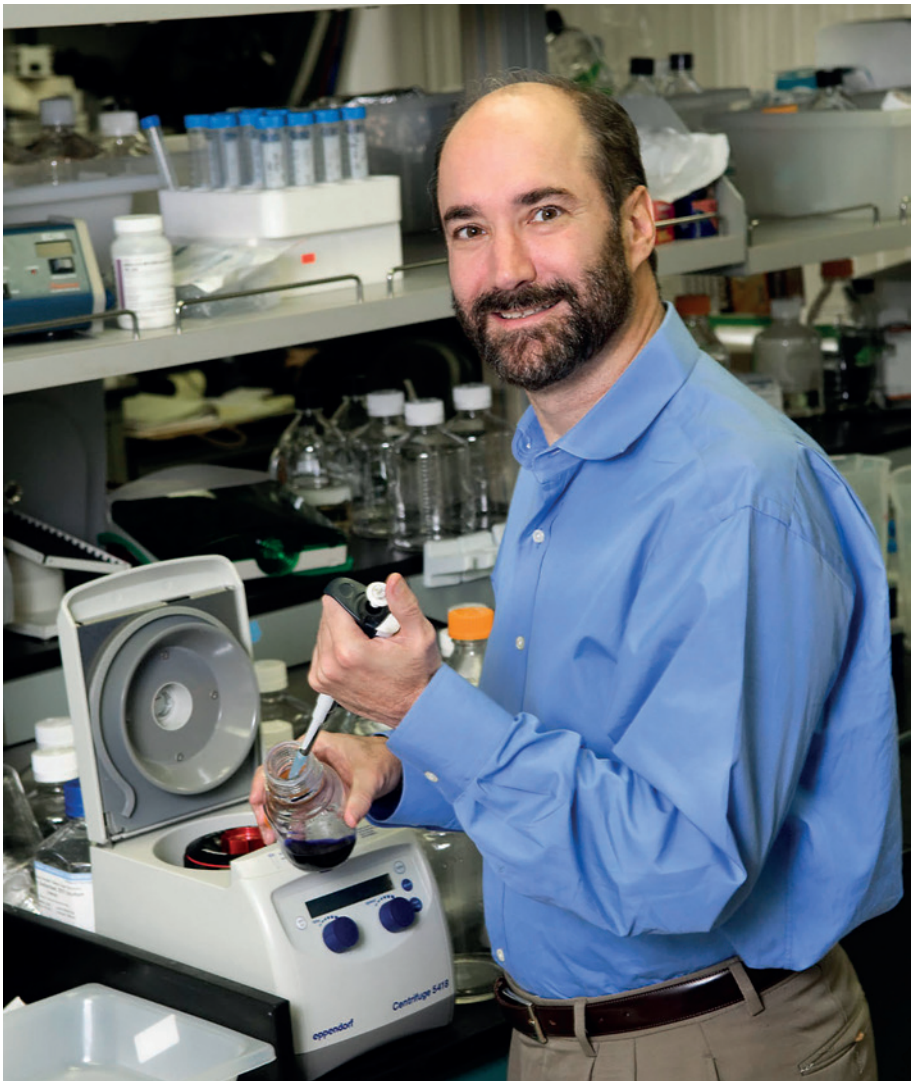


LEE ABEL/STANFORD DEPARTMENT OF GENETICS

Stanford researcher Michael Snyder analysed his own genome, RNA expression and protein production.

PROTEOMICS

# High-protein research

*The effort to catalogue proteins goes deeper in a push to make genetics research deliver practical benefits.*

BY NEIL SAVAGE

When Michael Snyder used the tools of '-omics' on himself, he was in for some surprises. Sequencing his genome, for instance, he discovered that he had a genetic predisposition for type 2 diabetes, even though he did not have any of the standard risk factors, such as obesity or family history of the disease. Over the next 14 months, Snyder, a molecular geneticist at Stanford University in California, repeatedly tested himself to monitor his RNA activity and protein production[1].

When he contracted a respiratory virus midway through the study, he watched as his protein expression changed and biological pathways were activated. Then he was diagnosed with diabetes — it looked to him as if the infection had triggered the condition. He also watched his proteins change during a bout of Lyme disease.

"I had no idea I'd turn out to be interesting," says Snyder, whose body has produced half a petabyte (500,000 gigabytes) of data so far. "It was just a proof of principle."

