Peptides you can count on

John JM Bergeron & Michael Hallett

Adjusting for proteotypic peptides offers a way forward for quantitative proteomics.

Two papers in this issue describe promising advances in mass spectrometry–based quantitative proteomics. Mallick *et al.*[1] and Lu *et al.*[2] report methods that should extend proteomics beyond merely exhaustively cataloguing the proteins present in a sample to providing quantitative estimates of protein abundance without requiring either the tagging of individual proteins or expensive isotope labeling.

In label-free methods, estimates of absolute abundance are obtained by counting the number of observed peptides. Peptides are identified by assigning the fragment ion patterns generated in the collision cell of a tandem mass spectrometer to amino acid sequences. These peptide sequences are usually determined by *in silico* fragmentation of the sequences contained in protein sequence databases.

However, not all peptides generated by proteolysis have the same likelihood of being detected. Indeed, for any given protein only a few 'proteotypic' peptides are reproducibly identified using a particular proteomic platform. Although proteotypic peptides will likely find many uses, one of the most attractive may be to use the ratio between the number of such proteotypic peptides and the number of peptides observed during a proteomics experiment as an index of protein abundance.

Quantitative proteomics approaches based on peptide counting were originally described by Pang *et al.*[3] and were first applied on a large scale by Blondeau *et al.*[4] The latter demonstrated the feasibility of determining the stoichiometry of stable protein complexes in samples in which the complexes were highly enriched. Using peptide counting of highly enriched clathrin-coated vesicles from rat liver and brain, they showed that the expected 1:1 stoichiometry of clathrin heavy and light chains was not observed in liver, a finding with important functional implications[4,5]. This methodology was subsequently applied to highly complex samples[6,7].

John J.M. Bergeron is in the Department of Anatomy and Cell Biology, 3640 University Street, McGill University, Montreal, Canada and Michael Hallett is in the Center for Bioinformatics, 3375 University Street, McGill University, Montreal, Canada.
e-mail: john.bergeron@mcgill.ca orhallett@mcb.mcgill.ca

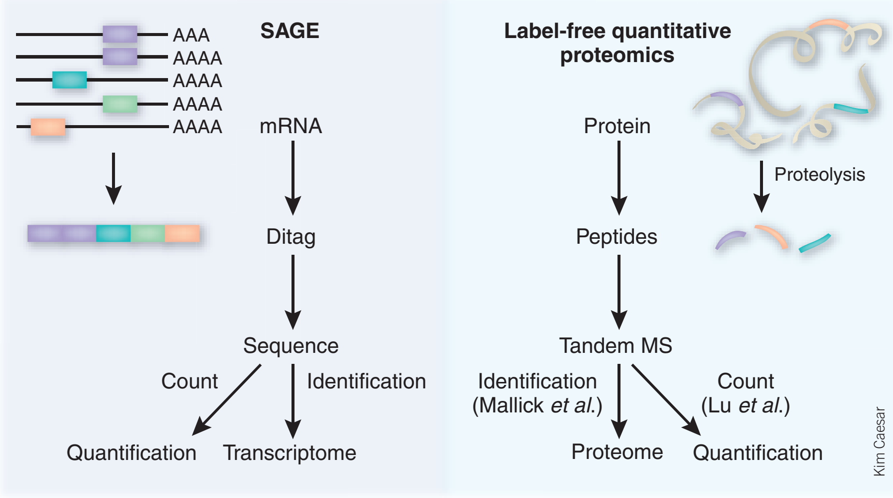

Figure 1 Comparison of the principles of SAGE (serial analysis of gene expression) with label-free quantitative proteomic strategies that do not require internal calibrants. In SAGE, a short sequence tag is sufficient to identify a transcript, quantification of which depends on the number of times a tag is observed. In proteomics, a short proteotypic peptide is sufficient to identify a protein and redundant peptide counting affords quantification.

