Title

Small, but mighty? Searching for human microproteins and their potential for understanding health and disease

Journal

Expert Review of Proteomics, 15(12)

ISSN

1478-9450

Authors

Rathore, Annie
Martinez, Thomas F
Chu, Qian
et al.

Publication Date

2018-12-02

DOI

10.1080/14789450.2018.1547194

HHS Public Access

Author manuscript

Expert Rev Proteomics. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Expert Rev Proteomics. 2018 December ; 15(12): 963–965. doi:10.1080/14789450.2018.1547194.

Small, but mighty? Searching for human microproteins and their potential for understanding health and disease

Annie Rathore, **Thomas F. Martinez**, **Qian Chu**, and **Alan Saghatelian**

Clayton Foundation Laboratories for Peptide Biology, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, United States

Keywords

microproteins; peptides; small proteins; small open reading frames (smORFs); proteomics; genomics

1. Introduction

Microproteins are a rapidly expanding class of peptides and small proteins translated from protein-coding small open reading frames (smORFs, less than 100–150 codons in length). Microprotein is a term that refers to peptides and small proteins that are translated from smORFs and can include known genes. Microprotein discovery and characterization reshapes our understanding of proteome composition and reveals new biological pathways [1]. Genomes contain thousands of open reading frames (ORFs), defined as the protein-coding sequence between an in-frame start and stop codon. Annotation of protein-coding ORFs from DNA sequences became paramount as whole-genome sequencing projects reached completion [2]. Excellent computational methods were developed and utilized to define genes, but these tools needed to establish parameters to reduce false positives. For this reason, most genome annotation pipelines required ORFs to be at least 300 nucleotides long (i.e., 100 amino acids) resulting in most smORFs being missed [2]. To get an idea on the challenge of assigning protein-coding genes without a length cutoff, Basarai, Hieter, and Boeke identified ~260,000 smORFs between 2–99 codons when plotting all ORFs in the yeast genome [3]. Today, it is clear that smORFs and their corresponding microproteins make up a sizable fraction of the genome and proteome. As new genes, very little is known about the structure and function of microproteins making these genes an incredible opportunity for discovering new biology.

Correspondence: Alan Saghatelian, Clayton Foundation Laboratories for Peptide Biology, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, United States, asaghatelian@salk.edu.

Declaration of Interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

2. Evidence for smORF translation

Pioneering studies of protein translation initiation first identified the presence of smORFs in the 5'-untranslated regions (UTRs) of some mRNAs, which are commonly referred to as uORFs [4]. uORFs regulate the translation of downstream ORFs by engaging the ribosome before it initiates at the downstream ORF. Translation of downstream ORF is thought to occur through leaky scanning, whereby the first uORF initiation site is skipped, or by re-initiation of ribosome scanning upon reaching the end of the uORF [4]. Though initially thought to be unimportant the translated sequence of the uORF was shown to matter in several cases. For instance, a uORF in the AdoMet carboxylase mRNA encodes a hexapeptide that when mutated effects the efficiency of translation repression of the downstream AdoMet carboxylase ORF [5]. This observation demonstrated that in some cases it is the interaction between the translated peptide and the ribosome that is mediating the activity of a uORF. More generally, this example and others like it show that uORFs/smORFs are translated and that the microprotein they encode are biologically active, providing the first evidence for functional microproteins. Since these findings smORFs have been identified other non-coding regions [1], including 3'-UTR and RNAs that were thought to be non-coding.

3. The smORFeome

These discoveries suggested that genomes contain protein-coding smORFs but determining the size of the smORFeome is a challenge. As mentioned, searching the yeast genome for all smORFs between 2–99 codons identified approximately 260,000 smORFs [3], but only a small fraction of these are likely translated. Improvements in computational methods, especially the use of homology between closely related species, provided some relief and have improved the prediction of bona fide smORFs. For instance, an informatics analysis of *Drosophila melanogaster* genome identified ~600,000 potential smORFs that were culled to ~400–4,500 smORFs by looking for conservation in another Drosophila species, *Drosophila pseudoobscura* [6]. The empirical discovery of smORFs took a huge step forward with the advent of Ribosome Profiling (Ribo-Seq) by Weissman and colleagues [7]. Ribo-Seq is based on the deep sequencing of ribosome-protected mRNA fragments called ribosome footprints, and these footprints are used to identify translated regions of the transcriptome. Ribo-Seq provides the only empirical method to measure transcriptome-wide translation comprehensively. Ribosome-profiling provided the first transcriptome-wide method to empirically identify novel smORFs and non-AUG start codons [8], which has been invaluable in the assignment of smORFs. Thus, improvements in smORF prediction and measurement have revealed the smORFeome to be a large missing fraction of the ORFeome —perhaps several thousand smORFs.

