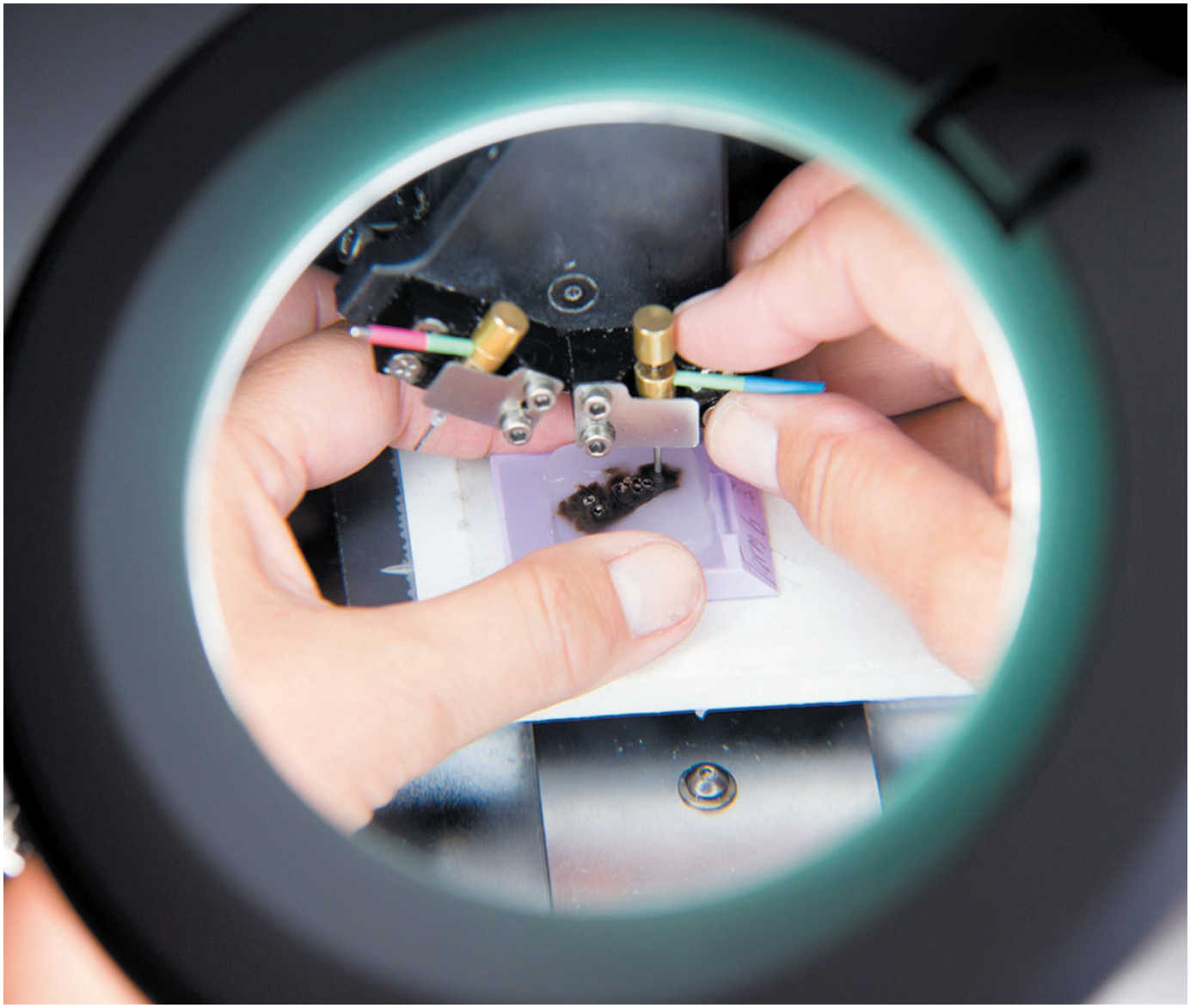
TECHNOLOGY FEATURE

AN ATLAS OF EXPRESSION

The first draft of the complete human proteome has been more than a decade in the making. In the process, the effort has also delivered lessons about technology and biology.