Extending previous approaches, Lu *et al.* describe Absolute Protein Expression (APEX) profiling, which uses statistical inference and statistics to predict proteotypic peptides and incorporate information on the repeated sampling of spectra from each protein in a shotgun proteomics experiment. The method has strong parallels to the measurement of mRNA levels by serial analysis of gene expression profiling, in which the number of tags sequenced from each mRNA species is a measure of mRNA abundance (**Fig. 1**).

In APEX, the absolute abundance of a protein is estimated from the number of peptides identified by mass spectra derived from the protein, but adjusting for the likelihood that a peptide is proteotypic increases the accuracy of quantitation. This is the first approach that permits comparison between abundance of various peptides obtained from experiments performed with mass spectrometers of different sensitivities, thereby enabling comparisons between different proteomic platforms. A second important aspect of their computation is to correct for the prior probability of observing each of these peptides. Intuitively, proteotypic peptides are likely the most often identified. It follows that the ability to measure protein abundance correctly rests largely upon the ability to estimate how proteotypic each peptide is.

Mallick *et al.* derive a distinct computational approach, also based on statistical inference, to predict proteotypic peptides. Although many proteotypic peptides for widely studied systems have been recognized through repeated identification in empirical studies, computational identification of proteotypic peptides will be particularly advantageous when the level of proteomic data lags far behind genomic sequence information, especially for targeted mass spectrometry experiments where it is desirable to specifically assess presence or absence of specific proteins. Starting with large data sets generated by the commonly applied mass spectrometric methods, Mallick *et al.* used close to 500 physicochemical properties of amino acids (for example, charge, hydrophobicity, prospensity to discriminate between frequently observed peptides and peptides seldom or never observed. A peptide was defined as being proteotypic if it was observed in at least half of the experiments in which the corresponding protein was observed.

Statistical analysis of their results indicated sets of properties for each typical mass spectrometry method that predicted the propensity of a peptide to be detected with a high degree of accuracy. The study therefore confirms the widely held notion that different types of mass spectrometric approaches detect different segments of a proteome and describes physicochemical properties that can be used to score the likelihood of a peptide being observed. Validation of the approach, which was trained using yeast proteins, with a human data set confirmed that the strategy is generally applicable and should enable prediction of proteotypic peptides for any protein, irrespective of the availability of empirical data. However, as the analysis revealed no proteotypic peptides for ~40% of the proteins tested, it is unclear whether proteotypic peptides can be defined for all proteins, at least using current technologies.

Clearly, these approaches open exciting new avenues for analysis and cross-comparisons of quantitative proteomics data that do not require protein tagging or the use of internal calibrants. Working with yeast and *Escherichia coli*, Lu *et al.* measured absolute abundance that span more than three orders of magnitude and showed that they can reliably infer less than twofold differences in protein levels. They also showed that mRNA expression data are a good proxy for protein levels in the majority of cases (and at a log-scale). Furthermore, APEX data allows for the investigation of protein degradation rates and the systematic identification of unusual regulatory events indicated by extreme protein-mRNA expression ratios.

Computational prediction of proteotypic peptides, as described by Mallick *et al.*, will substantially expand the scope of proteomic discovery in species for which the full genomic complement has been characterized, but where limited experimentally-derived proteomic data are available.

Mass spectrometry–based peptide counting approaches that consider the proteotypic propensities of peptides will have far-reaching implications for experimental design. They open new avenues for computational improvements to peptide-identification software, and they may enable more-realistic assessments of the minimal set of peptides that are likely to be observed in mass spectrometry–based experiments and that collectively define a proteome—the proteotypic proteome. With such approaches—and the rapidly increasing sensitivity of mass spectrometry technology[8]—it should not be long before we arrive at the holy grail of proteomics: the discovery of disease biomarkers from patient biofluids.

