Proteomics
www.proteomics-journal.com

Strategies and Challenges in Identifying Function for Thousands of sORF-Encoded Peptides in Meiosis

*Ina Hollerer, Andrea Higdon, and Gloria A. Brar**

Recent genomic analyses have revealed pervasive translation from formerly unrecognized short open reading frames (sORFs) during yeast meiosis. Despite their short length, which has caused these regions to be systematically overlooked by traditional gene annotation approaches, meiotic sORFs share many features with classical genes, implying the potential for similar types of cellular functions. We found that sORF expression accounts for approximately 10–20% of the cellular translation capacity in yeast during meiotic differentiation and occurs within well-defined time windows, suggesting the production of relatively abundant peptides with stage-specific meiotic roles from these regions. Here, we provide arguments supporting this hypothesis and discuss sORF similarities and differences, as a group, to traditional protein coding regions, as well as challenges in defining their specific functions.

1. Introduction

Modern technologies enable sequencing and assembly of genomes in a rapid fashion and at low cost. Deciphering the information encoded in these genomes to yield gene annotations, however, requires specific knowledge and the application of broad assumptions about features of coding and regulatory regions. These assumptions may evolve over time based on the accumulation of new knowledge that prompts re-evaluation of annotations to include, for example, new classes of small and long noncoding RNAs. Outside of a few specific cases,[1–3] however, the rules for identification of coding regions have remained surprisingly static and based on historical ideas of open reading frame (ORF) properties. These rules include the requirement for canonical AUG start codons at ORF beginnings and stop codons at ends, and a minimum codon length of 100. This conservative lower length cutoff reflected an effort to minimize false positive ORF calls and was generally adopted as default for all organisms. Genetic and proteomic approaches in diverse organisms have identified numerous examples of important shorter coding regions, such as the conserved large ribosomal subunit protein, Rpl41,[4] kinetochore components Hsk3 and Dad4,[5] or the polished rice (*pri*) short open reading frame (sORF) genes, involved in embryogenesis in Drosophila,[6,7] but these have generally been treated as exceptional cases. The development of ribosome profiling[8] provided the first unbiased and global view of translated regions and revealed condition-specific translation from formerly unannotated sORFs in various organisms, including yeast,[9] zebrafish,[10] fly,[11,12] and mouse.[13] In parallel, proteomic approaches have validated cellular accumulation, and occasionally function, for several sORF-encoded peptides, such as the human MRI-2 peptide.[13–19] The results of these studies suggest a general need to re-evaluate coding region annotation rules and present the exciting and daunting possibility that eukaryotic cells may contain a large set of overlooked and currently functionally mysterious components.

Dr. I. Hollerer, A. Higdon, Dr. G. A. Brar
Department of Molecular and Cell Biology
University of California
Berkeley, CA, USA
E-mail: gabrar@berkeley.edu
Dr. I. Hollerer, A. Higdon, Dr. G. A. Brar
California Institute for Quantitative Biosciences (QB3)
University of California
Berkeley, CA, USA

DOI: 10.1002/pmic.201700274

2. Pervasive sORF Translation in Meiotic Yeast Cells

One of the most dramatic cases of translated sORF discovery to date has been in budding yeast cells undergoing meiotic differentiation.[9] The meiotic program is the highly temporally regulated process by which haploid gametes are produced from a diploid precursor cell. We used ribosome profiling to probe translation during this process and noticed evidence for many short translated regions outside of annotated coding regions, specifically in meiotic cells relative to vegetative controls.[9] These regions fell into two categories: upstream ORFs (uORFs), which were positioned in 5' leaders of previously annotated genes, and independent sORFs, which were translated from transcripts that were either previously unrecognized or thought to be noncoding.[20] These transcripts included those that would be traditionally characterized as "intergenic," as well as "antisense," relative to traditional ORF-encoding transcripts. In total, we detected 2555 such independent sORFs, a remarkable number considering the thorough study of the genome of this organism, which previously was recognized to encode only around 6600 genes.

