Journal of Phylogenetics & Evolutionary Biology

Brocchieri, J Phylogen Evolution Biol 2016, 4:2
DOI: 10.4172/2329-9002.1000e120

Editorial Open Access

Discovering Elusive Small Genes

Luciano Brocchieri*

Department of Molecular Genetics and Microbiology and Genetics Institute, University of Florida, Gainesville, FL 32606, USA

The importance of short protein-coding genes (usually defined as no longer than 100 codons) and corresponding small proteins and peptides, in prokaryotic and eukaryotic organisms is becoming increasingly obvious as the pervasive role of small proteins as signaling molecules, and as regulators of protein expression and functionality is being uncovered (e.g., [1,2]).

In *E. coli*, the smallest known functional gene-product is a 29-amino acid peptide involved in K^+ transport (KdpF) [3]. ORFs with as few as 14 amino acids have been predicted to encode functional genes in *E. coli*, and with as few as 28 amino acids in *Saccharomyces cerevisiae* [4], while many short peptides have been identified, including a 13 aa peptide encoded within the *E. coli* Shiga-like toxin operon. Moreover, artificial constructs encoding just six amino acids were able to transcribe and result in functional gene-products involved in intracellular signaling in *B. subtilis* [5]. In eukaryotes, many important signaling molecules are short peptides, including various peptide hormones, cytokines and co-repressors or co-activators [6,7]. In eukaryotes, evidence is accumulating on the existence of widespread very short ORFs, called uORFs, located 5' of a reference gene, which post-transcriptionally regulate translation of the gene [8-10].

Computational Identification of Small Genes in Bacterial Genomes

In spite of the high sensitivity of computational prokaryotic-gene predictor methods, it is recognized that short genes are still often overlooked by published annotations. Computational prediction of short genes is risky. Short sequences contain less information that can signal their coding capacity, and small genes may not follow the same codon-composition properties of the average gene, and may encode for peptides with unique amino acid composition. Among the most common predictors, issues of specificity limit the ability of gene predictors to identify small genes with high sensitivity. For example, among popular predictors, Glimmer tends to predict many more short genes than others, but among the many predicted short genes the rate of false positives appears to be so high that a cut-off on the minimal length of predicted ORFs has been included in more recent versions of Glimmer [11], preventing prediction of very short genes.

In an attempt to identify coding regions missed by published annotations, we developed procedures to identify genomic regions with significant 3-base sequence periodicity, which when associated with ORF structures could signal the presence of a coding sequence [12]. These procedures were implemented in the N-Profile Analysis Computational Tool (NPACT), a web-based bioinformatics tool available at http://genome.ufl.edu/npact. We collected with NPACT all genes predicted by the annotations of 1000 prokaryotic genomes, by four other popular prediction methods, and ORFs identified by sequence 3-base periodicity, recording conservation of all genes across different phyla. We identified a total of 4,421,545 predicted genes, among which 889,837 ORFs were not included in the published genome annotations. Most of these excluded genes (83%) corresponded to ORFs no longer than 100 codons (Table 1). This collection of short ORFs would represent almost three-times as many short genes than currently annotated in prokaryotic genomes. How many of these ORFs are functional genes and not just false predictions, remains to be determined. Evolutionary conservation in sequence and in length provided evidence of persistence across genera and phyla for more than 13% of these putative small genes, a percentage that certainly suggests high levels of false positives, but also indicates that at least 97,836 small genes are currently missed from the 1000 genome annotations (Table 1).