HUMAN PROTEIN ATLAS



A researcher removes small sections from a tissue sample to analyse its protein expression for the Human Protein Atlas project.

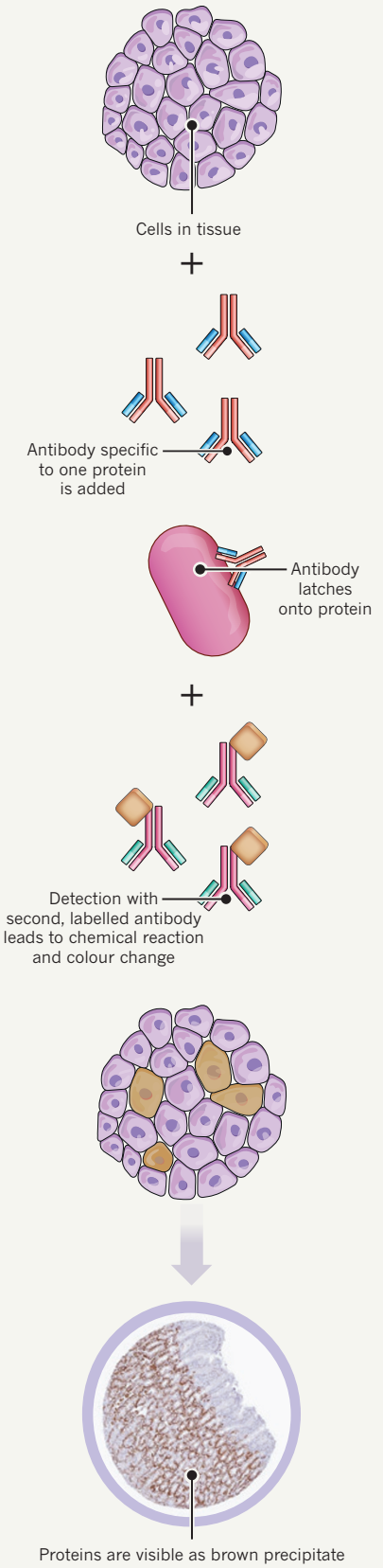
BY VIVIEN MARX

In 2003, two years after the Human Genome Project published a first-draft sequence of the roughly 20,000 genes that define *Homo sapiens*, a group of Swedish researchers launched the Human Protein Atlas (HPA) — a large-scale effort to map where the proteins encoded by those genes are expressed in the body's tissues and cells.

Few proteins had been localized in that way, explains Mathias Uhlén, a protein researcher at the Royal Institute of Technology in Stockholm, and the HPA's principal investigator. A comprehensive atlas of the human 'proteome' would set the stage for more-sophisticated study of protein function, he says. It would reveal the array of membrane proteins, which ferry molecules in and out of cells, and expose the 'secretome' — the proteins secreted by cells in health and in disease. It would guide exploration into the physiological impact of genetic variation. And it would help drug developers to predict where a candidate drug might interact with a protein or cause side effects (see 'Uses for the Human Protein Atlas').

EXPRESS YOURSELF

Cells in a given tissue may look the same but express different proteins. With the help of tissue stains and antibodies, proteins can be tagged and rendered visible.



MAP QUEST

Uses for the Human Protein Atlas

- Look at the spatial distribution of a protein in tissues or cells of interest.
- Study a protein in many different tissues.
- Search for protein-expression patterns that match those of a protein of interest.
- Select cell lines for experiments on the basis of their RNA-expression data and choose corresponding antibodies.
- Begin early studies of protein function.
- Compare normal and cancerous tissues to help to find biomarkers.
- Compare with protein-location data sets generated in other labs.
- Find antibodies for certain applications.
- Compare protein expression in cancerous and normal tissues.
- Use as a teaching tool for cell biology and histology.

