



ELSEVIER

Protein localization in proteomics

Trisha N Davis

A global analysis of the localization of 4156 yeast proteins has just been accomplished. Smaller scale analyses have been performed in a variety of organisms. These studies typically use green fluorescent protein as a tag for proteins in living cells. Improvements in the yellow and sapphire color variants will increase their utility. Reengineering of the red fluorescent protein has produced faster maturing tetrameric and monomeric variants not prone to aggregation. Techniques for high-throughput tagging of proteins include integration by homologous recombination, integration using mobile elements or recombinational cloning to produce plasmids expressing fusion proteins. Alternatives to localizing tagged proteins are to use antibodies or aptamers to detect the untagged protein.

Addresses

Department of Biochemistry, University of Washington, Box 357350, Seattle, WA 98195-7350, USA
e-mail: tdavis@u.washington.edu

Current Opinion in Chemical Biology 2004, 8:49–53

This review comes from a themed issue on
Proteomics and genomics
Edited by Michael Snyder and John Yates III

1367-5931/\$ – see front matter
© 2003 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.cbpa.2003.11.003

Abbreviations

CFP	cyan fluorescent protein
DsRed	red fluorescent protein from coral species <i>Discosoma</i>
GFP	green fluorescent protein
mRFP1	monomeric form of DsRed
ORF	open reading frame
YFP	yellow fluorescent protein

Introduction

Localization of proteins by light microscopy offers unique advantages for exploring a proteome. Under ideal circumstances, localization reveals not only where a protein is found but also when it is found there. Dynamic movements from one location to another can be followed as the cell proceeds through the cell cycle, or responds to environmental stresses or internal signaling pathways. Of the many techniques applied to study proteins, microscopy alone can view the whole intact cell. This global perspective can confirm whether localization of a protein *in vivo* is physically compatible with the information drawn from other proteomic methods [1^{••},2^{••}]. This article focuses on the use of wide field fluorescence microscopy to examine the localization of proteins in cells with an emphasis on large-scale or global analyses.

Recent improvements in fluorescent proteins

The discovery of green fluorescent protein (GFP) from *Aequorea victoria* has revolutionized the localization of proteins by allowing straightforward examination of proteins in living cells. Lippincott-Schwartz and Patterson [3[•]] give an excellent overview of GFP and the colored variants of GFP, including yellow fluorescent protein (YFP) and cyan fluorescent protein (CFP). Here, I discuss only the most recent developments. YFP and CFP is an excellent pair for comparing the localization of two proteins because they can be readily distinguished by standard filter sets. In yeast cells, YFP and CFP have approximately the same brightness as GFP but both bleach more rapidly [4]. One of the disadvantages of the original YFP (created by three mutations to GFP: S65G, S72A, T203Y) was its thermolability and sensitivity to pH and halide concentrations. Mutations Q69K and V68L reduced the pK_a such that the pH sensitivity of the fluorescence was less at the neutral pH found under physiological conditions [5]. The mutation Q69M found in the citrine version of YFP further reduced pH and chloride sensitivity, decreased the rate of photobleaching and improved the ability to fold at 37°C [6[•]]. Nagai and coworkers [7[•]] introduced the common folding mutations F64L, M153T, V163A and S175G into the original YFP and found that these also enhanced folding and decreased environmental sensitivity but did not decrease the rate of photobleaching. The mutation F46L further improved maturation at 37°C. YFP with the four folding mutations and F46L is called Venus, which provides about 10% greater brightness at 30°C than YFP (S65G, Q69K, V68L, S72A and T203Y) [4].

The sapphire form of GFP has an excitation peak at 399 nm and emission peak at 511 nm. Sapphire has had limited utility because of slow folding and maturation. Zapata-Hommer and Griesbeck [8] developed Turbo-Sapphire (T-Sapphire) with the mutations Q69M, C70V, V163A and S175G, which dramatically improved the rate of folding. With a pK_a of emission at 4.9, T-Sapphire is insensitive to pH changes in the physiological range. The rate of photobleaching was not reported.

The list of available colors was expanded with the discovery of the red fluorescent protein from the sea coral of the genus *Discosoma* (DsRed). DsRed has a primary emission peak at 583 nm and is naturally insensitive to pH changes in the physiological range [9]. The utility of DsRed has been diminished because it is a stable tetramer and, unlike GFP, tends to cause aggregation of tagged proteins. Moreover, it matures slowly with a half time of maturation in *Escherichia coli* of 11 h at 37°C and

longer at lower temperatures [10^{••},11]. Bevis and Glick [10^{••}] developed variants of DsRed that mature 15-fold faster and are less prone to aggregation. We have found the DsRedT1 version works well as a tag for proteins in yeast spindle pole bodies (SPB), although the rate of folding is still slow enough that the newly made SPB is not as bright as the old SPB [12]. DsRedT1 is 2.2-fold brighter than GFP [4].