Given the surprising apparent scale of meiotic sORF translation, we performed extensive analyses to determine if these results might reflect an artifact of the ribosome profiling

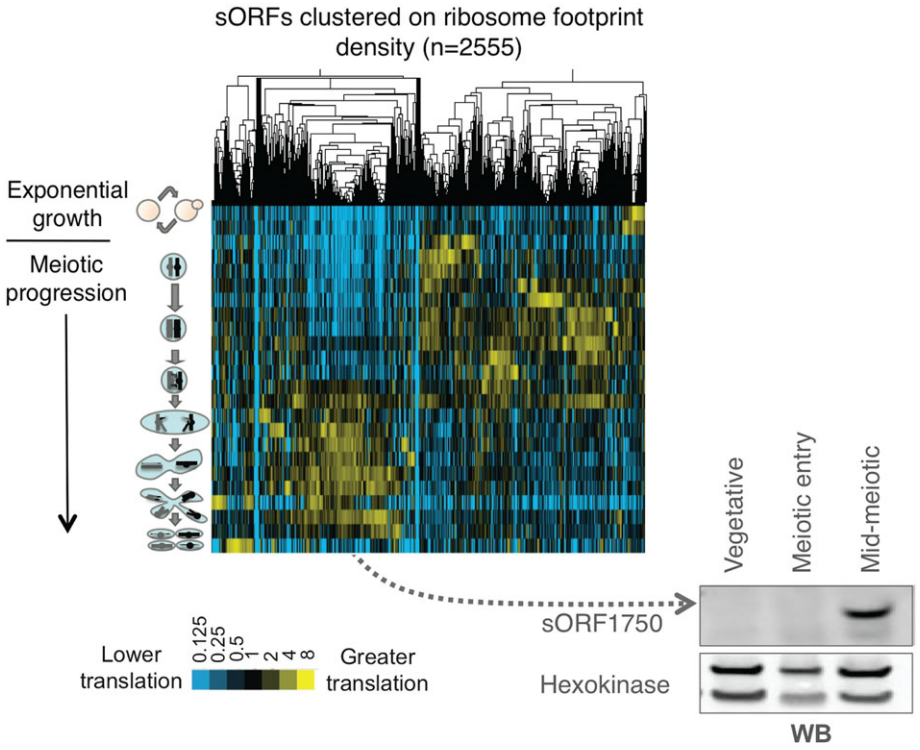

Figure 1. Ribosome profiling through budding yeast meiosis revealed regulated translation of 2555 meiotic sORFs encoded by freestanding transcripts. Ribosome footprints summed over each newly identified independent yeast sORF (columns) in exponentially growing cells (top rows) and cells at different stages of meiosis (below, represented by cartoons to the left of the dendrogram plot). Blue and yellow represent low and high translation, respectively. A Western blot confirming the predicted meiotic expression of an individual C-terminally GFP (Green Fluorescent Protein)-tagged sORF is shown on the right.

methodology. A simple approach that we employed for independent verification of translation was the C-terminal insertion of a tag-encoding sequence prior to the stop codon of newly predicted sORF loci and Western blotting for detection of the fusion protein. This strategy confirmed production of a stable tagged protein from these loci for many cases tested[13] (**Figure 1**), but is not easily scalable to test for protein production from all new sORFs and is likely to affect natural protein stability based on the large size of the tags tested relative to the predicted sORF-encoded peptides. We therefore employed several systematic approaches to assess the degree to which the features of sORF translation matched those seen for canonical ORFs. We noted that sORFs could be annotated using the same rules as classical ORFs, indicating bounding by canonical start and stop codons. We further found that the fragment sizes of the RNAs detected in ribosome profiling over the newly predicted meiotic sORFs were identical to that seen over canonical ORFs, which reflects the robust and specific biophysical properties of a translating ribosome as it protects the mRNA that it is decoding. The RNA fragments derived from sORFs could be clearly distinguished in size from contaminating noncoding RNA fragments, such as tRNAs and snoRNAs, that do not result from ribosome protection.[13] These results strongly suggested that approximately 2500 meiotic sORFs that we predicted by ribosome profiling data are indeed translated, but could not confirm resultant protein stability or, most importantly, provide evidence of functional significance for these sORFs.

