Data Deluge

Large-scale data collection and analysis have fundamentally altered the process and mind-set of biological research.

By Megan Scudellari | October 1, 2011

In the late 1970s, geneticist Robert Strausberg was an oddity. Instead of studying a single protein or gene, he focused on the expression patterns of the yeast mitochondrial genome. It was his first inkling of "what we could do if we had complete genomic information, though it wasn't being done at that time," recalls Strausberg, then at the University of Texas Health Science Center at Dallas.

A decade and a half later, Strausberg would tap back into that knowledge when James Watson invited him to the National Institutes of Health to lead the sequencing technology development program for the Human Genome Project (HGP). After he left the HGP, Strausberg initiated the National Cancer Institute's genomics project in 1997, and was soon an omics maestro, organizing collaborative teams to generate data sets of genomes, transcriptomes, epigenomes, and more. "It's a career path I never could have envisioned," he says, because it simply didn't exist when he started out. "It has transformed me as a scientist and the way I view the world."

And Strausberg is not alone. Today, the fundamental process of biology has changed, says computational biologist Eric Schadt, chief scientific officer of Pacific Biosciences, a DNA sequencing company. "There's definitely no question about it," he says. "I don't know that it's appreciated yet, but a major revolution is happening."

Biologists are no longer "heroes in isolation," querying single proteins or single genes, agrees genomics expert Stephen Friend, CEO of Sage Bionetworks. Instead, many biologists now work as part of collaborative systems, driven by high-throughput sequencing technologies, to study "omes"—collections of all the parts of a system. A genome, for example, is the collection of all the DNA in a cell or organism. "The need to embrace complexity has shifted the scale on which meaningful science can be done, from small to large," says Friend.

Genomics and beyond

Many scientists pin the start of the omics era to the HGP, a 13-year project formally begun in October 1990. "For biology, this was a different way of doing business," says Strausberg—biology's first real foray into "big science." So big that some critics called the pursuit "absurd" and "impossible." They were wrong on both counts.

[LOOKING BACK]

I expect that within a few years, our technology will be able to sequence one megabase/technician-year. At that rate 100 technicians could sequence the genome in 30 years.

—Harvard Nobel Laureate Walter Gilbert,
"Two Cheers for Human Gene Sequencing,"
The Scientist, October 20, 1986

Today, genomics still dominates other omics fields, though new ones seem to crop up on a daily basis. At Sage Bionetworks, an open-access repository for omics data, DNA data exists for 80 percent of the 10,000 individual human and mouse tissues and cell lines deposited, while only half the samples include RNA (transcriptomic) information, and less than 5 percent have proteomic or metabolomic data, says Friend. But that may soon change, as scientists continue to expand their scope to include more proteomic and metabolomic data, as well as dozens more data sets—from the antigenome to the vacuome.

One of the earliest omics study, in fact, involved a proteome, says John Weinstein, chair of bioinformatics and computational biology at the University of Texas MD Anderson Cancer Center in Houston. As the HGP was getting off the ground, the NCI began another "big science" project: a series of high-throughput DNA, RNA, and protein screens across 60 different human cancer cell lines, called the NCI-60. The idea of omics was so new that the first major paper out of the project, a database of protein expression across the cell lines, was initially rejected without review because the team presented data but not a hypothesis, says Weinstein, an author on the paper. "It was ahead of its time," he chuckles.

Since then, things have changed. Transcriptomics—the study of all the transcripts in a cell—is the latest, greatest "big business," says Mark Gerstein, a bioinformatician at Yale University. With the advent of next-generation RNA sequencing, or RNA-Seq, scientists can record all the gene transcripts in a cell—mRNAs, noncoding RNAs, and small RNAs—and quantify changing levels under different conditions, such as in diseased tissues. "To some degree, [transcriptomes] are fairly easy to interpret . . . and people are getting very practical knowledge from them," he adds. "We'll certainly see in the future RNA-Seq from every cell type and tissue in the human body."

Proteomics still awaits its own version of RNA-Seq, some fast technology to boost the output of protein data. Technology in the field is improving, says Eugene Kolker, chief data officer at Seattle Children's Hospital, who is involved in numerous proteomics studies. Still, for proteomics (and metabolomics, the profiling of metabolites to fully characterize a cell's metabolic pathways and processes) to be truly quantitative, he says, scientists must be able to accurately measure concentrations of proteins and other molecules in a sample, not just determine whether they are present or not—a refinement that is still in the works.