Extensive work by Campbell and co-workers [13^{••}] has overcome the obligate tetramerization of DsRed. They developed two versions, a monomer mRFP1 and a covalent tandem dimer (tdimer2(12)). The monomer has a lower extinction coefficient ($44\,000\text{ M}^{-1}\text{cm}^{-1}$ compared with $57\,000\text{ M}^{-1}\text{cm}^{-1}$), decreased quantum yield (0.25 compared with 0.79 for the original) and bleaches 30-fold faster than the original DsRed. However, mRFP1 matures 10-fold faster and so shows similar brightness in living cells. Alternative monomeric DsRed variants that may avoid the limitations of mRFP1 are being developed (B Glick, personal communication). The tdimer2 displays the more favorable quantum yield and bleaching kinetics of the original. It matures fivefold faster and has twice the extinction coefficient, which is expected because it contains two chromophores per polypeptide chain. The version tdimer2 offers an excellent choice for many applications, especially in slower growing cells, where the half time for maturation of 2 h would not be a significant factor.

Several corals and jellyfish produce fluorescent proteins and these are being developed into tools useful for research and development. A GFP from the stony coral *Galaxeidae* has a similar spectral profile to *A. victoria* GFP and has a low pK_a of fluorescence. However, like DsRed, it is a tetramer. Karasawa and co-workers [14] developed a monomeric form that shows promise, although a detailed measurement of its maturation rate was not presented.

Global analyses of protein localization

Fusions to fluorescent proteins or epitope tags

Large-scale analyses of protein localization using fusions to fluorescent proteins or epitope tags have been performed in budding yeast, fission yeast, *Drosophila*, tobacco, and Vero cells (a cell line derived from the kidney of the African green monkey) using human cDNAs [1^{••},2^{••},15[•],16–18,19^{••},20]. The most comprehensive is the construction of a collection of yeast strains expressing full-length chromosomally tagged GFP fusion proteins [2^{••}]. Analysis of this collection allowed localization of 75% of the yeast proteome. The genes were tagged by integration of the GFP cDNA at the 3'-end of each gene in its normal chromosomal locus and expressed under the control of each gene's endogenous promoter. Thus, localization artifacts due to overexpression were avoided. The only artifacts are those caused by tagging

the protein. This effect is likely to be minor because proteins tagged with GFP generally behave as the untagged protein. The obvious exceptions are those proteins whose localization depends on modification of the C-terminus (such as palmitoylation or farnesylation). Microscopic imaging was performed on live cells adhered to 96-well glass-bottom microscope plates and image capture was automated. The quality of the images is very good although not equivalent to those taken on single slides in smaller-scale studies. Image analysis was performed by two independent scorers, who classified the localization of each protein into one of 12 initial categories. These localizations were refined to include 11 additional categories. The fact that the data shows 80% agreement with previously reported localizations in the *Saccharomyces* Genome Database and 90% agreement with the localizations presented in a study of a smaller scale [1^{••}] indicates it is a rich and high quality dataset.

Can this be repeated for mammalian or other higher eukaryotic proteomes? Not yet, although recent developments suggest several strategies that might be feasible. One significant obstacle is that large-scale production of a collection of cell lines each with a defined gene chromosomally tagged at the 3'-end is not yet possible. The hurdles to be overcome include identification of the open reading frames (ORFs), identification of the alternatively spliced forms, and the cost of making cell lines in which a gene is tagged by homologous recombination. These hurdles did not exist for the global analysis of protein localization in yeast because splicing of RNAs is rare in *S. cerevisiae*, and virtually all integration into the genome occurs by homologous recombination.