4. The microproteome

The integration of proteomics with genomics, or proteogenomics, complements smORF discovery approaches by providing direct translation evidence of microproteins. An early example of this approach took transcripts from the RefSeq database and translated them in all possible reading frames to create a proteomics database that was then searched with

proteomics data to identify novel smORFs [9]. This proof-of-concept study identified four non-annotated ORFs including one smORF. The advent of next-generation RNA sequencing (RNA-Seq) and the use of RNA-Seq data to create a custom proteomics database, which is commonly referred to as proteogenomics, led to dramatic improvements in the number of new microproteins discovered. In two separate studies, our application of proteogenomics led to the discovery 323 microproteins from human cell lines and tissues [10, 11]. Likewise, another study combined smORF prediction with proteomics to find 1259 microproteins from human tissues [12]. These experiments provided substantial evidence that at least some microproteins are stable and abundant enough to be detected by proteomics, but also underscore a discordance in total numbers between proteomics (hundreds), Ribo-Seq (hundreds to thousands), and computational (thousands) methods. Several factors may limit the proteomic detection of microproteins including the low abundance or stability of microproteins, and the technical challenge that microproteins often yield a single tryptic peptide making their detection more difficult than longer proteins.

The various methods for smORF discovery offer different strengths. Proteogenomics provides strong evidence for a stable, abundant microproteins, but misses most microproteins. Ribo-Seq provides a superior overview of the smORFeome, but does not provide evidence that translation is producing is stable or abundant. We believe that Ribo-Seq is better for smORFeome-wide analyses but when deciding to study a single smORF/microprotein evidence of translation at the protein level (proteomics or immunochemistry) is required.

5. Functional microproteins

With the elucidation of a large unannotated smORFeome and microproteome it became important to understand whether any of these genes are functional so as to justify their investigation [1]. The discovery that inspired most studies into human smORFs/microproteins came from the world of insect developmental biology. Investigation of mRNAs that were regulated during development uncovered a gene called Tarsal-less [13] or polished rice [14] (tal/pri) that was found to encode four smORFs translating to three 11- and one 32-amino acid microproteins that regulate fly development. Tal/pri demonstrates that smORFs produce functional proteins, albeit small microproteins, and leads to the hypothesis that genomes might harbor fundamentally important smORFs that were still not annotated. Testing this hypothesis led the discovery of important mammalian smORFs, such as Cell Cycle Regulator of Non-Homologous End Joining (CYREN) [15]. CYREN inhibits non-homologous end joining repair (NHEJ) during the S and G2 phases of the cell cycle to enable slower but more accurate homologous recombination (HR) to dominate. Upon DNA damage in S and G2 phase, CYREN activity preserves DNA integrity by reducing chromosomal rearrangements through its inhibition of NHEJ. CYREN reveals a microprotein with fundamental role in DNA repair, a process that is intimately linked to cancer.

