Biology's Dry Future

The explosion of publicly available databases housing sequences, structures, and images allows life scientists to make fundamental discoveries without ever getting their hands "wet" at the lab bench

Most life scientists single-mindedly focus their careers on a particular organism or disease—even just a specific molecular pathway. After all, it can often take months of training to master growing a particular cell type or learn a new laboratory technique. Atul Butte, however, wanders from topic to topic—and reaps scientific successes along the way. Though only 44 years old, he has earned tenure at Stanford University's School of Medicine in Palo Alto, California, based on advances in diabetes, obesity, transplant rejection, and the discovery of new drugs for lung cancer and other diseases.

Butte's lab is different, too. It isn't crowded with cell cultures and reagents. His tools look like those of an engineer or software developer: Most often, he's simply working on a Sony laptop, although at times he does turn to a large computer cluster at Stanford and supercomputers elsewhere when in need of massive processing power. Instead of growing cells and sequencing DNA, Butte, his students, and postdocs sift through massive databases full of freely available information, such as human genome sequences, cancer genome readouts, brain imaging scans, and biomarkers for specific diseases such as diabetes and Alzheimer's.

Many call this type of research "dry lab biology," to contrast it with the more hands-on "wet" traditional style of research. Although statistics on the number of dry lab biologists are hard to come by, these data hunters believe they are a growing minority. Butte is one of its top practitioners. Using publicly available data, for example, 2 years ago Butte and his colleagues surveyed the activity of large sets of genes in people affected by 100 different diseases and in cultured human cells exposed to 164 drugs already on the market. By comparing patterns of genes flipped on or off by the diseases and by the drugs, the team drew unexpected connections. They found clues

CREDIT: ANDREW J LENARDS

Published by AAAS

New miners. New database construction and analysis tools from the iPlant Collaborative (*left*) allow digging through plant and microbial genomes, helping plant biologists around the world improve their understanding of basic biology and advance crop breeding.

CREDIT: ORBERT VON DER GROEBEN/STANFORD SCHOOL OF MEDICINE

that a drug now prescribed for ulcers might also be a useful lung cancer treatment, for example, and that an antiepileptic compound would fight two forms of inflammatory bowel disease (see chart, p. 188). Subsequent lab studies of animals offered support for both inferences. And last month, Butte's group reported in *Cancer Discovery* that a similar approach suggested that the antidepressant drug imipramine would be effective against small-cell lung cancers resistant to standard chemotherapy—a finding that has already prompted the launch of a clinical trial. "This is an exciting time to be doing biological research on a dry bench," Butte says.

And not just for Butte. The growth of publicly accessible data troves on genome sequences, gene activity, and protein structures and interactions has opened new territory for biologists. Seizing on advances in computational power, data storage, and software algorithms able to separate the wheat from the chaff, dry lab researchers are making fundamental discoveries without ever filling a pipette, staining a cell, or dissecting an animal. Thanks to a National Science Foundation–funded initiative called the iPlant Collaborative, for example, there's an emerging generation of data-analyzing "plant biologists" who have never gotten their hands dirty digging in soil or watering seeds. And the National Institutes of Health (NIH) recently announced plans to sink $96 million into boosting analysis of big data. "There is a transformation happening in biology," says Daniel Geschwind, a neurogeneticist at the University of California, Los Angeles.

"You basically don't need a wet lab to explore biology," agrees David Heckerman, a computational scientist at Microsoft Research in Los Angeles. None of these dry lab biologists believe that advances in data sciences will replace the traditional approach. Rather, they argue that the two dovetail with one another like never before, each propelling the other forward. "I'm like a kid in a candy store," Butte says. "There is so much we can do."

Data for all

Big data is certainly nothing new to science. (*Science* had a special package on the topic in the 11 February 2011 issue.) The Large Hadron Collider at CERN generates 15 petabytes (10^{15}) of data every year it's in operation. Astronomy's Sloan Digital Sky Survey contributes terabytes (10^{12}) yearly as well. Big data isn't even all that new to biology. As of the end of August, for example, NIH's 31-year-old gene sequence database, GenBank, held some 167 million gene sequences containing more than 154 billion nucleotide bases.

Nor is the marriage of computational science and biology novel on its own. Researchers have amassed large-scale basic biology data sets for years—unimaginatively dubbed genomics, proteomics, metabolomics, and so on—and combed them in search of novel insights into complex biological pathways and disease.

"I'm like a **kid in a candy store.** There is so much we can do."