An alternative strategy for chromosomally tagging genes uses protein tags inserted in mobile elements that randomly insert into the genome. The site of insertion is determined after the fact by sequencing using primers complementary to the tag. This technique was demonstrated in a large-scale analysis in yeast by the pioneering work of Ross-Macdonald and coworkers [21] and further refined by Kumar and co-workers [15[•]] to localize 1083 yeast proteins. A 'protein trap' strategy using GFP as a mobile artificial exon was used to create 600 independent lines in *Drosophila* [18]. Similar technology has also been developed for mammalian cells using a retroviral vector to introduce epitope tags or GFP into genes and has been termed CD-tagging [22]. CD-tagging was used to create several hundred cell clones each with a GFP tag inserted into the genome. A key feature of the protein trap strategy and CD-tagging is that the GFP is inserted as its own exon complete with signals for splicing. The tag is likely to be inserted into the middle of a protein, but has no stop codon and so translation of the entire protein can continue. No prior knowledge of the genome structure or definition of the ORFs is required, although the genome sequence is important for later identification of the

insertion sites. Presumably, multiple alternatively spliced forms of a protein will be tagged by a single event.

Another approach is to make a library of genomic fragments or cDNAs fused to GFP or epitope tags as has been done in fission yeast, budding yeast, human and plants [15[•],16,17,19^{••},20]. Recombinational cloning systems such as the Gateway system make large-scale cloning projects feasible. For fungal systems, where the ORFs are readily defined and introns rare, entire ORFs are amplified from the genome [15[•]]. For other organisms, cDNAs can be made by random priming of mRNA, which can lead to fragments of proteins being labeled [19^{••}] or a large-scale identification of full length cDNAs is required [20]. Expression from plasmids often results in production of large amounts of protein. Simpson and co-workers [20] imaged cells at multiple time points after transfection to identify any effects of the increasing expression levels.

Antibodies and aptamers

Technologies that detect a protein itself rather than a fusion protein prevent mislocalizations due to the presence of the tag or abnormal expression levels of the fusion proteins. Antibodies detect proteins directly but are difficult to make in large scale by standard methods. Two problems to overcome are the production of the antigen and the production of the antibody. Although not for this reason, large-scale antigen production is already being performed by the structural genomics centers funded by the protein structure initiative (<http://www.nigms.nih.gov/psi/>). The goal of these centers is to express, purify, crystallize and solve the structures of many of the proteins predicted from genomic sequences from a variety of organisms, including human. Although it is not clear what fraction of the proteins will meet the stringent criteria required to obtain structures, many more will meet the much less stringent criteria required for antigens, basically the ability to be purified not even necessarily in a soluble form. A more formal connection between the structural genomics centers and laboratories interested in global analyses of protein localization is desirable.

Phage display or other display technologies offer the best opportunity for high-throughput production of antibodies but require some modifications to be adapted for protein localization (see excellent recent reviews [23[•],24]). The display technologies can also be used to select antibodies by intracellular genetic selection. Purification of antigen is not required. Instead, both the antibodies and the antigen are expressed in cells and confer a growth phenotype if interaction occurs [23[•]].

Aptamers offer an alternative to antibodies for high affinity probes. Aptamers are short single-stranded DNA or RNA sequences selected *in vitro*. A recent improvement in the selection method used magnetic bead-bound His-

tagged recombinant protein to select DNA aptamers with K_d values for their target in the nanomolar range [25[•]]. Combining this technique with the large-scale recombinant expression of ORFs from *C. elegans* [26], could produce a very useful set of high affinity probes. The use of aptamers as probes in protein localization is in its infancy, but recent work shows promise for localization in mammalian cells [27].

Microscopy and image analysis

Global analyses of protein localization require high-throughput microscopy and image analysis, with the former being simpler to attain. Huh and co-workers [2^{••}] automated image acquisition on 96-well slides using a script in Metamorph imaging software (Universal Imaging Corporation, <http://www.universal-imaging.com/index.cfm>). The microarray-driven gene expression system described for mammalian cells would enable the quick analysis of a collection of tagged cDNAs [28]. In this system the cDNAs are spotted on a microarray and then mammalian cells are plated on the slide. Each spot on the microarray contains a collection of cells expressing a given cDNA. In plants, leaves are inoculated with a collection of tobacco mosaic viruses expressing GFP fusion proteins. Multiple lesions form on each plant and each lesion represents expression of a single GFP fusion, so several hundred GFP-fusions can be screened daily [19^{••}].

The greater challenge is high-throughput image analysis. The yeast images produced by Huh and co-workers were analyzed by humans [2^{••}]. This is the best method, but is tedious even for proteomes the size of yeast. There is great interest in image analysis for clinical diagnostics. Applying these specialized image analysis systems to the identification of the multiple subcellular patterns will be challenging. The high-throughput fluorescence imaging and analysis methods being developed to screen chemical libraries [29^{••}] are likely to be adaptable to the more complex problem of automated protein localization.