He has since expanded his study to 100 people, collecting measurements from the proteome and 13 other '-omes', including the proteome and transcriptome of the microorganisms that inhabit their bodies. He hopes that he and others can collect these deep profiles from a million patients, and apply the tools of big data to tease out differences that predict disease and provide a finer-grained understanding of various conditions. He also hopes that they can break conditions down into subtypes by their proteomic profiles. "There are probably 100 different types of diabetes," Snyder says.

Snyder's experience shows the power of using '-omics' to improve our understanding of biology, says William Hancock, a protein chemist at Northeastern University in Boston, Massachusetts.

## PRACTICAL GENETICS

Genes provide the instruction manual for biological processes, but it is the proteins they create that turn those instructions into reality. Huge international efforts are under way to identify proteins, map their locations in tissue and cells, count how many are produced in particular circumstances, and describe the various forms they can take. And the oceans of data from these searches will uncover biomarkers for diseases and provide targets for drugs to treat various conditions. By combining proteomics with genomics, transcriptomics, metabolomics and other '-omics', scientists may further deepen their understanding of biology on a molecular level.

Proteomics brings genetic information to a practical level, says Gilbert Omenn, a bioinformatician at the University of Michigan in Ann Arbor and chair of the global Human Proteome Project (HPP). The idea of the project is to create a "complete parts list" of the human body, he says, "to fill in the many blank spots between knowing that a gene has something to do with a disease process and knowing how it really works".

That is quite a parts list. The human body contains roughly 20,000 genes that are capable of producing proteins. Each gene can produce multiple forms of a protein, and these in turn can be decorated with several post-translational modifications: they can have phosphate or methyl groups attached, or be joined to lipids or carbohydrates, all of which affect their function. "The number of potential molecules you can make from one gene is huge," says Bernhard Küster, who studies proteomics at the

NATURE.COM
Find a review of how proteomics affects cell biology here:
go.nature.com/akxnp7

Technical University of Munich in Germany. "It's very hard to estimate, but I wouldn't be surprised to have in one cell type 100,000 or more different proteins."

## GLOBAL MAPPING

Proteomics research is an international enterprise. The Human Proteome Organization created two complementary HPP projects, both of which use mass spectrometry. One, the Chromosome-based HPP, divided the 24 chromosomes among 19 countries. Japan, for example, is tackling chromosome 3 and the X chromosome, and Iran is studying the Y. The second, the Biology/Disease-driven HPP, is looking for proteins in specific tissues and organs, focusing on those that are relevant to diseases such as diabetes and colon cancer. A separate global project, the Human Protein Atlas, relies on antibodies with fluorescent molecules or other tags attached that bind to specific proteins to identify them.

There are also some significant national efforts. China is investing heavily in proteomics research, with one example being a new national laboratory called PHOENIX, which was set to open in October with annual funding of US$10 million.

Whatever the technical approach, mapping the human proteome is no easy task. The genome is simple in comparison — it is assembled with just four nucleic acids and changes little over a person's lifetime, except in the special case of cancer. Proteins, on the other hand, vary over time, changing during exercise, disease and menstrual cycles, for example. Another complication is that the most abundant protein can be about 10 billion times as common as the least. "You have one genome and you have a gazillion proteomes, depending on the environmental situation," says Hancock, who is co-chair of the Chromosome-based HPP.

"There is no such thing as a human proteome in one person, let alone in many people," says Küster. Last year, his group published a draft map[2] of a human proteome based on 16,857 mass-spectrometry measurements of human tissue, cell lines and body fluids. They also created a database, ProteomicsDB, to provide analysis of the data.

## TOO MUCH DATA?

Just figuring out how to handle the volume of proteomics data is tough. The Human Protein Atlas, for instance, collects images of tissues with tagged antibodies. Each image takes up tens of megabytes, and compressed jpeg files about 10 megabytes in size are made available for online distribution.

Meanwhile, the European Bioinformatics Institute (EBI) in Hinxton, UK, is creating ELIXIR, a distributed-computing infrastructure designed to share proteomics and other biology data among research institutions in Europe. "ELIXIR doesn't want to create a huge database — they want to link different groups and different countries," says Mathias Uhlén, a microbiologist at the KTH Royal Institute of Technology in Stockholm, Sweden. The EBI is already the repository for the Protein Identifications (PRIDE) database, which collects mass-spectrometry data generated by multiple research groups.

