

---

## AWARD ADDRESS

# 2004 Irving Sigal Young Investigator Award

---

JONATHAN S. WEISSMAN<sup>1</sup> AND ERIN K. O'SHEA<sup>2</sup>

<sup>1</sup>Department of Cellular and Molecular Pharmacology and <sup>2</sup>Department of Biochemistry and Biophysics, Howard Hughes Medical Institute (HHMI), University of California-San Francisco (UCSF), San Francisco, California 94143-2240, USA

In recognition of the fact that we are corecipients of the 2004 Irving Sigal Young Investigator Award from The Protein Society, we have chosen to write jointly about the work we have done together in the area of proteomics.

We first met in Peter Kim's lab at the Whitehead Institute for Biomedical Research and the Massachusetts Institute of Technology (MIT) nearly 15 years ago when we were both graduate students. Like many other people in the Kim lab at that time, including Peter himself, we both had strong backgrounds in the physical sciences—E.K.O. was a graduate student in Chemistry and J.S.W. was a graduate student in the Physics Department at MIT—and were attracted to the Kim lab with the possibly naive hope that we would be able to combine quantitative analytical approaches with experiments to address problems of immediate biological importance. Both of us worked on projects involving the study of protein structure, function, and folding: E.K.O. investigated the structure and folding of a dimerization motif known as the “leucine zipper” and J.S.W. investigated the folding pathway of bovine pancreatic trypsin inhibitor (BPTI). In addition to providing a superb place to obtain a classical training in the rigors and craftsmanship of protein biochemistry, the Kim lab, by virtue of the fact that it was situated at the Whitehead, also exposed us to many of the most exciting advances in cell and developmental biology. It was this combination that has allowed us, and a remarkable proportion of our contemporaries in the Kim lab, to succeed in our future careers.

After leaving the Whitehead, E.K.O. pursued a brief postdoc in the laboratory of Robert Tjian at the University of California, Berkeley, and Ira Herskowitz at UCSF, and J.S.W. worked with Arthur Horwich at Yale University be-

fore both of us joined the faculty at UCSF. Ostensibly our specific research interests diverged: E.K.O. worked on transcription and signal transduction, and J.S.W. worked on chaperonin-mediated protein folding and prion-based inheritance. However, we both retained the common goal of going beyond simple phenomenological models to explain complex cellular behaviors. We also both naturally gravitated to using the common bakers yeast *Saccharomyces cerevisiae* as a model organism because its simplicity and technical advantages provided the most hopeful platform for such approaches, especially after the genome sequence of *S. cerevisiae* was completed and the introduction of microarray technology made it possible to readily measure the abundance of all of the messages in a cell.

Being at UCSF in the mid- to late 1990s, we were near the epicenter of the explosion of microarray technology that made it possible to probe the internal state of a cell with a precision that was previously unimaginable. Both of our labs extensively integrated such measurements in our own studies. That said, we were also acutely aware that such measurements of mRNA levels were an imperfect proxy for the information we really wanted, the abundance and activity of the proteins that were responsible for carrying out the large majority of cellular activities. Beyond the well-documented inability of the mRNA changes to fully account for changes in protein levels, many of our own studies focused on protein regulation involving changes in subcellular localization and covalent modifications such as phosphorylation—information inaccessible by microarray studies. It was the inability of microarrays to track such changes at the protein level that had been, and continues to be, the driving force behind efforts to develop robust proteomics-based approaches. There have been remarkable technical advances, especially in the area of mass spectrometry, toward such efforts. Nonetheless, it was also evident that despite such progress, even for a simple organism such yeast, the long-term challenge of proteomics, to define the identities, quantities, structures, and functions of complete complements of proteins, and to characterize how these properties vary in different cellular contexts, remained far out of the reach of

---

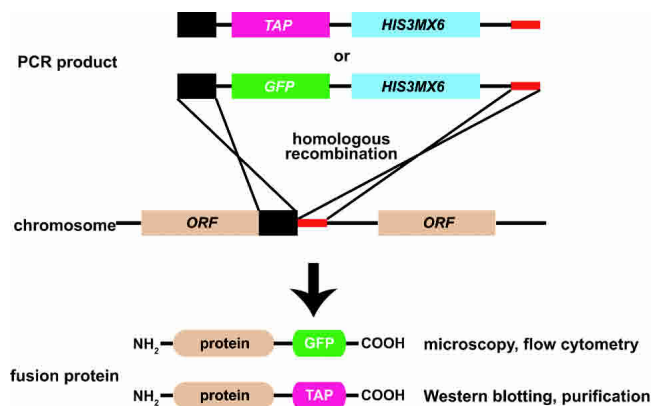
Reprint requests to: Jonathan S. Weissman, Howard Hughes Medical Institute, Department of Cellular and Molecular Pharmacology, University of California-San Francisco, Genentech Hall S472C, 600 16th Street, San Francisco, CA 94143-2240, USA; e-mail: weissman@cmp.ucsf.edu; fax: (415) 514-2073.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.041134604>.

present technologies. At the same time it was also clear from a broad range of individual examples that a rather mature and simple approach—the use of epitope tags to provide a common physical handle on a protein—offered a robust, highly sensitive way of exploring the abundance, localization and function of even poorly expressed proteins. Moreover, two seminal studies had illustrated the potential power of expanding this strategy to the whole yeast proteome: Michael Snyder's group had used transposons to make random fusions between yeast proteins and a reporter protein such as  $\beta$ -galactosidase or green fluorescent protein (GFP), and Eric Phizicky's group had made a nearly complete library allowing overexpression of the yeast proteins as N-terminal fusions to glutathione-S-transferase (GST). As important and innovative as these studies were, we also felt that they did not fully exploit the power of the yeast system which made it possible to express full length epitope-tagged proteins from their endogenous chromosomal location at natural levels.