—Atul Butte, Stanford University School of Medicine

But many of these early efforts were run by large consortia of researchers, who often had rights to first mine the data before releasing them to the public. So much of that information is now public, however, that it's opened the door for researchers who never participated in those consortia. "Now it's possible to ask big-data questions with data that is extant in the public domain," says Ed Buckler, a research geneticist who specializes in maize genetics at the U.S. Department of Agriculture's Agricultural Research Service in Ithaca, New York, and Cornell University.

Asking those questions requires specialized algorithms and software, capable of handling massive data sets, and those are improving even as the data proliferate. Heckerman and his Microsoft Research colleagues, for example, made a splash recently with a software advance that eases large-scale searches within genetic databases, such as those used to compare entire genomes in what are known as genome-wide association studies (GWAS). These efforts examine DNA of large numbers of ill people and healthy controls, looking for genetic fingerprints linked to disease. Those fingerprints can be subtle, because most diseases are unlike the simple traits of classical genetics—the colors of Mendel's peas, for example—in which each trait maps to a single gene. "When people first started doing GWAS they thought this would be really easy," Heckerman says. "The problem is that Mendel's peas are the exception not the rule."

Instead, the genetics behind most traits and diseases, such as diabetes and prostate cancer, is far more complex, with small contributions from many genetic changes having an additive effect. "To uncover these weak signals you need tons of data. You need tens of thousands or hundreds of thousands of people," Heckerman says. "But there is a catch. When you analyze lots of data, there is hidden structure," in which separate individuals share a multitude of genetic similarities. But in many cases, these similarities are due to two individuals being more closely related than others, instead of sharing common disease genes. "That wreaks havoc with data. You get tons of what looks like signals. But when you look closer it evaporates."

One way around this has been to use a data analysis approach called a linear mixed model. The approach's mathematical rigor helps reduce false positives, but the computing power needed for it grows as a cube of the number of subjects being analyzed. That's no problem when analyzing a few dozen people or so, but if you want to comb through tens of thousands of genome samples, "forget about it," Heckerman says.

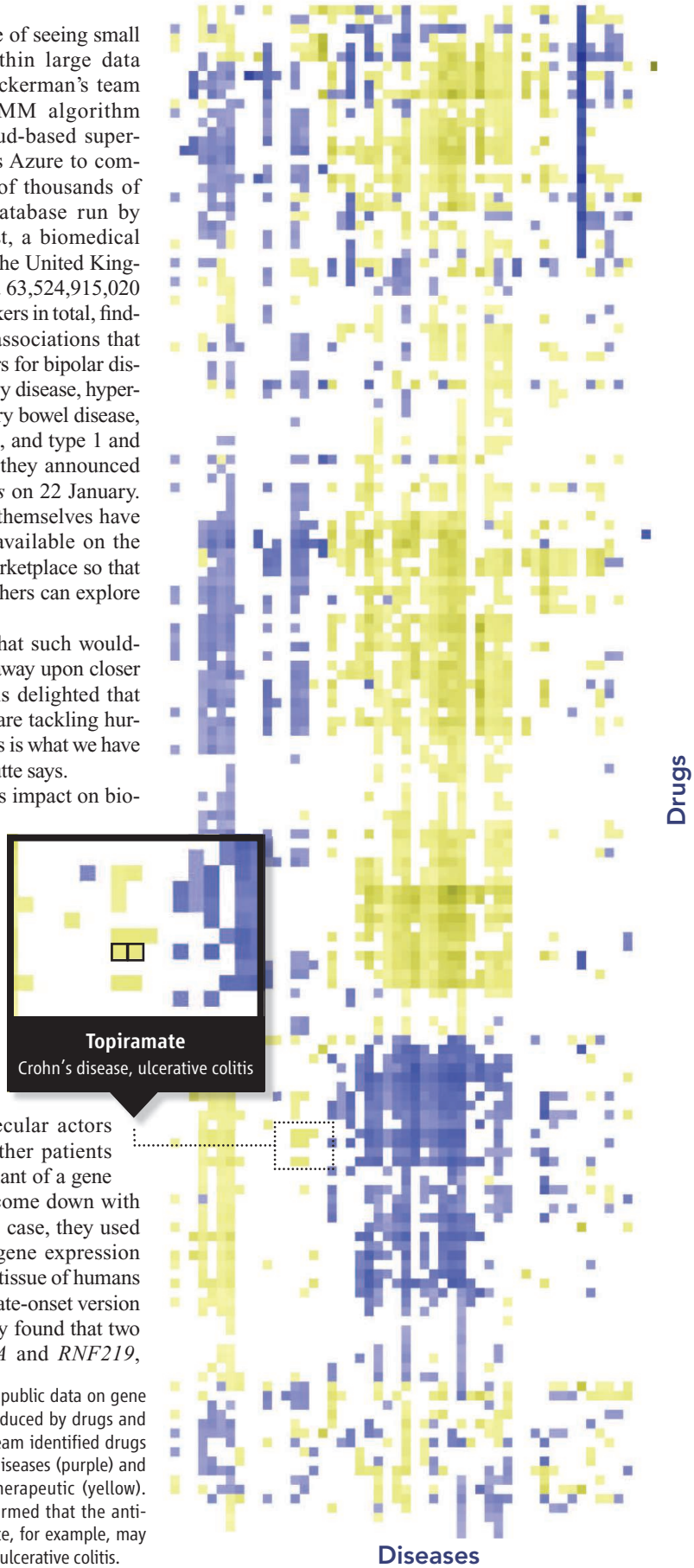
After grappling with the problem for some time, Heckerman and his colleagues came up with what he calls simple "algebraic tricks" to convert the problem to one that scales linearly with the number of subjects, making it tractable to crunch large data sets. The result, an algorithm dubbed FaST-LMM, reduces confounding results, increases the size of the samples that can be processed, and thereby