3. Evidence for Meiotic Function of sORFs

In assessing the likelihood that approximately 2500 newly identified sORF-encoded peptides perform function in meiotic yeast cells, their sheer number is both a strength and a weakness. It allows analyses of sequence feature trends, but precludes systematic detailed molecular analysis. We have first focused, therefore, on bulk analyses of sORF features. One strong piece of evidence in favor of a functional role for the meiotic sORF-encoded peptides as a group is their expression levels. Both at the mRNA and translation level, the range of expression for these short genes is broadly comparable to that of canonical genes, constituting in sum approximately 10–20% of the cell's translational capacity, depending on specific meiotic stage. Second, sORF translation is highly regulated, and to a degree similar to canonical genes. Within meiosis, they are translated in precise temporal windows of action (Figure 1), with a variety of discrete patterns observed, on par with the diversity of translation patterns seen for characterized ORFs. Few of the meiotically translated sORFs are translated in vegetative cells, providing a possible explanation for why so few of these regions have been flagged for function in the many genetic screens performed in this well-studied eukaryote. Finally, analysis of the base and amino acid composition of meiotic sORFs and their predicted encoded peptides show contents that are generally similar to canonical ORFs (with some exceptions discussed below) (**Figure 2**).

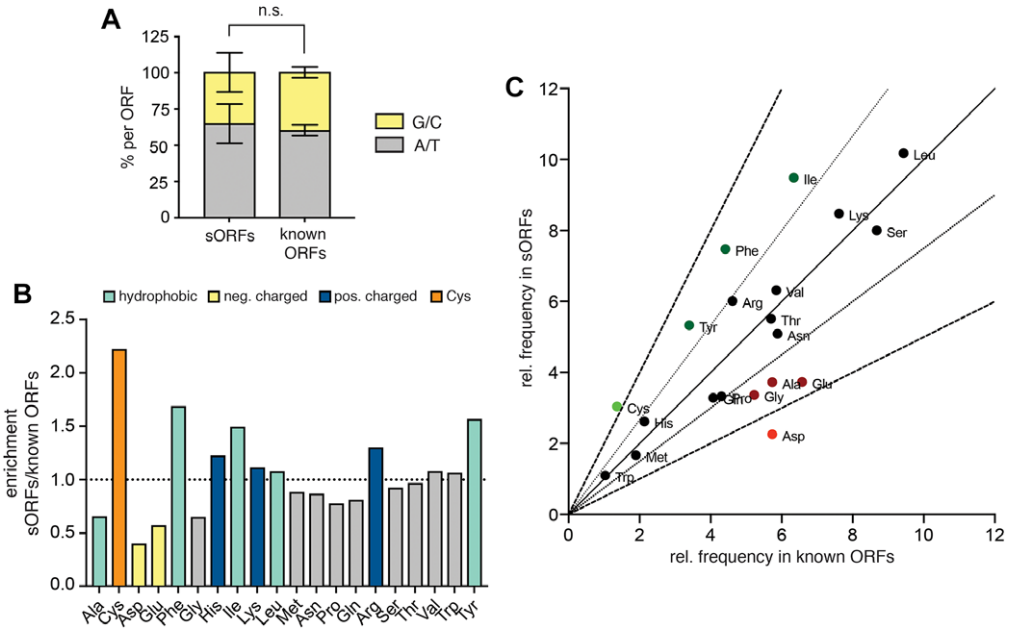

Figure 2. The overall amino acid composition of sORFs and known ORFs is similar, with interesting exceptions. A) G/C and A/T content of meiotic sORFs (independent sORFs and AUG-initiated uORFs) and known ORFs. Chi-square test was applied to calculate statistical significance. Start and stop codons were computationally removed from each ORF prior to analysis. B) Amino acid composition of peptides produced from sORFs (independent sORFs and AUG-initiated uORFs) relative to those of proteins translated from known ORFs. Fold enrichment of individual amino acids in sORF-encoded peptides compared to proteins translated from known ORFs. C) Enriched or disenriched amino acids for all sORFs (independent and AUG-initiated uORFs) are shown in green and red, respectively, with lines indicating the fold difference relative to known ORFs (solid line = no enrichment, dotted line = 1.33-fold change, bold dotted line = 2-fold change). For the analyses in (B) and (C), the content of all ORFs in each category was pooled, after computational removal of the start codon-encoded methionine.