The task of constructing such strains was seemingly a large one with which we had no experience and would not have thought to consider except for the fact that in the summer of 2000 we both had the good fortune to be chosen as Investigators in the HHMI as part of a national search. Gerald Rubin, the vice president of HHMI, urged the new HHMI investigators to use this opportunity and the freedom and resources it provided not simply to do more of the same thing we had been doing, but rather to start novel areas that have the potential to have far broader impact.

It was in that context that we decided to think seriously about approaches to carrying out proteome-wide studies. We fairly quickly came to two conclusions: First, that the appropriate use of robotics would make it possible to create comprehensive libraries of tagged proteins with a relatively small number of people in a short period of time (months, not years). Indeed it was clear that the effort in creating such a library would be far less than the collective efforts that were going on in many different yeast labs to tag their favorite proteins in a piecemeal manner. Second, once created, such libraries would make it possible to explore the yeast proteome in unprecedented depth and precision. Indeed it was precisely because so many labs were actively engaged in tagging and analyzing individual proteins that made us confident that such comprehensive libraries would be of enormous immediate value to the community. We also made an early strategic decision to start from scratch (which meant among other things committing to synthesizing 18,000 oligonucleotides) rather than trying to make a less than ideal library by taking advantage of existing oligo sets. The design and synthesis of new oligos allowed us to create tagged yeast strains in which each polypeptide was fused precisely at its carboxy terminus and was expressed from its endogenous promoter (Fig. 1). This design was critical, as it gave us the best possible chance of obtaining functional



**Figure 1.** Strategy for the construction of epitope-tagged yeast strains. Gene-specific PCR products containing DNA coding for the epitope tag (GFP or TAP) and a selectable marker (*HIS3MX6*), with ends homologous to the desired site of insertion in the genome, were transformed into yeast. Homologous recombination with target genes on the yeast chromosome leads to the generation of yeast strains, each of which expresses a single fusion protein epitope-tagged at its C-terminus with GFP or TAP.

proteins expressed at natural levels. Moreover, because we were working in haploid strains in which the tagged protein replaced the endogenous one, we could directly test the functionality of the tagged proteins. Finally, we decided to make simultaneously two different libraries: one in which proteins were fused to GFP, allowing the subcellular localization of the protein to be determined, and a second in which proteins were fused to a high affinity epitope tag, termed the tandem affinity purification (TAP)-tag, which made it possible to detect proteins with extremely high sensitivity by Western blot analysis as well as to rapidly purify protein complexes.

With the libraries in hand we then set out to use them to define the intracellular location and estimate the absolute abundance of the yeast proteome (Ghaemmaghami et al. 2003; Huh et al. 2003). The localization studies were based on microscopic analysis of the GFP-tagged strains. Early on we made the critical decision that for proteins that could not be unambiguously localized on the basis of the pattern of fluorescent protein we would carry out co-localization studies using RFP-tagged proteins of known subcellular localization. While this greatly increased the effort required to complete our studies, it also dramatically increased the value of the data. In total, we were able to obtain localization for ~4200 proteins. Prior to our work the collective individual effort of many hundreds of laboratories had defined the localization of ~3200 of the yeast proteins. With our studies we were able to define the localization for 70% of those proteins that were not previously characterized.

We also took advantage of the tagged strains to provide a census of the yeast proteome. Specifically, using immunodetection of the TAP tag together with microscopy of the GFP-tagged strains, we defined the complement of proteins

expressed during log-phase growth. Remarkably, we find that about 80% of the proteome is expressed during normal growth conditions. Using this expression information, together with a novel metric termed the codon enrichment coefficient (CEC), we were able to systematically identify misannotated genes. Finally, we used quantitative Western blotting to estimate the absolute abundance of each detectable protein. These studies revealed that protein abundances range from fewer than 50 to more than  $10^6$  molecules per cell. Many proteins, including essential proteins and most transcription factors, are present at levels that are neither readily detectable by other proteomic techniques nor predictable by mRNA levels or codon bias measurements.

A critical aspect of our efforts that we had not fully appreciated when we began was ensuring that all of the data and strains would be available as a resource to the community. To allow rapid access to the data, including the actual fluorescence micrographs, we made them available in an easily accessible form on a Web site (<http://yeastgfp.ucsf.edu>) we host. The task of distributing the strains was naturally a much greater task. Ultimately we settled on using a third-party company (Invitrogen for GFP and Open Biosystems for the TAP-tagged strains). More than 100 copies of the entire collection and several hundred individual strains have been distributed this way, suggesting that the strains are serving their intended purpose as a resource for the yeast community.

We have continued to pursue several projects exploiting the strain collections to explore the dynamics, localization and activity of the yeast proteome. Our most mature efforts include (1) comprehensive characterization of protein lifetimes in cells, (2) dynamic measurements of changes in protein abundance and subcellular localization, and (3) measurements of protein expression in single cells using flow cytometry. It is certain that other labs will find many creative uses for the strains and the data we have generated.

### Acknowledgments

We acknowledge the HHMI and the David and Lucile Packard Foundation for their support of this work. We also thank the following talented people in our laboratories who carried out the localization and abundance experiments, and who helped in the construction of the strain collections and database: Archana Belle, Kiowa Bower, Adam Carroll, Noah Dephoure, James Falvo, Luke Gerke, Sina Ghaemmaghami, Rusty Howson, Won-Ki Huh, and Felix Lam.

### References

- Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. 2003. Global analysis of protein expression in yeast. *Nature* **425**: 737–741.
- Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. 2003. Global analysis of protein localization in budding yeast. *Nature* **425**: 686–691.