

Minireview

The subcellular localization of the mammalian proteome comes a fraction closer

Jeremy C Simpson and Rainer Pepperkok

Address: Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Correspondence: Jeremy C Simpson. Email: simpson@embl.de

Published: 23 June 2006

Genome Biology 2006, **7**:222 (doi:10.1186/gb-2006-7-6-222)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2006/7/6/222>

© 2006 BioMed Central Ltd

Abstract

Another step along the road towards determining the subcellular localization of a complete mammalian proteome has been taken with a study using cellular fractionation and protein correlation profiling to identify and localize organellar proteins. Here we discuss this new work in the context of other strategies for large-scale subcellular localization.

The landmark achievements of the complete sequencing of the human and mouse genomes are becoming a distant memory. Their importance has rightly been lauded, but the use of these resources to gain a comprehensive understanding of the human proteome at a functional level has only just started. The identification of all potential open reading frames (ORFs) is doubtless the minimum information required to study the proteome, and is an essential prerequisite to contemporary functional genomics and systems biology approaches. In this context, one logical step towards our understanding of the proteome is the global determination of subcellular protein localization and how it may change, for example, as a result of extracellular stimuli or during development. Despite many parallel and complementary efforts, this goal has still not been achieved for any mammalian proteome.

Tag and tell

On the face of things this may seem somewhat surprising, as the 'localizome' for the budding yeast *Saccharomyces cerevisiae* was reported back in 2003 [1], effectively as a consequence of the availability of the yeast genome sequence. In this elegant work the authors systematically genetically fused the green fluorescent protein (GFP) with 97% of the organism's ORFs, then used fluorescence microscopy to classify the locations of the tagged proteins. An important aspect

of this study was that the proteins were expressed from their endogenous promoters, thereby providing additional confidence in the results.

Such tagging and visualization approaches are undoubtedly powerful and have already been applied to a wide range of organisms, including mammals (reviewed in [2-4]), but they also have limitations. The tag may interfere with correct protein localization, and this can occur regardless of whether the tag is a whole protein (for example, GFP) or a small epitope (for example, the Myc epitope). But although this is true for some proteins, the direct visualization of each and every protein in a living cell is clearly a legitimate goal. What then are the alternatives? One possibility is the systematic generation of antibodies against the entire proteome and their use in immunofluorescence localization methods. Although this approach uses fixed rather than living cells, and can also suffer from the dangers of mislocalization, this time by antibodies recognizing similar or overlapping epitopes, the visualization of endogenous proteins at 'normal' expression levels is an exciting prospect. A pioneering effort in this respect is the recent work by Mathias Uhlen and colleagues [5], who have generated and tested more than 700 antibodies against human proteins. In this study, the protein localization information is mainly obtained at the tissue level by immunohistochemistry, but the antibodies could readily be used for immunofluorescence analysis at the subcellular level.

Divide and identify

A quite different approach towards proteome localization uses cellular fractionation followed by mass spectrometry (MS) to identify the protein composition of the fractions. This is the strategy used in work recently published in *Cell* by Matthias Mann and colleagues [6], which attempts to create a 'mammalian organelle map' using mouse liver cells. This general approach has become possible as a result of significant advances in MS-based organelle proteomics, an area that has recently seen a huge increase in activity. Projects to isolate the Golgi complex, clathrin-coated vesicles, and mitochondria, among many other organelles, followed by MS and protein identification, have yielded impressive lists of proteins associated with these cellular structures (reviewed in [7]). In its simplest form, however, this approach requires purification of the organelle of interest to a high degree of homogeneity from the remainder of the cellular content. In general, the greater the number of biochemical separation steps used, the higher the purity, but this comes at the expense of loss of valuable material. Organelle proteomics of this type also isolates the organelle from its cellular context, and at best can only provide a snapshot of the resident proteins at any particular point in time. Proteins transiently associated with the organelle, for example those involved in inter-organelle communication, are therefore most likely to be missed by such approaches.