But scientists often disagree about whether to keep the raw data or throw it away. "The methods for identifying proteins from raw data are constantly improving, so it makes sense to keep the raw data if you can — but it does take lots of space," says Conrad Bessant, a bioinformatician at Queen Mary University of London. The argument on the other side, he says, is that "the field is advancing so quickly that why would you look at a five-year-old data set? You might as well run the analysis again, because the instruments are so much better."

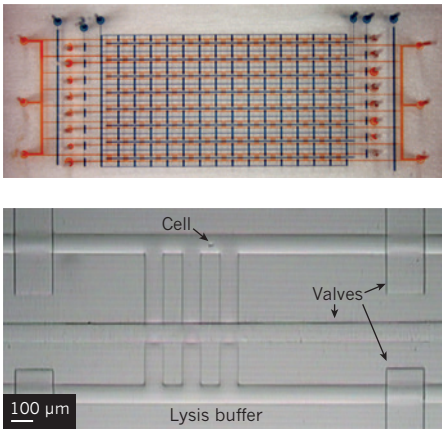
## FILLING IN THE MAPS

Proteome data are far from perfect, however. In the issue of *Nature* last May in which Küster's group reported their results, another group of scientists from the United States and India published a draft map[3] said to cover about 84% of the protein-coding genes in the human genome. Both maps were based on mass spectrometry: an enzyme digests proteins and produces peptide sequences about 7 to 30 amino acids long, and the mass of these peptides is used to deduce the protein's composition. And both projects ended up reducing the number of proteins they claimed to have found, after other scientists called into question some of their interpretations[4]. Mass spectrometry is a probabilistic method, says Omenn, and there is no way to exclude the possibility that two different proteins produced the same peptide sequence.

*"When it comes to big data, it's easier to generate the data than to get knowledge out of it."*

The Human Protein Atlas's antibody-based detection, on the other hand, is non-probabilistic, as it tags individual proteins. The advantage of this approach, argues Uhlén, one of the creators of the Atlas, is that it shows precisely in which organs, tissues and even cells the proteins are located. "What we are providing is a map of where the proteins are," Uhlén says. "That gives you hints about the function of the proteins."

Recent years have seen a push to develop microfluidic chips on which to perform antibody-based single-cell proteomics. This approach is particularly important when the cells of interest are rare, as in the case of circulating tumour cells. It also allows investigators to study differences between populations of the same cell type. For example, if one tumour cell makes many more copies of a particular protein than its neighbour, or the proteins in one cell have a methyl group attached whereas those in another cell do not, this could explain how the tumour develops drug resistance, leading to possible targets for therapeutics.




**A proteomics chip (top) profiles individually labelled cells in its microchambers (bottom).**

However, even the antibody approach has limitations, as some antibodies can bind to more than one protein, creating misleading results. "An even harder problem is knowing what data are of good quality and what are not," says Uhlén. "When it comes to big data, it's easier to generate the data than to get knowledge out of it."

Then there are the missing proteins. Roughly 15% of human genes that should encode proteins have had no associated protein identified[5] — that means there are nearly 3,000 missing proteins. In some cases, this may be because they occur in small amounts or in only tiny areas of tissue. Without a complete catalogue of proteins, the overall picture of human proteomics remains fuzzy.

Computing with incomplete or inaccurate data could lead researchers astray, Hancock worries. "Bringing biology and mathematics together is a match made in hell," he says. "Biology is wet and dirty and messy."

But as measurement techniques improve and scientists amass more findings, "the picture is going to get sharper and sharper," Hancock adds. And the sheer volume of data available to sift through will continue to soar as measurement techniques improve. "We get all kinds of data from many different experiments," Bessant says. "It doesn't take long until you get hundreds of gigabytes or terabytes of data." ■

**Neil Savage** *is a freelance science and technology writer based in Lowell, Massachusetts.*

1. Chen, R. *et al. Cell* **148,** 1293–1307 (2012).
2. Wilhelm, M. *et al. Nature* **509,** 582–587 (2014).
3. Kim, M.-S. *et al. Nature* **509,** 575–581 (2014).
4. Ezkurdia, I., Vázquez, J., Valencia, A. & Tress, M. *J. Proteome Res.* **13,** 3854–3855 (2014).
5. Horvatovich, P. *et al. J. Proteome Res.* **14,** 3415–3431 (2015).