4. Unusual Features of Meiotic sORF-Encoded Peptides

The most obvious difference between meiotic sORFs and canonical ORFs is their size, with sORFs showing a mean length of 22, compared to 485 codons for previously characterized genes.[9] This makes quantification of sORF-encoded peptides difficult, as traditional MS might produce no tryptic peptides of detectable length or may result in only one or two theoretically detectable peptides per sORF-protein that are not efficiently captured in practice. Even in detectable cases, the concentration of such peptides will be small relative to the cellular peptide pool. It is known that even for well characterized, highly expressed sORF-derived proteins (like Rpl41, which is 25 amino acids in length), quantification by traditional MS is not typically feasible. Consistently, we currently are able to detect very few of the predicted sORF-encoded peptides by standard MS.[21] Traditional conservation analyses do not suggest robust selection for most sORF regions, although these analyses are also not fully independent of length. The short length of these proteins further has implications for their likely cellular functions. We observe, for example, that catalytic domains are not predicted in the 2555 meiotic sORF-encoded peptides, which is not surprising given that short amino acid strings would likely not be capable of folding into multi-domain structures and an isolated catalytic domain might be expected to cause many nonspecific and detrimental cellular effects.

5. What Might sORF-Encoded Proteins Do?

With so few functionally defined sORF-encoded peptides in eukaryotes, we can only speculate about the potential functions of meiotic sORF-encoded proteins based on their sequence and expression patterns. Given that they show classical as well as non-canonical features, sORF-derived peptides might either act like traditional proteins or exert non-classical functions, or both. Our computational analyses showed that the overall amino acid compositions of meiotic small peptides and proteins translated from known ORFs are similar (Figure 2B and C). A few residues, however, are clearly relatively over- and under-represented in sORF-encoded peptides (Figure 2B and C). The chemical properties of these residues might reflect specific cellular functions of sORF-encoded proteins. Most strikingly, peptides derived from meiotic sORFs are predicted to be rich in cysteine residues (Figure 2B and C), an amino acid whose highly reactive thiol group (R–SH) equips it with unique chemical properties. These include the ability to regulate redox potential or promote intra- and intermolecular interactions by disulfide bond formation and ion coordination, which may prove to be a hint to a subset of cellular roles and/or organizations of these small proteins.

An additional valuable observation from our meiotic dataset was that regulated and meiotically-enriched translation could be seen for short ORFs that initiated internal to and in-frame with classical ORFs. This set of sORFs would be predicted to encode truncated isoforms of traditional proteins and was not

easily annotated by our original ribosome profiling approach due to the difficulty in distinguishing ribosome footprint reads derived from these internal sORFs and those from the overlapping, canonical ORFs. The instances that could be identified most readily were those in which the canonical ORF was fortuitously very lowly translated in meiotic cells, revealing fewer than 100 such cases with high confidence. The reverse case, in which a lowly translated vegetative ORF revealed a truncated sORF translated at the same locus, however, has not been yet detected in our datasets, providing further evidence for some particular cellular importance for sORF translation in meiosis. More importantly, this subset of internal sORFs allows hypothesis-based experiments to be performed to deduce meiotic sORF function, leveraging existing knowledge of the domain structure and known function of the classical genes at these loci. In several cases, we observe that these truncated protein isoforms encode only the region of the full-length protein containing a known protein interaction domain, which we predict might be capable of regulating protein–protein interactions between canonical proteins in a short and isolated form.[9,22] This function may be specific, for example, enabling or precluding a canonical signaling interaction, or nonspecific. The large number of small proteins predicted to be translated from independent sORFs in meiotic cells could, for example, potentially function in part as a "cellular blocking buffer," inhibiting non-specific interactions without affecting the functions of the proteins they bind to, which might be especially important in times of rapid cellular change, such as during meiosis. It is thus intriguing that, in addition to meiosis, the most pervasive cases of sORF translation have been observed in dynamic and/or developmental contexts,[10,12,15,23] situations in which rapid cellular changes are common and important.