The project's annual releases of preliminary maps and data have already yielded surprises, says Fredrik Pontén, a pathologist at Uppsala University in Sweden and a co-founder of the HPA. For example, the maps show that some 3,500 genes encode proteins specific to one tissue type or a small group of tissues, and that the testis contains one-third of those proteins — more than in any other organ[1] (see 'Our proteins, ourselves').

In November, the team plans to upload data completing the first draft of the human proteome. In total, it includes nearly 15 million high-resolution micrographs of stained tissues and cells and presents data on about 80% of human proteins. This information was gleaned from 46 human cell lines and tissue samples from 360 people — 44 normal tissue types from 144 people, and the 20 most common types of cancer, from 216 people. "What's most exciting is the scale of it," says Pontén. Another project — dubbed the subcellular protein atlas — will locate proteins within cells, and will be completed next year.

The samples used for the HPA came from people who agreed to donate tissue while they were being treated for various conditions. In accordance with Swedish research-ethics guidelines, all samples have been anonymized so that they cannot be traced to their donors.

The HPA is not the only protein-mapping venture. Other efforts include the Human Proteome Project, which is loosely affiliated with the HPA but uses some different strategies, and projects in individual labs or among groups of researchers. Two drafts of the human proteome, based on mass spectrometry, are presented in this issue of *Nature*[2,3]. In this approach, tissues are processed and their proteins broken into fragments. The fragments are ionized and then separated according to their mass and charge, which helps to measure, sort and identify the proteins.

The HPA is the only large-scale mapping effort based on antibody-profiling methods, in which chemical stains and antibodies are used to locate and identify proteins in tissues (see 'Express yourself'). The project's scientists used a 'brute force' approach that includes some automation but requires many manual steps. Applying their techniques to the many proteins in the body is daunting — the sum of human proteins exceeds the number of genes by far, and could run in the millions.

With funding from the Knut and Alice Wallenberg Foundation in Stockholm, the HPA researchers — 140 scientists in 13 groups working mainly in Stockholm, Uppsala or Mumbai, India — divided up the work. They scaled up standard proteomics approaches, such as immunohistochemistry, in which antibodies and stains are used to visualize proteins in tissue samples, and western-blot assays to check the specificity of antibodies.

There is a need for such corroboration — often, different approaches yield contradictory results. The researchers have spent time, effort and resources to ensure that the antibodies they use work as expected, and to develop reliable ways to interpret and classify the stained samples, Uhlén says.

SCALING UP

The researchers have also struggled with the general lack of automation in proteomics, says Caroline Kampf, a proteomics researcher and director of the HPA's Uppsala site, who joined the project when it began. That situation has improved over the past decade.

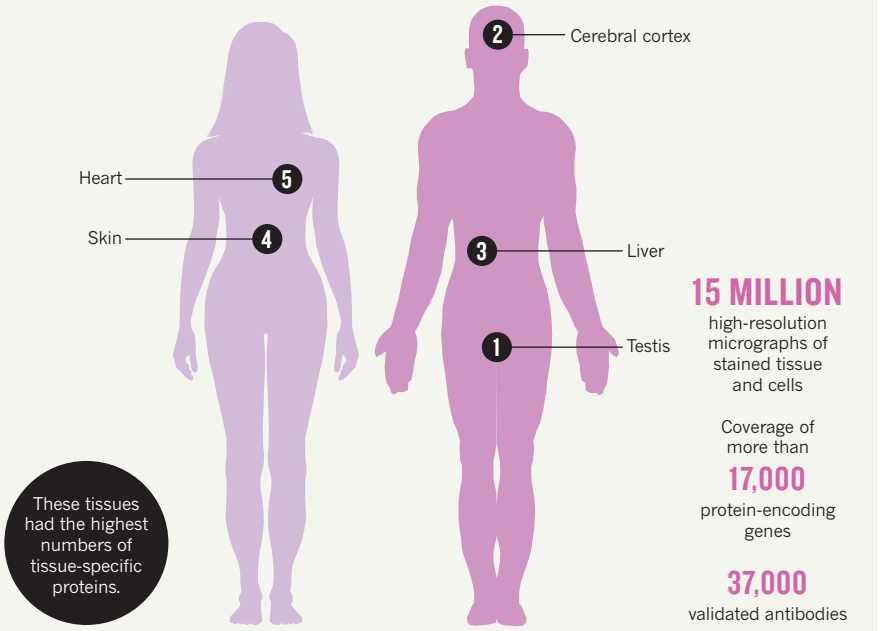
"The Human Protein Atlas has delivered many lessons about tissue staining and large-scale projects."