In the recent study in *Cell* by Foster *et al.* [6], Mann and his group have sought to avoid some of these problems by using protein correlation profiling to study multiple organelles simultaneously. This technique is described in earlier work from the same group that identified novel centrosomal components [8]. In that study, they disrupted cells by biochemical techniques, obtained a crude centrosome preparation, and then subjected this to gradient centrifugation. The fractions obtained were digested with protease and the resulting peptides analyzed by MS. The abundance of each peptide in every fraction was determined, and the abundances were then compared to abundance profiles of peptides from well known resident centrosomal proteins. The correlation between such profiles could then be used to indicate the likelihood that the unknown protein is localized to the centrosome, and the likely deviation expressed as a χ^2 value. In total, 23 novel centrosomal proteins were identified by this technique, and their localization was validated by GFP tagging and microscopic analysis. One major advantage of protein correlation profiling over the organellar fractionation techniques noted above is that it can potentially be applied to crude cell extracts, and data can be obtained from organelles that are difficult to purify to homogeneity biochemically. Furthermore, protein correlation profiling analyses proteins expressed at endogenous levels, it does not require antibodies, and it can be applied at either the cellular or the tissue level.

The new work by Foster *et al.* [6] applied this profiling approach to whole mouse liver, and created reference

peptide profiles for ten organelles or subcellular structures, including the endoplasmic reticulum, Golgi complex, different classes of endosomes and proteasomes. Analysis of continuous sucrose gradients resulted in the identification of over 22,000 peptides, corresponding to 2,200 proteins, of which 1,400 were localized with a high degree of confidence. Comparison of these results with non-proteomic-based localization annotations in the UniProt and Gene Ontology (GO) databases indicated a remarkable accuracy of 87%. In addition, Foster *et al.* [6] extended their analysis to include mRNA expression data from 44 mouse tissues, which revealed subsets of coexpressed organellar genes.

One of the more striking results from this work is the large number of proteins that appear to localize to more than a single organelle (for example, almost 40% of the proteins identified as belonging to either the cytoplasm or the proteasome were also found in other fractions). Although not entirely unexpected, this is a very important observation, and one that would inevitably be missed by single-organelle proteomics strategies. The problem is, of course, to dissect out those proteins that truly localize to multiple compartments from those that show such a pattern as a result of limitations in the experimental procedure. The separation of certain organelles, for example those that migrate at similar densities in a sucrose density gradient, suffers from the technical restrictions of the fractionation procedure, and indeed Foster *et al.* [6] observed this effect in some of their results. Critically, the success of the biochemical fractionation approach relies on proteins remaining stably associated with their *bona fide* organelle of residence during isolation. For example, the Rab family of small GTPases comprises more than 60 closely related proteins that are central regulators of membrane traffic, each of which is highly specifically localized to particular membranes (reviewed in [9]). As such, they are believed to be one important determinant of organelle identity and therefore function. Of the 14 Rab proteins localized by the protein correlation profiling analysis of Foster *et al.* [6], eight were reported to be at least partially present in the plasma membrane fraction, despite the fact that the majority of these have been reported to be present only on internal organelles. Careful interpretation of these data and their complementation by other methods is therefore important.

Correctly defining the localization of some other classes of proteins by protein correlation profiling analysis is also likely to be somewhat problematic. These include cytoskeletal proteins, peripheral membrane proteins, and proteins that only transiently interact with membranes. Cytoskeletal elements and their regulatory factors are not permanently associated with organelles, but help to define their identity. Although the profiling study of Foster *et al.* [6] correctly identified many actin and tubulin subunits in the soluble cytosolic fraction, this reveals little about their true function as major structural components of the cell, or their crucial and dynamic interaction with organelle membranes.