Conclusions

To increase the likelihood that the localization of a tagged protein represents the normal distribution of the protein several criteria should be met. To minimize artifacts from abnormal protein abundance the fusion protein should be expressed from its normal chromosomal locus under control of its native promoter. The fusion should replace the native protein to avoid competition for binding partners and to establish that the fusion is functional.

Huh and co-workers [2^{••}] satisfied these conditions in their impressive study examining the localization of proteins in budding yeast. However, the production of similar large-scale collections will be challenging in systems less amenable to genetic manipulation. Several alternatives exist. Mobile elements can be used to insert artificial

exons encoding protein tags randomly in the genome. This strategy is less likely to produce artifacts due to abnormal expression levels; however, multiple splice variants are likely to be tagged by a single event. If a collection of cDNAs is available then a library of fusions can be made by recombinational cloning. Although expression is often abnormal from these constructs, the localization can still be informative if monitored shortly after transfection before overexpression occurs. Neither of these strategies results in production of strains where the tagged gene is the only copy of the gene, which is especially difficult in obligate diploids. Analysis of the significance of this caveat would be useful. Antibodies or aptamers identify the untagged protein expressed at normal levels. Adaptation of these technologies for large-scale protein localization would be fruitful. Given the importance of understanding protein localization for understanding function, high quality localization datasets for multiple organisms will be important for a full description of each proteome.

Acknowledgement

I thank Eric Muller for helpful discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Hazbun TR, Malmström L, Anderson S, Graczyk BJ, Fox B, Riffle M, •• Sundin BA, Aranda JD, McDonald WH, Chun CH *et al.*: **Assigning function to yeast proteins by integration of technologies.** *Mol Cell* 2003, in press.
Characterization of essential uncharacterized yeast ORFs integrating protein localization with three other proteomic technologies, mass spectrometric identification of co-purifying proteins, two-hybrid analysis and protein structure prediction.
 2. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, •• Weissmen JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425**:686-691.
Localization of 75% of the yeast proteome using full length chromosomally tagged GFP fusion proteins. This is the most comprehensive large-scale analysis of protein localization ever performed.
 3. Lippincott-Schwartz J, Patterson GH: **Development and use of fluorescent protein markers in living cells.** *Science* 2003, **300**:87-91.
Excellent review of fluorescent proteins.
 4. Muller EGD, Davis TN: **Protein localization by cell imaging.** In *Proteomics for Biological Discovery*. Edited by Veenstra TD, Yates JRI: Wiley and Sons; 2004:in press.
 5. Miyawaki A, Griesbeck O, Heim R, Tsien RY: **Dynamic and quantitative Ca²⁺ measurements using improved cameleons.** *Proc Natl Acad Sci USA* 1999, **96**:2135-2140.
 6. Griesbeck O, Baird GS, Campbell RE, Zacharias DA, Tsien RY: • **Reducing the environmental sensitivity of yellow fluorescent protein. Mechanism and applications.** *J Biol Chem* 2001, **276**:29188-29194.
Simple mutation of Q69M in YFP has many beneficial effects including reduced environmental sensitivity, increased rate of folding at 37°C and reduced rate of photobleaching.
 7. Nagai T, Ibata K, Park ES, Kubota M, Mikoshiba K, Miyawaki A: • **A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications.** *Nat Biotechnol* 2002, **20**:87-90.
Introduction of folding mutations into YFP increases rate of maturation and decreases environmental sensitivity.
 8. Zapata-Hommer O, Griesbeck O: **Efficiently folding and circularly permuted variants of the Sapphire mutant of GFP.** *BMC Biotechnol* 2003, **3**:5.
 9. Baird GS, Zacharias DA, Tsien RY: **Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral.** *Proc Natl Acad Sci USA* 2000, **97**:11984-11989.
 10. Bevis BJ, Glick BS: **Rapidly maturing variants of the Discosoma •• red fluorescent protein (DsRed).** *Nat Biotechnol* 2002, **20**:83-87.
Development of a rapidly maturing DsRed.
 11. Mizuno H, Sawano A, Eli P, Hama H, Miyawaki A: **Red fluorescent protein from Discosoma as a fusion tag and a partner for fluorescence resonance energy transfer.** *Biochemistry* 2001, **40**:2502-2510.
 12. Yoder TJ, Pearson CG, Bloom K, Davis TN: **The Saccharomyces cerevisiae spindle pole body is a dynamic structure.** *Mol Biol Cell* 2003, **14**:3494-3505.
 13. Campbell RE, Tour O, Palmer AE, Steinbach PA, Baird GS, •• Zacharias DA, Tsien RY: **A monomeric red fluorescent protein.** *Proc Natl Acad Sci USA* 2002, **99**:7877-7882.
Development of a rapidly maturing monomeric version of DsRed and of a tandem covalent dimer.
 14. Karasawa S, Araki T, Yamamoto-Hino M, Miyawaki A: **A green-emitting fluorescent protein from Galaxeidae coral and its monomeric version for use in fluorescent labeling.** *J Biol Chem* 2003, **278**:34167-34171.
 15. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, • Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y *et al.*: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16**:707-719.
Large-scale immunolocalization of 1083 proteins epitope tagged by transposon insertional mutagenesis and 2022 epitope-tagged proteins expressed as full length ORFs cloned into an expression vector.
 16. Ding DQ, Tomita Y, Yamamoto A, Chikashige Y, Haraguchi T, Hiraoka Y: **Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library.** *Genes Cells* 2000, **5**:169-190.
 17. Sawin KE, Nurse P: **Identification of fission yeast nuclear markers using random polypeptide fusions with green fluorescent protein.** *Proc Natl Acad Sci USA* 1996, **93**:15146-15151.
 18. Morin X, Daneman R, Zavortink M, Chia W: **A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in Drosophila.** *Proc Natl Acad Sci USA* 2001, **98**:15050-15055.
 19. Escobar NM, Haupt S, Thow G, Boevink P, Chapman S, Oparika K: •• **High-throughput viral expression of cDNA-green fluorescent protein fusions reveals novel subcellular addresses and identifies unique proteins that interact with plasmodesmata.** *Plant Cell* 2003, **15**:1507-1523.
High-throughput analysis of protein localization in tobacco using tobacco mosaic virus expressing GFP fusion proteins.
 20. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S: **Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing.** *EMBO Rep* 2000, **1**:287-292.
 21. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symoniatis D, Umansky L *et al.*: **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature* 1999, **402**:413-418.
 22. Jarvik JW, Fisher GW, Shi C, Hennen L, Hauser C, Adler S, Berget PB: **In vivo functional proteomics: Mamm genome annotation using CD-tagging.** *BioTechniques* 2002, **33**:852-854,856,858-860.
 23. Bradbury A, Velappan N, Verzillo V, Ovecka M, Chasteen L, • Sblattero D, Marzari R, Lou J, Siegel R, Pavlik P: **Antibodies in proteomics I: generating antibodies.** *Trends Biotechnol* 2003, **21**:275-281.
Outstanding review of the methods with potential for high-throughput generation of antibodies.
 24. Bradbury A, Velappan N, Verzillo V, Ovecka M, Chasteen L, Sblattero D, Marzari R, Lou J, Siegel R, Pavlik P: **Antibodies in**