6. Challenges in Determining the Functions of sORF-Encoded Peptides

The major challenges in evaluating cellular roles for sORFs are that (1) there are many of them, (2) they could affect any number of cellular phenotypes, and (3) they might have a high level of functional redundancy. There is currently a gap between the analytical approaches that have validated sORF translation and the detailed functional dissection that has been achieved for only a small fraction of these peptide products. Hypothesis-driven functional analyses are significantly more challenging for these cases than for "hits" from a genetic screen, which are identified based on some known functional feature that can be subsequently investigated in further detail. We believe that new genome-wide approaches could provide an alternate route to determine cellular roles for these new short meiotic genes, allowing testing of function for many sORFs, and for many phenotypes in parallel. CRISPRi,[24] which makes use of a catalytically dead version of the Cas9 nuclease (dCas9), for example, offers the possibility to inhibit the transcription of sORF-encoding mRNAs in a dominant fashion to investigate the functional relevance of each single knockdown in yeast meiosis through pooled screening. The CRISPRi technique can further be applied to inhibit the expression of multiple sORF genes in parallel allowing testing for possible functional redundancies of individual meiotic sORFs. Such large-scale analyses could thus circumvent the earlier-mentioned challenges faced in determining sORF function and provide an attractive and unbiased approach to begin to unravel the cellular importance of the thousands of sORFs that are translated during meiosis.

7. Concluding Remarks

Beyond a few specific instances in which sORF-encoded peptide function has been defined, the general roles for these small proteins have remained mysterious in all eukaryotic systems in which they have been observed. Given the sheer number of sORFs seen to be translated during the meiotic program in budding yeast, this system seems a promising one to investigate the general set of functions that can be mediated by small proteins during a natural developmental process. Indeed, the functions of many classes of traditional gene products, in meiosis and beyond, were first determined in this single-celled eukaryote. The task of systematically defining sORF functions is a big one, however. It is worth noting that to date, over 1400 long and traditionally annotated budding yeast ORFs retain generic systematic naming, reflecting no known function for their encoded proteins, usually despite strong evidence for their expression and decades of intensive study of this organism. It is perhaps not surprising then, that despite multiple lines of strong evidence for translation of thousands of novel short ORFs in eukaryotes, our knowledge of the cellular functions of their peptide products is in its infancy. Their short length is an impediment to some traditional functional approaches, but there is great promise in the power of new proteomic and genomic experimental and computational tools, which have been so valuable in discovery of the pervasiveness of sORF translation, in beginning to delineate their function and progressively illuminating the surprising number of remaining dark corners of the eukaryotic genome, even in the simple budding yeast.

Abbreviations

sORFs, short open reading frames; uORFs, upstream ORFs

Acknowledgements

Our lab's work on sORFs is supported by NIH funding (DP2-GM-119138), and investigator awards from the Alfred P. Sloan Foundation and Pew Charitable Trusts to G.A.B. A.H. is supported by an NSF pre-doctoral fellowship.

Conflict of Interest

The authors declare that they have no conflict of interests.