Bumps in the road

Notable Papers

F. Sanger, A.R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *J Mol*

With big science come big challenges. The first is finding ways to tame the data. “It’s now becoming unruly,” says Kolker. “We produce so much data, we’re not even always sure what we produce.” In fact, institutions often begin omics studies without sufficient storage capacity, software, or personnel to complete the work, says Weinstein. “There’s a tendency for institutions to buy the hardware and find that they can churn out terabytes and then to think afterwards about the problems of the data.” (See “Harnessing the Cloud,” October 2010; “At the Tipping Point,” February 2011; “Sequence Analysis 101,” March 2011 in *The Scientist.*)

Biol,94:441-48, 1975.

E.S. Lander et al., “Initial sequencing and analysis of the human genome,” *Nature*, 409:860-92, 2001.

M. Tyers, M. Mann, “From genomics to proteomics,” *Nature*, 422:193-97, 2003.

R. Goodacre et al., “Metabolomics by numbers: acquiring and understanding global metabolite data,” *Trends in Biotechnology*, 22:245-52, 2004.

M.Y. Hirai et al., “Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*,” *PNAS*, 101:10205-10, 2004.

But getting a handle on a single data set isn’t enough. Omics data will be most valuable when integrated, scientists agree, by layering various data sets, including DNA, RNA, and protein data, to get a comprehensive view of a cell’s activities. Yet that concept is so new and difficult that fewer than 20 people know how to integrate the layers into correct mathematical models, says Friend. Kolker agrees: “Our ability to cope with data analysis to produce meaningful information and knowledge is lagging.”

Despite the challenges, omics research is so popular that *OMICS: A Journal of Integrative Biology*, a publication launched 15 years ago with only 4 issues a year, now publishes 12 issues annually and is bursting at the seams with content, says editor-in-chief Kolker. “It’s becoming too much,” he says. “We’re trying to cover a really huge area.” The field is advancing so quickly, he adds, that the journal regularly receives papers on new omics of which they’ve never heard.

The impact of omics has spread beyond biology, to other research fields. The omics mind-set—of having and analyzing all the data for a single thing—has even spread to popular culture: “Culturomics,” the application of high-throughput data collection and analysis to the study of human culture, debuted in *Science* in December 2010, with an analysis of cultural trends within digital texts.

“Our ability to sequence and analyze genomes has gone far beyond what we all expected,” says Strausberg. Today, that observation is beginning to seem like an understatement.

OMICS A-Z

The completion of the Human Genome Project at the turn of the 21st century was a defining moment in omics. But these days, new omics fields crop up every day. Here’s a list of the hottest omics fields and the researchers who have made them so.

Genomics

The study of the full complement of an organism's DNA or specific parts of DNA sequences within the genome in order to map the positions and sequences of its nucleotides

- **Frederick Sanger and Alan Coulson:** MRC Laboratory of Molecular Biology, The Sanger Institute
- **Leroy Hood:** Institute for Systems Biology
- **Walter Fiers:** Ghent University, Belgium

Glycomics

The study of a cell's or organism's entire complement of sugar molecules, whether bound in more complex biomolecules or free, especially those that may play some role in diseases including cancer, Parkinson's, Alzheimer's, and AIDS

- **Carolyn Bertozzi:** University of California, Berkeley
- **Ram Sasisekharan:** Massachusetts Institute of Technology

Metabolomics

A way to characterize metabolic pathways in an organism or cell by studying chemical fingerprints, as represented by small molecules called metabolites, which result from cellular reactions

- **Douglas Kell:** University of Manchester
- **David Wishart:** University of Alberta, Canada
- **Jeremy Nicholson:** Imperial College London

Nutrigenomics

An outgrowth of genomics, transcriptomics, and proteomics, this field focuses specifically on the way that foods and general nutrition influence metabolic pathways, homeostatic control, and gene expression.

- **Michael Müller:** Wageningen University, The Netherlands
- **Raymond Rodriguez:** University of California, Davis

Pharmacogenomics

The study of how genetic variation influences responses to drugs or other chemical treatments

- **Richard Weinshilboum and Liewei Wang:** Mayo Clinic

Proteomics

The large-scale study of the functions and structures of proteins, especially the full body of proteins expressed in a cell, tissue, or organism

- **Patrick O'Farrell:** University of California, San Francisco
- **Donald Hunt:** University of Virginia

Transcriptomics

The study of the complete set of RNA transcripts produced by a particular genome at any one time—a global method for looking at gene-expression patterns

- **Stephen Fodor:** Affymetrix
- **Lubert Stryer**Stanford University

Vaccinomics

The study of tailoring vaccines to an individual's genetic profile with the goal of affecting the most robust immune response to a given pathogen

- **Gregory Poland:** Mayo Clinic
- **David Reif:** National Center for Computational Toxicology, US Environmental Protection Agency

Megan Scudellari is a freelance correspondent for **The Scientis**