Tech News

The Human Proteome Project Takes Shape Down Under

The 'omics suffix is synonymous with 'Big Science.' Genomics, transcriptomics, metabolomics—thanks to the Human Genome Project, any branch of molecular science with those five letters tacked onto the end conjures images of biology writ large.

Most often, the science doesn't merit the hype; 'omics experiments rarely yield genome-scale datasets. But this past September, during the Human Proteome Organization's (HUPO) annual conference in Sydney, Australia, the world's proteomics community came together to embrace the challenge implicit in its name. They officially launched the Human Proteome Project (HPP), an attempt to characterize the expression, abundance, and localization of at least one representative isoform of each of the 20,300-odd protein-coding genes in the human genome.

Bolstered by "three pillars"—mass spectrometry (the defining technology of proteomics), antibodies, and knowledge integration (or informatics)—the HPP aims to "give biological meaning to the Human Genome Project," according to Gil Omenn of the University of Michigan, who is codirecting one of the HPP pilot projects. While DNA gets most of the attention, proteins do most of the work: they read DNA and translate proteins, catalyze chemical reactions, and transmit signals. Proteins give cells their characteristic shape, and help them differentiate. More to the point, protein abnormalities are the proximal cause of most genetic diseases, and are the primary targets of most therapeutics.

"Proteins represent the machinery of life," says Eric Deutsch of the Institute for Systems Biology in Seattle. Yet on a large-scale 'omics level, they are frustratingly difficult to detect and quantify, especially for non-experts—a function of their vast heterogeneity, different ionization efficiencies, post-translational modifications, and wide dynamic range, among other things. According to one recent review, "about 35% of predicted human proteins have yet to be observed reliably by [mass spectrometry]" (3). Even relatively simple samples can prove challenging. In 2009, HUPO conducted a test, asking 27 labs to analyze a sample containing 20 different proteins in equimolar amounts. Only seven labs correctly identified the 20

Eyes on the prize. Swiss Institute of Bioinformatics group leader Amos Bairoch, a member of the Human Proteome Project working group, is coordinating one of the effort's three "pillars." Credit: Nicolas Righetti, Swiss Institute of Bioinformatics.

proteins exactly, though post-hoc analysis of the data indicated that was mostly a problem of data interpretation and reporting, not collection (1).

All agree the problem is infinitely more complex than the human genome ever was; after all, the proteome is a protean thing. Nevertheless, says Omenn, the community feels that overall, advances in technology, expertise, and informatics now make the human proteome a tractable problem. "After years of discussion, we think the time is really ripe," he says.

Which Human Proteome?

That's not to say it will be easy. The HPP naturally evokes comparisons to the Human Genome Project, and according to Tommy Nilsson, director of the proteomics and systems medicine program at the Research Institute of the McGill University Health Centre in Quebec (who helped craft the vision for the HPP in the seminal 2008

"Barbados whitepaper"), the two "are very similar in their spirit." But beyond the obvious similarities of scale and scope, the two projects couldn't be more different.

For one thing, proteomics data are more complex. DNA sequencers literally sequence DNA, but mass spectrometers identify proteins by smashing them into smaller and smaller pieces and then matching the resulting masses against precomputed lists of all possible proteins that could have generated them. The result "is a statistical analysis," explains Amos Bairoch, a group leader at the Swiss Institute of Bioinformatics, who is helping to coordinate the informatics pillar, and this means that results can vary with the quality of the sample preparation, mass spectrometer, reference library, and data analysis.

Moreover, the Human Genome Project was not open-ended, Nilsson says. It had a clear goal, and easily understood milestones: sequencing every one of three billion bases that make up the human genome. The

How to build a reference library. To develop the mass spectrometry pillar's library of verified SRM assays, human proteins are digested with proteinases in silico to create lists of product peptides. Those peptides are synthesized and characterized using tandem mass spectrometry to identify the best quantitative assays. Credit: Paola Picotti, ETH Zurich.

simplicity of that vision, coupled with the promise of inevitable medical breakthroughs, made the HGP easy to sell to funding agencies, and to the general public that was footing the bill.

By comparison, says Bairoch, the HPP is a far tougher pitch. For one thing, there is no single human proteome. The human body contains “some 230 different cell types ... each presumably expressing only a subset of the 20,300 predicted proteins,” he, Nilsson, and others wrote in 2010, in a review entitled “Mass spectrometry in high-throughput proteomics: Ready for the big time” (3). (“We wanted to have a question mark in the title, but we were prompted not to,” confides Nilsson.) Between splice variants and post-translational modifications, environmental influences and divergent differentiation programs, the proteome is a moving, evolving target. “It’s a fractal system,” Bairoch says, “the more you go deep, the more you have to do.” Just one class of modification, glycosylation, represents “an enormous permutational landscape,” says Nilsson.

And just what does it mean, exactly, to complete a proteome? Simply inventorying proteins isn’t the answer, says Ruedi Aebersold, professor for molecular systems biology at the Swiss Federal Institute of Technology in Zürich, Switzerland; rather, researchers want to “understand what [proteins] are doing in the cell and how the proteome is dynamically changing.”