Published by AAAS

increases the chance of seeing small signals hidden within large data sets. Last year, Heckerman's team used this FaST-LMM algorithm on Microsoft's cloud-based supercomputer known as Azure to compare the genomes of thousands of individuals in a database run by the Wellcome Trust, a biomedical research charity in the United Kingdom. They analyzed 63,524,915,020 pairs of genetic markers in total, finding a host of new associations that may serve as markers for bipolar disorder, coronary artery disease, hypertension, inflammatory bowel disease, rheumatoid arthritis, and type 1 and type 2 diabetes, as they announced in *Scientific Reports* on 22 January. These associations themselves have been made freely available on the Windows Azure Marketplace so that independent researchers can explore them further.

Butte cautions that such would-be links often fade away upon closer inspection, but he is delighted that software engineers are tackling hurdles in biology. "This is what we have been hoping for," Butte says.

Dry lab biology's impact on biomedicine extends well beyond GWAS studies. Researchers led by Asa Abeliovich at Columbia University, for example, reported in *Nature* on 1 August that they used a big-data approach to discover new molecular actors that influence whether patients with a common variant of a gene known as *APOE4* come down with Alzheimer's. In this case, they used publicly available gene expression data sets from brain tissue of humans with and without a late-onset version of Alzheimer's. They found that two genes, called *SV2A* and *RNF219*, have abnormally low activity in people who develop the disease.



Drug hit. By analyzing public data on gene expression patterns produced by drugs and diseases, Atul Butte's team identified drugs that might exacerbate diseases (purple) and those that might be therapeutic (yellow). Follow-up studies confirmed that the anti-epilepsy drug topiramate, for example, may treat Crohn's disease or ulcerative colitis.

Combined with other clues to the genes' functions, the finding suggests that they act as previously undiscovered players in the molecular network that regulates intracellular accumulation of amyloid precursor protein. Amyloid collects in dense plaques in patients' brains and may play a causal role in the disease. Abeliovich's team confirmed the result in lab studies of mice, and then moved on to people—still in a dry lab. The team analyzed publicly available neuroimaging data of Alzheimer's patients and showed that variations in *RNF219* are correlated with the amount of amyloid that accumulates in their brains.

The work not only raises hopes of new drug targets for fighting dementia, but it may also help doctors stratify patients into groups that may one day benefit from different Alzheimer's treatment programs, as they do today for patients with several types of cancer. The experiment, Geschwind notes, was impressive because of the combination of database mining, lab validation, and imaging analysis of now standardized brain scans. "Five years ago they would never have been able to do this," he says.

Beyond biomedicine

The rapid rise in the number of plants that have had their whole genomes sequenced and made public has enabled plant biologists to produce their own dry lab discoveries. Buckler and his colleagues, for example, have been exploring disease resistance across the many species of maize, or corn. In one recent paper, they compared the genomes of 103 different maize species, analyzing 1000 different regions of DNA both within genes and nongene regions of the chromosomes. They linked certain traits, such as variation in disease resistance and in when the plant flowers, to specific patterns of the noncoding DNA. Now, Buckler says, his group and others are helping plant breeding programs improve disease resistance and other traits by singling out which offspring have nongene coding DNA signatures that promote desired traits. "Big data is already having a day-to-day effect on how people are breeding crops," Buckler says.

