-quarters of these were by researchers outside the project network who accessed TCGA data for their own work.

Ethan Cerami, a bioinformatics engineer and computational biologist at Memorial Sloan–Kettering Center in New York, is using the databases to study glioblastomas. Cerami and his colleagues developed an algorithm to determine whether the genetic alterations in glioblastomas clustered within specific pathway networks. When they used that algorithm to automatically process TCGA glioblastoma data, they uncovered previously undetected pathway networks and genes that appear to drive the cancer.

Peter Park, Ph.D.

"The scope of the data being generated has spurred us to rethink everything from basic data storage to pathway and network analysis of the data," Cerami wrote in an e-mail.

Another cancer researcher, Peter Park, Ph.D., from Harvard Medical School in Boston, has several TCGA data–based studies published or in the works. He and his colleagues analyzed the glioblastoma data and found a microRNA that, by inhibiting expression of certain key

genes, fosters tumor aggressiveness and is tied to decreased survival. "The availability of clinical data has been especially helpful as they allow us to concentrate on validating interesting features we have found in our analysis rather than on generating data," Park said in a phone interview.

Potential Obstacles

One obstacle to open-source science forums is that they depend on researchers' willingness to share their data and other resources, when obtaining grants, and career advancements often depends on individual recognition for work. For this reason, some open-access forums, such as OSDD, give attribution to all contributors on its website.

Another obstacle is that researchers' datasets are often in different formats. For example, the word *gene* can mean two different things in two different databases. Developing standards and infrastructure to deal with inconsistent or uncomparable data may be a costly but necessary endeavor. One suggestion at the IOM workshop, from Richard Bookman, Ph.D., of the University of Miami, was that the IOM devise a set of standards on the sharing of data, materials, tools, and collaboration. Federal, state, and other funding agencies could then use the standards as guidance when shaping grant programs.

Data Quality

Some researchers are also concerned about the quality of the data and biospecimens in open-source science. Bioinformatician Keith Baggerly, Ph.D., of the University of Texas M. D. Anderson Cancer Center in Houston, has used the data on TCGA and other government-sponsored websites. "I don't trust the data to be perfect—sometimes there are odd findings," he said in a phone interview. "In most cases the data are all correctly labeled, but sometimes they are not, and there are often no checks to ensure that for people using the data."

Those checks should include a posting of the raw data, the processed data, and the steps taken between the two, he added. "Open-source science has the potential to be of high quality as much as anything else out there as long as there is documentation for the results posted—how people got to their conclusions," Baggerly said.

In his experience with mislabeled data on TCGA's website, finding the right person to notify was easy, and the errors were quickly corrected, Baggerly said. But for nongovernment open-source science sites, how to report errors and whether those errors are corrected once they are reported is not always clear, he added. And some sites may not make all their data completely open to public scrutiny.

Computational biologist John Quackenbush, Ph.D., of the Dana–Farber Cancer Institute in Boston, criticizes OSDD for not being broadly open to the public. "There are places for crowd-sourcing approaches such as wikis. The challenge is to build open-source science sites intelligently, with enough checks and balances to make sure that the gene assignments or other conclusions are correct—the systems and processes have to be well engineered and curated if they are going to ultimately be successful," Quackenbush wrote in an e-mail.

Munos argues that the researchers generating the data provide the best check on data quality, because they constantly review each other's contributions. The hope is that "the glare of scrutiny will . . . keep people honest," Munos said.

Others say open access may actually solve quality issues in scientific data. According to Nguyen, cell line contamination—a

Selected Open-Source Science Websites

Biomarkers Consortium

Qualifies biomarkers and makes its results available in the public domain.
http://www.biomarkersconsortium.org/index.php?option=com_content&task=section&id=5&Itemid=39

Cancer Genome Atlas

Aims to provide for the public domain the highly characterized genomes of 20 tumor types—500 cases each—and matched normal tissue to facilitate the future discovery of pharmaceutical and diagnostic targets in cancer.
http://cancergenome.nih.gov

Oncomine

Cancer microarray database and Web-based data-mining platform aimed at facilitating discovery from genomewide expression analyses.
https://www.oncomine.org/resource/login.html

Open-Source Drug Discovery (OSDD)

Provides a global virtual platform where researchers can collaborate and collectively discover drug therapies that cause major health care problems in the developing world.
http://www.osdd.net

Pathway Commons

Offers a point of access to biological pathway information collected from public pathway databases.
http://www.pathway/commons.org

Personal Genome Project

Aims to provide a public database of the genomes and phenotypes of 100,000 people.
http://www.personalgenomes.org

Sage Bionetworks

Provides an Internet-based commons for biomedical data, as well as integrative models for human diseases.
http://sagebase.org

Science Commons

Provides policy guidelines and legal agreements to make research data and resources, such as biospecimens, cell lines, and model animals, easier to find and share. The Science Commons also recently began a Neurocommons project, which will create an open-source blending of neuroscience databases and information portals.
http://sciencecommons.org

World Community Grid

A global public computing grid that hosts Help Conquer Cancer, whose mission is to automate the processing of X-ray crystallography images of proteins thought to play a role in human cancers.
http://www.worldcommunitygrid.org/research/hcc1/overview.do
http://www.worldcommunitygrid.org

perennial problem in biomedical research—is more likely to be detected if researchers have open access to the materials and can test them for contamination. "The availability of these materials to the scientific community is crucial in order to validate results and detect potential problems with past studies," Nguyen said.

"Open source science has the potential to be of high quality. . .as long as there is documentation for the results posted. . ."

Sustainable Funding

Many public science forums are public–private partnerships, supported by a combination of funding from the government, pharmaceutical firms, charitable institutions, and patient advocacy organizations. For example, the Signaling Gateway, a website for information on cell signaling proteins, is supported by the National Institute of General Medical Sciences (part of NIH), as well as Genentech and the Nature Publishing Group. The OSDD currently receives most of its funding from the government of India but is seeking more funding from international agencies and philanthropic organizations.

But many resources that support open-source science are short-lived, according to Nguyen, and without some steady source of funding for the effort it takes to maintain these forums, the commons will fail.

"So what happens is you do the research and then you throw away the data. That is a wasteful way for us to use limited funding dollars and not leverage the potential for this stuff to be in the commons," Nguyen said.

And sharing can be expensive: Integrating data and maintaining and shipping biospecimens all involve substantial costs, let alone the time costs for people to create and support open-source science websites, several participants at the IOM conference pointed out. The FDA, NCI, and other government agencies, along with industry and academic researchers, spend many hours participating in meetings and carrying out other activities that further public domain efforts, yet no specific funds are earmarked to support such activities.

At the IOM workshop, Ray Woosley, M.D., Ph.D., of the Critical Path Institute, a nonprofit that creates collaborations between FDA scientists and the medical product industry, noted that collaborations often are never hatched or fail because of insufficient funding. "People have tried to share and pool placebo data for many years, and have actually gotten the data from companies, but they have not been able to get the funding to actually use it," he said.


Keith Baggerly, Ph.D.

Despite those obstacles, the open-source science movement is something that many are watching, especially those who think this model can spark innovation and speed drug development. But both advocates and critics say that it is not likely to succeed unless it is accompanied by changes in corporate culture and the behavior of other stakeholders.

"Some of the norms in science of how you share data have to change," Nguyen said.

 DOI: 10.1093/jnci/djq321