One possible approach, laid out in “Ready for the big time” (3), involves sampling repeatedly until each element of the proteome has been detected at least once. “By combining fractionation and replicate analyses to collect as much data as possible, it is possible to develop saturation curves,” the authors explain. That was the strategy taken in a 2008 study, where Matthias Mann and his colleagues at the Max Planck Institute for Biochemistry in Martinsried, Germany, produced what may be the first complete comparative proteome of a eukaryotic

organism. The analysis used a combination of three parallel approaches to identify 4399 proteins from the haploid and diploid states of the yeast, *Saccharomyces cerevisiae* (2).

Adopt a Chromosome

To achieve that end for the human proteome, the HPP’s first phase will use a so-called “gene-centric” approach to detect at least one protein isoform from each of the protein-coding genes in the proteome, currently numbering 20,253, according to UniProt (www.uniprot.org). Later stages will switch to a protein-centric approach, which will address the many different splice variants and post-translational modifications of those different proteins. These variants could number in the millions.

To engage the worldwide scientific community, divvy up the workload, and fund the effort, the project will proceed via an adopt-a-chromosome model, in which consortia in individual countries will accept responsibility for all the proteins on specific

“It’s a fractal system. The more you go deep, the more you have to do.”

chromosomes. According to William Hancock of Northeastern University in Boston, MA—one of the co-PIs leading a team focused on chromosome 17—the approach “gives a problem with a boundary. Proteomics at this stage is so large that you need to have initial goals that are achievable.”

As of the Sydney meeting, nine chromosomes had been at least provisionally spoken for: chromosomes 3 (Japan), 8 (China),

BioTechniques is the first peer-reviewed journal with open access to the whole of the life science community.

100% life science
100% peer-reviewed methods
100% “everyday” practical information

Maybe it’s time you get your own free subscription copy too, or maybe your subscription has recently expired. Don’t worry, it’s easy and free to subscribe or renew to *BioTechniques*.

Go to:
www.BioTechniques.com/subscribe

13 (Korea), 17 (USA), 18 (Russia), 19 (Sweden/Japan), 21 (Canada/Sweden), X (Japan), and Y (Iran). Canada may take on one more chromosome (6), says McGill's John Bergeron, a past president of HUPO who coauthored the Barbados white-paper. Other European consortia have also expressed interest, says Pierre Legrain of the French Alternative Energies and Atomic Energy Commission—who heads up the HPP working group—but adding that to

Ruedi Aebersold, of the Swiss Federal Institute of Technology in Zurich, is guiding the HPP's mass spectrometry pillar. Credit: ETH Zurich.

date, only projects in Korea, Russia, and Iran are actually up and running.

But not everyone believes this chromosomal plan is necessarily the best approach. From a technical point of view, proteomics experiments cannot be constrained in that way; they tend to pick up whatever proteins are in a sample, regardless of which chromosomes those proteins came from. Aebersold calls it "somewhat misguided" to expect that ongoing large-scale projects should stop after completing one discrete chromosome's worth of work. Yet according to Omenn, who also is on the chromosome 17 team, the adopt-a-chromosome strategy enables researchers—who will ultimately have to convince funding agencies for support—to dovetail the HPP mission with their existing projects. For instance, chromosome 17 includes several breast cancer-related genes, which complements nicely with Hancock's own research. Similarly, the Korean team chose chromosome 13 for its role in diabetes.

Three Pillars

The HPP as discussed in Sydney will differ slightly from the vision laid out in

an article published in early 2010, which initially articulated the HUPO community's plans (4). That initial article had three objectives: a "protein parts list," "protein distribution atlas," and "protein pathway and network map." In contrast, the organizational scheme derived in Sydney includes pillars for mass spectrometry (parts list), antibodies (distribution atlas), and knowledge integration (informatics). According to Legrain, that change reflects the fact that the anticipated biological annotation of the proteome should fill in the interactome data without specifically targeting that kind of work directly.

The antibody pillar, led by Mathias Uhlen of the Royal Institute of Technology in Stockholm, will produce at least one antibody for every human protein for use in Western blotting, immunoprecipitation, and tissue localization experiments. That work is already ongoing; as of the Sydney meeting, some 13,150 antibodies targeting more than 10,100 proteins had been produced and tested against 48 different tissues and 20 tumor types for version 7 of the Human Protein Atlas (HPA; www.proteinatlas.org), where both localization data and links for obtaining the antibodies are available.

On the mass spectrometry side, Aebersold—along with Robert Moritz at the Institute for Systems Biology in Seattle, WA—has worked with colleagues to assemble a collection of proteotypic peptides and fragment ion reference spectra for nearly all human proteins that can be used in targeted proteomics experiments. Essentially, says Aebersold, these peptides provide positive control samples for every possible human protein, even ones that have never been experimentally observed. "If a protein is in a sample and this sample is subjected to mass spectrometry analysis, we know how this protein would look like to the mass spectrometer," he says. "So we basically have a complete human proteome reference library for MS spectra." In Sydney, Aebersold announced that he had generated nearly 19,960 such peptides already, representing more than 98% of the predicted human proteome.