The task of locating proteins begins with preparation of tissue samples that Kampf and her team receive from Uppsala Biobank, which oversees management of the tissues. The scientists consult research literature and protein databases and also study the tissue's messenger RNA (mRNA), a molecule that carries the information used to manufacture proteins from a DNA sequence.

In 2012, the researchers began using high-throughput mRNA sequencing, or 'RNA-seq', which has enabled them to more quickly obtain data about the set of genes expressed as proteins in a given tissue. This approach has helped to validate results from antibody-based

OUR PROTEINS, OURSELVES

The Human Protein Atlas is based on data from 46 cell lines and from samples of normal and cancerous tissues from 360 people. The five tissues found to have the highest numbers of tissue-specific proteins are shown.



15 MILLION
high-resolution micrographs of stained tissue and cells

Coverage of more than
17,000
protein-encoding genes

37,000
validated antibodies

HUMAN PROTEIN ATLAS

Mathias Uhlén directs the Human Protein Atlas.

tissue profiles, says Kampf.

Tissue preparation involves preserving tissue samples in blocks of paraffin and then processing them into microarrays, groups of tiny tissue samples arranged in a grid to enable scientists to test for the presence of many proteins. Only a fraction of the tissue is used for the microarray; the rest is kept in a repository[4]. For the microarray, cylindrical 'cores' of tissue of around 1 millimetre in diameter are punched out of the paraffin blocks. These cores are embedded in rows and columns in another paraffin block, which is then sliced into thin sections and placed on a slide for staining. The researchers produce around 100 slides a day.

TECH SUPPORT

Although production of tissue microarrays can be automated, says Kampf, plenty can go wrong, and skilled technicians are needed for troubleshooting and for their craftsmanship. For example, she says, a technician needs the right touch to know when to stop tissue punching, or when the punch is stuck. The degree of intervention required depends in part on the sample — differences in texture mean that some tissue types are more challenging than others. Skin, for example, is tougher than fatty breast tissue. To accelerate production, the scientists group similar tissue types for similar processing steps. Some steps still require manual labour. "But when punching cell lines or punching cancers, you can more easily let a machine do it because [the tissue has] a more homogeneous composition," Kampf says.

Once a tissue sample is on the slide, it is treated with reagents that bind specifically to one protein and not another. Another challenging aspect of the procedure is that the concentration of a protein in a given tissue can vary within and between samples. In the body, protein abundance varies by as much as 1-million-fold. The abundant proteins can swamp out the more scarce ones and make them hard to detect.

There are many types of affinity reagents, both biological and synthetic. The HPA uses polyclonal antibodies, which recognize portions of specific proteins. They are produced by injection of an antigen into laboratory animals and later harvesting of the antibodies the animals produce in response. These antibodies are not identical to one another.

"It is always easier to interpret your results if you spent time and effort validating reagents."

The HPA teams have developed methods of antigen design and antibody purification that maximize the specificity with which an antibody latches on to a protein, raising the probability that it will locate one protein only. To date, researchers have validated 37,000 antibodies for the project.

The antibodies developed in the course of the HPA project are available through Atlas Antibodies, an HPA spin-out in Stockholm. The company now sells 17,000 polyclonal antibodies developed by the HPA, and expects to add another 2,000 by November, says Marianne Hansson, the firm's chief executive officer. What is special about this collection of antibodies is that they are produced in a uniform fashion and are 'proteome-wide', says Henrik Wernérus, chief scientific officer at Atlas.