A surprising aspect of the work by Foster *et al.* [6] is the relatively small number of proteins positively identified as associated with organelles. Clearly this work was an enormous undertaking, but it has resulted in experimentally determined localization information for probably less than 10% of the proteome. Despite the potential of protein correlation profiling, the impressive recent improvements in MS and peptide identification, and their application at the tissue level, the weakest link in this study is the reliance on the initial steps of traditional subcellular fractionation and gradient centrifugation. These limitations will require further refinement if protein correlation profiling is to be the methodology of choice for global subcellular localization analysis of complex mammalian proteomes.

A question of cellular complexity

This approach nevertheless takes us another step closer to the subcellular localization of the complete mammalian proteome. Perhaps we should ask why this task is still not complete, considering the many noteworthy efforts that are under way. One answer could be the great size and complexity of mammalian genomes, but we rather favor the explanation that it is more a problem of biology, not simply of numbers. In higher eukaryotic cells, compartmentalization is an essential feature that enables the sequestering of specific biochemical reactions to a defined environment. Compartmentalization is predominantly achieved through membrane-bounded organelles, although it can occur through highly localized concentration of proteins (at the centrosome, for example). In particular, in mammalian cells, the special reorganization of organelles coupled with their more specialised roles in different cells types, adds additional complexity to protein localization. Furthermore, in living cells these compartments are not static; rather, the interchange of small molecules, lipids and proteins between them is essential to preserve their functionality. Organelle constituents may be structural or dynamic, and can be distributed evenly throughout the entire organelle or only be present in concentration gradients or local hot spots. The resulting distinct physical and biochemical properties of the proteins involved mean that the technique used to study them must preserve them and their equilibrium as much as possible. A single methodology is unlikely to achieve this.

Bioinformatic tools continue to play a role in this quest (reviewed in [10]), and are helpful in supporting and extending large-scale experimental datasets. In addition, comprehensive data mining needs to be used more, so that all published localization information is collated: the LOCATE database for the mouse proteome is a good example [11]. As the results of Foster *et al.* [6] show, no one approach can be completely successful, and it will only be through the combination of different large-scale subcellular identification methodologies that the complete organelle map will be drawn.

Acknowledgements

We would like to acknowledge funding by the Federal Ministry of Education and Research (BMBF) in the framework of the National Genome Research Network (NGFN-2 SMP-Cell FKZ01GR0423).

References

1. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425**:686-691.
2. Pepperkok R, Simpson JC, Wiemann S: **Being in the right location at the right time.** *Genome Biol* 2001, **2**:reviews1024.
3. Simpson JC, Pepperkok R: **Localizing the proteome.** *Genome Biol* 2003, **4**:240.
4. O'Rourke NA, Meyer T, Chandy G: **Protein localization studies in the age of 'omics'.** *Curr Opin Chem Biol* 2005, **9**:82-87.
5. Uhlen M, Bjorling E, Agaton C, Szigartyo CA, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C, et al.: **A human protein atlas for normal and cancer tissues based on antibody proteomics.** *Mol Cell Proteomics* 2005, **4**:1920-1932.
6. Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, Mann M: **A mammalian organelle map by protein correlation profiling.** *Cell* 2006, **125**:187-199.
7. Yates JR III, Gilchrist A, Howell KE, Bergeron JJM: **Proteomics of organelles and large cellular structures.** *Nat Rev Mol Cell Biol* 2005, **6**:702-714.
8. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M: **Proteomic characterization of the human centrosome by protein correlation profiling.** *Nature* 2003, **426**:570-574.
9. Zerial M, McBride H: **Rab proteins as membrane organizers.** *Nat Rev Mol Cell Biol* 2001, **2**:107-117.
10. Donnes P, Hoglund A: **Predicting protein subcellular localization: past, present, and future.** *Genomics Proteomics Bioinformatics* 2004, **2**:209-215.
11. Fink JL, Aturaliya RN, Davis MJ, Zhang F, Hanson K, Teasdale MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD: **LOCATE: a mouse protein subcellular localization database.** *Nucleic Acids Res* 2006, **34(Database issue)**:D213-D217.