6. The future of smORF/microprotein research

smORFs and microproteins represent a frontier in biochemistry, molecular biology, and physiology that is at its inception. It is likely that many more microproteins await discovery and characterization. Based on the current state of the field, several future directions seem likely. First, methods for the elucidation of smORFs are not designed to provide insight into the functions of these genes, only that they exist. With so many smORFs already discovered higher throughput methods in the form of gain- or loss-of-function screens with smORFs are needed to find the most interesting smORFs/microproteins for further investigation. Second, the integration of smORFs into big data will provide additional methods to identify the most interesting smORFs. For example, combining smORF discovery with GWAS data can identify disease-associated smORFs, or mining expression profiling data can identify smORFs that are up or down regulated in different diseases. Lastly, smORF and microprotein research is still basic research, and even though some of the pathways regulated by microproteins are associated with prevalent diseases, such as DNA repair and cancer, it must still be demonstrated that research into smORFs/microproteins can be translated to benefit humanity. We believe that smORFs/microproteins will impact medicine in several ways. Our biochemical studies, for instance, demonstrate that microproteins that utilize short sequences (usually 2–4 amino acids) [15] to bind to more massive protein complexes to regulate biology. Interactions that utilize short peptide interactions are amenable for small molecule inhibition, and, therefore, microprotein-protein interactions will reveal new druggable targets for medicine.

In conclusion, research into microproteins so far has proven valuable in revealing new genes and biological mechanisms for the regulation of significant processes, while also expanding the limits of what was considered a typical gene or protein. We suspect that continued investigations will begin to find more smORFs/microproteins linked to human disease, and in the future new medicines may emerge from these studies.

Funding

This paper was supported by funding from Timken Sturgis Foundation Award and Salk Women in Science Fellowship (A. Rathore), the George E. Hewitt Foundation (Q. Chu), NIH Ruth L. Kirschstein National Research Service Award (NRSA) Individual Postdoctoral Fellowship Award (T.F.M, F32 GM123685). The Helmsley Center for Genomic Medicine, NIH National Cancer Institute Cancer Center Support Grant P30 (CA014195 MASS core, A. Saghatelian), R01 (GM102491, A. Saghatelian). Additional support was provided by the Leona M. and Harry B. Helmsley Charitable Trust grant (A. Saghatelian), and Dr. Frederick Paulsen Chair/Ferring Pharmaceuticals (A. Saghatelian). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

References

1. Saghatelian A, Couso JP. Discovery and characterization of smORF-encoded bioactive polypeptides. Nat Chem Biol. 2015;11(12):909–916. [PubMed: 26575237]
2. Burge CB, Karlin S. Finding the genes in genomic DNA. Curr Opin Struct Biol. 1998;8(3):346–54. [PubMed: 9666331]
3. Basrai MA, Hieter P, Boeke JD. Small Open Reading Frames: Beautiful Needles in the Haystack. Genome Research. 1997;7(8):768–771. [PubMed: 9267801]

4. Gray NK, Wickens M. Control of translation initiation in animals. Annu Rev Cell Dev Biol. 1998;14:399–458. [PubMed: 9891789]
5. Hill JR, Morris DR. Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame. J Biol Chem. 1993;268(1):726–31. [PubMed: 8416975]
6. Ladoukakis E, Pereira V, Magny EG, et al. Hundreds of putatively functional small open reading frames in Drosophila. Genome Biology. 2011;12(11):R118–R118. [PubMed: 22118156]
7. Ingolia NT, Ghaemmaghami S, Newman JRS, et al. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009;324(5924):218–223. [PubMed: 19213877]
8. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes. Cell. 2011;147(4):789–802. [PubMed: 22056041]
9. Oyama M, Kozuka-Hata H, Suzuki Y, et al. Diversity of Translation Start Sites May Define Increased Complexity of the Human Short ORFeome. Molecular & Cellular Proteomics. 2007;6(6): 1000–1006. [PubMed: 17317662]
10. Ma J, Ward CC, Jungreis I, et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J Proteome Res. 2014;13(3):1757–65. [PubMed: 24490786]
11. Slavoff SA, Mitchell AJ, Schwaid AG, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol. 2013;9(1):59–64. [PubMed: 23160002]
12. Vanderperre B, Lucier JF, Bissonnette C, et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS One. 2013;8(8):e70698. [PubMed: 23950983]
13. Galindo MI, Pueyo JI, Fouix S, et al. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 2007;5(5):e106. [PubMed: 17439302]
14. Kondo T, Hashimoto Y, Kato K, et al. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nature Cell Biology. 2007;9:660. [PubMed: 17486114]
15. Arnoult N, Correia A, Ma J, et al. Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN. Nature. 2017;549:548. [PubMed: 28959974]