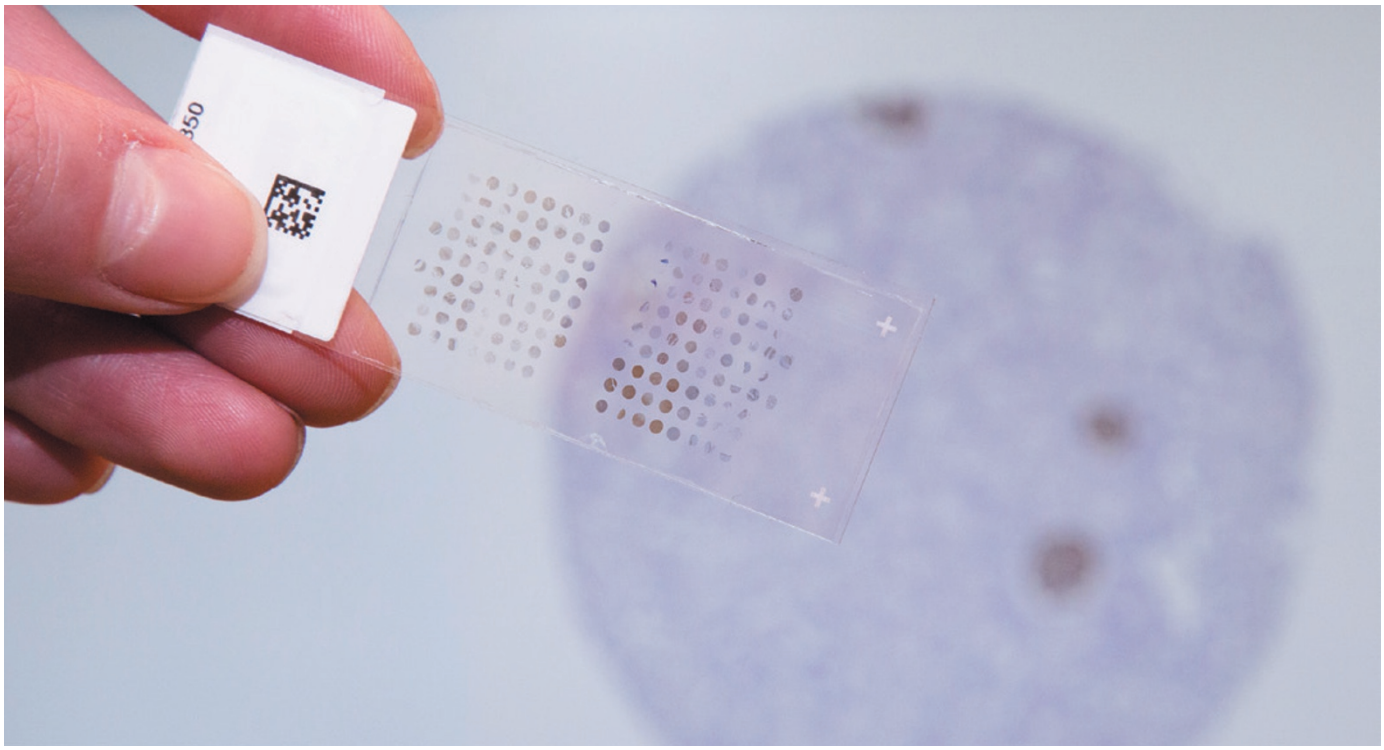
One validation test for antibodies is the tried-and-true western blot, in which proteins are propelled through a gel by an electric field and separated into fragments of different molecular weights. The fragments are then transferred to a material such as nitrocellulose paper and probed with a primary antibody, which binds the protein of interest, followed by a secondary antibody that binds the primary one and bears a fluorescent or enzymatic tag to enable detection of the protein.

Each step in the mapping process takes time, and there are bottlenecks. For example, the project's tissue slides quickly piled up. An automated system cannot readily emulate a pathologist changing the focus on a microscope to study a tissue's protein-expression patterns. Since the HPA began, automated slide readers have become available and are now used to supplement the educated eyes of the project's pathologists.