According to Aebersold, providing a complete set of pre-validated, consistent reagents and tools will enable the broader biology community. For instance, a researcher new to the field could visit the Peptide-Atlas web site (www.peptideatlas.org), download the correct reference spectra for a particular protein of interest, and use those to unequivocally assess the abundance of those proteins. The alternative, he says, is for new researchers to reinvent the wheel

Pippin Prep™

DNA Size Selection System
For NGS Library Construction

- Tight Size Distributions
- High Recoveries
- Minutes of Hands-On-Time

888.744.2244

www.sagescience.com

each time. "The net effect," Aebersold says, "would be that better-quality and more-consistent proteomic data can be produced in many, many labs that currently do not have the ability to produce such datasets."

The informatics pillar will focus on the development of a human proteomics knowledgebase. Called neXtProt (a beta version was just released at <http://beta.nextprot.org>), it will aggregate HPP data, as well as disease-related information, genetic variants, and expression profiles in a single unified resource. "neXtProt is not going to be specifically about the Human Proteome Project," Bairoch says, "it's a platform for having information on human proteins."

In other words, neXtProt will be the glue that binds the project together and gives it biological context. "This would be basically the connection to the two other pillars and would allow researchers to seamlessly go beyond antibody-based and mass spectrometry-based measurements," Aebersold says.

Other bioinformatics initiatives in the HPP will (i) aim to build a kind of "corporate" web portal for the project, (ii) encourage the deposition of correctly formatted and annotated data in existing data repositories, and (iii) collect, reanalyze, and curate those data via the ProteomExchange framework at PeptideAtlas, UniProt, and the GPMdb (gpmdb.thegpm.org).

One thing that has yet to be settled for the HPP is the question of standards. The HUPO plasma proteome project used a unified set of samples to keep all participating researchers on the same page. The difficulty for the HPP, however, is that there is no one cell type, tissue type, or even developmental stage that all researchers can focus on if they want to capture everything. Chromosome 17, for instance, contains a large stretch of olfactory receptor genes, which are expressed only in nasal epithelium; it likely will require collaboration with non-proteomics researchers working on such specialized cell types to pick up proteins like those, says Omenn. Furthermore, the consortium will not be dictating protocols and equipment that must be used to collect the data. HUPO lacks the authority to dictate terms in any event; as Aebersold puts it, there will be no HPP "protocol police."

That said, technical guidelines are coming, according to Legrain, possibly building on data submission policies at journals like *Molecular & Cellular Proteomics*. "In the next three or four months, we hope to have official guidelines for the HPP experiments," he says. The Canadian National Proteomics

Network is already devising guidelines for Canadian researchers, as well as a proposal for a Canadian contribution to the HPP. This effort may eventually influence HPP guidelines as a whole. One possible element of such guidelines, Bergeron says, could be the 20-protein sample used in the HUPO Test Sample Study (1). "We would certainly want everyone who is using a platform and a technology to go through the eye of the needle of a test sample to ensure that when samples have been given to various investigators using each of these technologies, that they are at least agreeing, in the context of the test samples, in terms of what results they're getting."

Of course at the moment, much of this is still conjecture. Ongoing projects like the Human Protein Atlas are moving ahead, but few HPP projects have yet to start. Now, says Legrain, the challenge is to convince funding agencies this is an effort worth supporting. Legrain estimates that the cost of the HPP could total \$1 billion over 10 years. Money has already been kicked in (or at least promised) for pilot projects in Korea and China, and up to €10 million has been set aside by the European Commission for human proteomics. Conversations are ongoing with funding agencies in the United States, Canada, and elsewhere.

Nevertheless, Legrain hopes that by the end of 2011 he will have guidelines locked down, funding mechanisms falling into place, and additional national consortia queuing up for chromosomal projects. Such items will surely be on the agenda when the HPP working group meets again this month in Busan, South Korea.

References

1. **A.W. Bell et al.** 2009. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* 6:423-430.
2. **L.M.F. de Godoy et al.** 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455:1251-1254.
3. **T. Nilsson et al.** 2010. Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nat. Methods* 7:681-685.
4. **HUPO.** 2010. A gene-centric human proteome project. *Mol. Cell Proteomics* 9:427-429.

Written by Jeffrey M. Perkel, Ph.D.

BioTechniques 50:149-155 (March 2011)
doi 10.2144/000113624

To purchase reprints of this article, contact:
carmelitag@fosterprinting.com

The days of comparing apples to apples are over!

Introducing Universal Mycoplasma Detection.

Not all PCR-based kits
are the same.

- Superior sensitivity
- Confirmed specificity
- Broad detection range
- Results in less than 3 hours

See for yourself:

www.atcc.org/MDK1

THE ESSENTIALS OF LIFE SCIENCE RESEARCH
GLOBALLY DELIVERED™