1. Mallick, P. *et al. Nat. Biotechnol.* **25**, 125–131 (2007).
2. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. *Nat. Biotechnol.* **25**, 117–124 (2007).
3. Pang, J.X., Ginanni, N., Dongre, A.R., Hefta, S.A. & Opitek, G.J. *J. Proteome Res.* **1**, 161–169 (2002).
4. Blondeau, F. *et al. Proc. Natl. Acad. Sci. USA* **101**, 3833–3838 (2004).
5. Girard, M., Allaire, P.D., McPherson, P.S. & Blondeau, F. *Mol. Cell. Proteomics* **4**, 1145–1154 (2005).
6. Ishihama, Y., *et al. Mol. Cell. Proteomics* **4**, 1265–1272 (2005).
7. Gilchrist, A., *et al. Cell* **127**, 1265–1281 (2006).
8. Olsen, J. *et al. Cell* **127**, 635–648 (2006).

A fluid means of stem cell generation

Alan Trounson

Stem cells in amniotic fluid may represent an attractive alternative to embryonic and adult stem cells.

Pluripotentiality—the ability of a cell to form all the cells of the body—is generally considered to be confined to embryonic stem (ES) cells of the preimplantation embryo, embryonal carcinoma cells and embryonic germ cells of the primitive gonad[1]. Rare multipotential or pluripotential stem cells have also been isolated from cultured bone marrow cells[2] and spermatogonial cells of the testis[3]. In this issue, Atala and colleagues[4] describe a stem cell from another source, amniotic fluid, that can be directed into a wide range of cell types representing the three primary embryonic lineages of mesoderm, ectoderm and definitive endoderm (**Fig. 1**) If these results are confirmed by independent laboratories, amniotic fluid–derived stem (AFS) cells may become an important source of cells for regenerative medicine given their apparent advantages of accessibility and multipotentiality over embryonic and adult stem cells, respectively.

Amniotic fluid is known to contain a heterogeneous population of cell types derived from fetal tissues and the amnion. Atala and colleagues captured these cells from amniocentesis samples that were collected for prenatal genetic diagnosis of disease and chromosomal abnormalities, a procedure that is usually performed at 16–20 weeks of pregnancy. Normally, 10–20 ml of fluid is recovered, and the cell sample is divided into a test sample and back-up samples to be used if suitable preparations are not obtained from the original test sample. Using discarded back-up samples, the authors isolated AFS cells by selection for expression of the membrane stem cell factor receptor c-Kit, a common marker for multipotential stem cells.

AFS cells represent about 1% of the cells found in amniotic fluid. After a week of slow proliferation in culture, they adhere to plastic culture flasks and can be passaged at 70% confluence every two to three days. They have a high renewal capacity and can be expanded for over 250 doublings without any detectable loss of chromosomal telomere length. Moreover, the cells are not feeder-cell dependent and should be amenable to bulk culture for research and therapeutic applications.

Atala and colleagues show that AFS cells may be directed into a range of cell types with typical tissue characteristics. AFS cells can become nestin-positive neural stem cells, and then dopaminergic and glutamate-responsive neurons. AFS cell–derived neural stem cells grafted into the lateral cerebral ventricles of twitcher mutant mice dispersed throughout the brain parenchyma, preventing the distortion and neoplasia expected in these animals. In appropriate medium, the AFS cells also form functional osteoblasts that produce bone-like material when embedded in alginate/collagen scaffolds and grafted to immunodeficient mice. Other cell types were obtained, including putative hepatocytes capable of expressing liver proteins such as albumin, α-fetoprotein, hepatocyte nuclear factor and growth factor, and of secreting high levels of urea. These data are strongly indicative of the coordinated function of hepatic molecular pathways.

Although the origin of AFS cells has not yet been elucidated, it is likely that they derive from the amnion, a predictably attractive source of stem cells with multilineage

Alan Trounson is at the Australian Stem Cell Centre and the Monash Immunology and Stem Cell Laboratories, Building 75, Monash University, Clayton, Victoria 3800, Australia. e-mail: alan.trounson@med.monash.edu.au