Keywords

meiosis, small peptides, sORFs, yeast

Received: August 14, 2017

[1] M. R. Hemm, B. J. Paul, T. D. Schneider, G. Storz, K. E. Rudd, *Mol. Microbiol.* **2008**, *70*, 1487.
[2] A. Kumar, P. M. Harrison, K. H. Cheung, N. Lan, N. Echols, P. Bertone, P. Miller, M. B. Gerstein, M. Snyder, *Nat. Biotechnol.* **2002**, *20*, 58.
[3] M. M. Kessler, Q. Zeng, S. Hogan, R. Cook, A. J. Morales, G. Cottarel, *Genome Res.* **2003**, *13*, 264.
[4] K. Suzuki, T. Hashimoto, E. Otaka, *Curr. Genet.* **1990**, *17*, 185.
[5] J. M. Li, Y. Li, S. J. Elledge, *Mol. Cell Biol.* **2005**, *25*, 767.
[6] T. Kondo, S. Plaza, J. Zanet, E. Benrabah, P. Valenti, Y. Hashimoto, S. Kobayashi, F. Payre, Y. Kageyama, *Science* **2010**, *329*, 336.
[7] J. Zanet, E. Benrabah, T. Li, A. Pelissier-Monier, H. Chanut-Delalande, B. Ronsin, H. J. Bellen, F. Payre, S. Plaza, *Science* **2015**, *349*, 1356.
[8] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science* **2009**, *324*, 218.
[9] G. A. Brar, M. Yassour, N. Friedman, A. Regev, N. T. Ingolia, J. S. Weissman, *Science* **2012**, *335*, 552.
[10] A. Pauli, M. L. Norris, E. Valen, G. L. Chew, J. A. Gagnon, S. Zimmerman, A. Mitchell, J. Ma, J. Dubrulle, D. Reyon, S. Q. Tsai, J. K. Joung, A. Saghatelian, A. F. Schier, *Science* **2014**, *343*, 1248636.
[11] J. L. Aspden, Y. C. Eyre-Walker, R. J. Philips, U. Amin, M. A. S. Mumtaz, M. Brocard, J. P. Couso, *Elife* **2014**, *3*, e03528.
[12] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, A. J. Giraldez, *EMBO J.* **2014**, *33*, 981.
[13] N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne, S. E. Jackson, M. R. Wills, J. S. Weissman, *Cell Rep.* **2014**, *8*, 1365.
[14] S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, *J. Biol. Chem.* **2014**, *289*, 10950.
[15] N. Stern-Ginossar, B. Weisburd, A. Michalski, V. T. Le, M. Y. Hein, S. X. Huang, M. Ma, B. Shen, S. B. Qian, H. Hengel, M. Mann, N. T. Ingolia, J. S. Weissman, *Science* **2012**, *338*, 1088.
[16] A. G. Schwaid, D. A. Shannon, J. Ma, S. A. Slavoff, J. Z. Levin, E. Weerapana, A. Saghatelian, *J. Am. Chem. Soc.* **2013**, *135*, 16750.
[17] B. Vanderperre, J. F. Lucier, C. Bissonnette, J. Motard, G. Tremblay, S. Vanderperre, M. Wisztorski, M. Salzet, F. M. Boisvert, X. Roucou, *PLoS One* **2013**, *8*, e70698.
[18] S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. A. Budnik, J. L. Rinn, A. Saghatelian, *Nat. Chem. Biol.* **2013**, *9*, 59.
[19] J. Ma, C. C. Ward, I. Jungreis, S. A. Slavoff, A. G. Schwaid, J. Neveu, B. A. Budnik, M. Kellis, A. Saghatelian, *J. Proteome. Res.* **2014**, *13*, 1757.
[20] A. Lardenois, Y. Liu, T. Walther, F. Chalmel, B. Evrard, M. Granovskaia, A. Chu, R. W. Davis, L. M. Steinmetz, M. Primig, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1058.
[21] Z. Cheng, M. Jovanovic, A. Regev, G. A. Brar, Unpublished data.
[22] I. Hollerer, A. Uebersohn, A. Higdon, G. A. Brar, Unpublished data.
[23] H. Li, C. Hu, L. Bai, H. Li, M. Li, X. Zhao, D. M. Czajkowsky, Z. Shao, *DNA Res.* **2016**, *23*, 571.
[24] L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman, L. S. Qi, *Cell* **2013**, *154*, 442.