Kampf says that the first slide scanners her team used could handle around 10 slides a day. But she needed to scan more than 100 slides a day. Now, she says, a scientist can "go home, and when you are back in the morning, they are all there".

But even as automated slide scanning emerged, Kampf and her staff still had to load the slides manually. And the first auto-loaders were far from perfect. Kampf recalls coming into the lab after a coffee break to find the floor covered in glass because an instrument's

HUMAN PROTEIN ATLAS



HUMAN PROTEIN ATLAS

A slide of tissue sections to be stained and incubated with antibodies to see which proteins are expressed.

grippers had smashed the slides against the wall. She spent many an evening on the phone with companies in various time zones to address technical issues.

Kampf says that scaling up is not just about data generation and analysis. "You encounter many other things that you haven't expected," she says. The staining patterns in tumour tissue, for example, can be difficult to interpret, in part because such tissues contain a mix of normal and diseased cells. Cancer cells also show variation — within a tumour as well as between tumour types and patients. Given this variability, data need to be carefully evaluated.

OBJECT LESSONS

Pontén says that the HPA has delivered many lessons about immunohistochemistry and large-scale projects. The scientists have found, for example, that RNA-seq data can help to shore up the results of immunohistochemistry. And they have grappled with differing results between immunohistochemistry and western blots. Whereas the proteins analysed in a western blot have been denatured into two-dimensional fragments, those probed in tissue samples are more likely to retain their three-dimensional (3D) structure. That distinction could lead to differing results, as some antibodies recognize proteins only in their 3D form.

One solution has been to produce the antibodies in a different way, says Pontén. The antigens used conventionally to create antibodies are peptides, short strings of amino acids. In their approach to making antigens, the HPA team designed longer protein fragments. The fragments are 50–150 amino acids in length and are selected to have the highest likelihood of yielding unique proteins, and to contain two or three epitopes — the sites that antibodies recognize on a protein. This multiplicity increases the likelihood that the resulting antibodies will work in western blots, immunohistochemistry and other techniques, Pontén says.

Caroline Kampf says that slide scanners could once handle 10 slides a day but now can handle more than 100.

The results of these assays depend not only on antibody affinity, but also on the relative abundance of a target protein, says Uhlén. Because it can be tough to detect low-abundance proteins, results from antibody staining should be validated with a different antibody for the same protein, for example, or by checking the RNA-seq data to be sure that the RNA that gives rise to the protein in question is present in the tissue, he says.

Antibody validation has "turned out to be even more important than expected", says Emma Lundberg, who directs the subcellular protein atlas project, which shows where in a cell a particular protein is present. Antibodies can have affinity for proteins other than their targets, leading to cross-reactivity that can confuse protein-mapping results. "In the end, it is always easier to interpret your results if you spent some time and effort validating your reagents," she says.

Kampf says that she has, over the years, received many e-mails from scientists asking when their favourite proteins will be published in the atlas. All she could tell them was to be patient. The process, from designing an antigen to having fully validated results ready to upload to the HPA website, takes 9–12 months on average.

The researchers still have to refine the map, says Pontén. There are many low-abundance proteins, among others, to find. And some proteins have been overlooked because they are expressed only at certain times during development, says Pontén. Others have gone undetected because they are present in tissues that the HPA has only begun to collect, such as the retina.

The team hopes that the research community will help with spotting inaccuracies in the map. Pontén says that it will take at least another five years of curation to offer the research community HPA data that have been analysed and assessed to the level of 'textbook' authority. ■

Vivien Marx *is technology editor for* Nature *and* Nature Methods.

1. Fagerberg, L. *et al. Mol. Cell Proteomics* **13,** 397–406 (2014).
2. Kim, M.-S. *et al. Nature* **509,** 575–581 (2014).
3. Wilhelm, M. *et al. Nature* **509,** 582–587 (2014).
4. Kampf, C. Olsson, I., Ryberg, U., Sjöstedt, E. & Pontén, F. *J. Vis. Exp.* **63,** e3620 (2012).