CREDIT: *SCIENCE TRANSLATIONAL MEDICINE*/AAAS

Published by AAAS

It's also helping answer more esoteric questions about plants. David Sankoff, a mathematician at the University of Ottawa, has tapped the whole genome sequences of some 30 flowering plant species to try to reconstruct the general genome architecture—not the specific DNA sequence—of the common ancestor of all flowering plants that lived some 120 million years ago. They recently reported a big step in that direction. By analyzing and comparing the presence of duplicate and triplicate copies of genes found within modern eudicots, one key branch of flowering plants, Sankoff's team concluded that the common ancestor had seven chromosomes and between 20,000 and 30,000 genes, making it a significantly smaller genome than many modern plants. Although such discoveries aren't likely to impact plant breeding or other commercial interests, "it's a really fun aspect of genetics work," says Eric Lyons, a plant geneticist at the University of Arizona in Tucson, who developed a comparative genomics database and software infrastructure used by Sankoff and his colleagues.

"You basically **don't need a wet lab** to explore biology."

—David Heckerman, Microsoft Research

CREDIT: MICROSOFT RESEARCH

Playing well together

Dry lab biology still faces plenty of growing pains. Among the most challenging is gaining access to other people's data. In many cases, researchers who have spent their careers generating powerful data sets are reluctant to share. They may be hoping to mine it themselves before others make discoveries based on their work. Or the data may be raw and in need of further analyses or annotation. "These are really hard problems," Butte says. "We need better systems to reward people that share their data."

A lack of common standards also handicaps the field. Not only do research groups file their data using different software tools and file formats, but also in many cases the design of the experiments—and therefore precisely what is being measured—can differ. Butte and others argue that dealing with multiple file formats is somewhat cumbersome but that the problem is surmountable. But it can be harder to account for differences in experimental design when comparing large data sets.

Years of work to standardize experiments, analysis, and interpretation of experiments involving tools such as DNA and RNA microarrays and proteomic mass spectrometry are beginning to pay off, Butte says. Heckerman agrees. Biological data, he says, are becoming "very standardized."

As the volume of publicly available data grows, so do concerns about genetic privacy. Geneticists have shown that even anonymous data can be "reidentified"—and any leaks can reveal not only the medical conditions of patients themselves, but also genetic predispositions to disease that other family members may share. In this case, however, at least one potential solution is already in place. In order to get access to the National Center for Biotechnology Information's database of genotypes and phenotypes (dbGaP), which archives studies such as GWAS associations and molecular diagnostic assays that attempt to link genes to traits, researchers must register and ask for approval. Furthermore, all such requests are made public, so that it's transparent who is attempting to gain access to the data and for what purpose.

To address these challenges—as well as take advantage of the scientific opportunities at the crossroads between big data and biomedical research—NIH announced this summer that it was launching a new project called Big Data to Knowledge (BD2K). With an initial funding of $96 million over 4 years, BD2K has dual aims. It will establish a series of centers to push the development of novel algorithms and other methodology to make discoveries, and it will also create a series of working groups across NIH's institutes to deal with the trouble spots of data standards, access, and privacy. Other efforts to grapple with these tough issues exist as well, including a global alliance of more than 70 institutions in 40 countries that was launched in June 2013 to make more digital data freely available.

Dry lab biology could receive a further boost from an upcoming U.S. requirement that databases be open to the community. On 22 February, a memo from John Holdren, the director of the U.S. Office of Science and Technology Policy (OSTP), asked the heads of executive departments and agencies within the federal government to come up with new strategies to encourage access to federally funded science and data. The memo drew attention at the time for its call for increasing open access to scientific publications. But what went largely unnoticed is that the memo also called for digital data from federally funded unclassified research projects to be stored in publicly available databases. OSTP officials say they have the agency recommendations now and are in the process of reviewing them.

While a potential boon for biology's data miners, access to unprecedented data sources will likely exacerbate problems with data standardization and issues of patient privacy, Butte says. It could also create new headaches for those required to submit their data. They will either have to take time themselves, or hire assistants, to manage the data sets and prepare them for deposition in a public source. And that could take dollars and expertise away from actual research. Particularly in small labs, this may be a significant impact, says Peter Lyster, a program director in the Division of Biomedical Technology, Bioinformatics, and Computational Biology at the National Institute of General Medical Sciences in Bethesda, Maryland. "At some point, it's a zero-sum game."

That's only for the wet labs that generate the data, he adds. For the new breed of dry lab biologists, the combination of new tools, new policies, and burgeoning databases holds nothing but opportunities. Says Heckerman: "I think we're full steam ahead at this point."

–ROBERT F. SERVICE

Published by AAAS