proteomics II: screening, high-throughput characterization and downstream applications. *Trends Biotechnol* 2003, **21**:312-317.

25. Murphy MB, Fuller ST, Richardson PM, Doyle SA: **An improved method for the *in vitro* evolution of aptamers and applications in protein detection and purification.** *Nucleic Acids Res* 2003, **31**:e110.

Describes a high-throughput method for production of aptamers, small single-stranded RNA or DNA molecules specific for binding proteins. If combined with the high-throughput expression of ORFs as described for *C. elegans* ORFs in Huang *et al.*, it could result in the large-scale production of specific high-affinity probes with many applications.

26. Huang RY, Boulton SJ, Vidal M, Almo SC, Bresnick AR, Chance MR: **High-throughput expression, purification, and**

characterization of recombinant *Caenorhabditis elegans* proteins. *Biochem Biophys Res Commun* 2003, **307**:928-934.

27. Stanlis KK, McIntosh JR: **Single-strand DNA aptamers as probes for protein localization in cells.** *J Histochem Cytochem* 2003, **51**:797-808.
28. Ziauddin J, Sabatini DM: **Microarrays of cells expressing defined cDNAs.** *Nature* 2001, **411**:107-110.
29. Yarrow JC, Feng Y, Perlman ZE, Kirchhausen T, Mitchison TJ: **Phenotypic screening of small molecule libraries by high throughput cell imaging.** *Comb Chem High Throughput Screen* 2003, **6**:279-286.

Development of high-throughput methods for microscopy and image analysis.