Experimental Data on Expression of Small Genes

Small proteins are technically difficult to study because of their poor resolvability and high diffusibility during electrophoresis and/or column chromatography, low intracellular concentrations of many small proteins, possession of reduced number of amino groups and dye/isotope accepting elements per molecule, and interference by protein degradation products. However, global information on gene expression can in principle be obtained by genome-wide transcriptomics, whereas expression of predicted proteins can be recognized by mass spectrometry-based proteomics. While highly informative, these approaches are however not devoid of limitations in both eukaryotic and prokaryotic species, which are aggravated in the case of small genes. In bacteria in particular, quantification of protein expression through transcriptome analysis is limited by the polycistronic nature of the bacterial mRNA and by difficulties in defining the boundaries of operon structures (where are the coding regions in a polycistronic mRNA?), and are confused by the large amount of

anti-sense and non-sense transcription in bacteria [13]. Furthermore, predicting protein expression by mRNA abundance depends on the assumption that the amount of expression of a protein is proportional to the level of transcription of its gene. In fact, comparisons between RNA-seq and mass spectrometry measurements recently quantified a significant discrepancy between mRNA expression and steady-state protein levels [14], suggesting that post-transcription control of translation plays a significant role in determining steady-state levels of proteins in the cell.

Translatome Analysis by Ribosome Profiling

Significant improvements in uncovering expression of proteins of any length and on a genomic scale, can be achieved using the technique of "ribosome profiling" [15]. This technique overcomes limitations of transcriptomics analysis by providing genome-wide quantification at codon resolution of protein-translation activity. Ribosome profiling (or RIBO-seq) is based on deep sequencing mRNA segments engaged in translation that are protected by the ribosome from degradation (the ribosome "footprints"). By this technique, only regions of the mRNA that are actively translated are represented in sequencing libraries, allowing comprehensive quantitative determination of *in vivo* synthesis of translation products, i.e., definition of the "translatome". Thus, in contrast to RNA-seq, RIBO-seq reads identify the exact position

***Corresponding author:** Luciano Brocchieri, Department of Molecular Genetics and Microbiology and Genetics Institute, University of Florida, Gainesville, FL 32606, USA, Tel: 352.284.5414; E-mail: lucianob@ufl.edu

Received May 07, 2016; **Accepted** May 10, 2016; **Published** May 17, 2016

Citation: Brocchieri L (2016) Discovering Elusive Small Genes. J Phylogen Evolution Biol 4: e120. doi:10.4172/2329-9002.1000e120

of each expressed protein-coding sequence. The ribosome-profiling sequencing technology can provide deeper measurements and more accurate quantification than mass spectrometry proteomics, and can provide information not only on the amount of protein produced in given conditions but also on the dynamics of protein expression [15].

In contrast to computational gene prediction and to other experimental approaches, the identification power of ribosome profiling is independent on gene length, allowing detection of very short expressed genes and regulatory peptides (Figure 1). Furthermore, translation initiation sites (TIS) can be identified exploiting the activity of inhibitors stalling ribosomes at or proximal to TISs [16-18], a strategy also referred to as global translation initiation sequencing or GTI-seq [18]. Ribosome profiling provides the opportunity not only to identify precisely coding region, but also to uncover events and mechanisms of post-transcriptional control of protein expression in response to environmental stimuli. By genome-wide profiling, previously unrecognized widespread post-transcriptional regulation of gene translation and translational response to stress have been newly identified providing ample evidence of post-transcriptional regulation in eukaryotes (e.g., [10,14]). New fundamental biological processes have been discovered [19,20] and opportunities for biotechnological innovation have been identified [21]. It was demonstrated that in vertebrates the majority of expressed genes are associated with the translation of peptides encoded by uORFs in the 5'UTR or by internal out-of-frame ORFs (AltORFs) [10]. By bridging the gap between global measurements of mRNA and protein levels, ribosome profiling provides the most advanced tool for accurately and directly measuring levels of protein expression, and can provide the necessary information for building an optimal protein sequence search database for MS-based proteomics [22].

Although identification of small proteins by experimental data on expression and functionality is facilitated by the ever-growing availability of these high throughput genomic methods, it is unlikely that computational gene predictions will be soon superseded by experimental methods, whose sensitivity is limited by the necessity to identify the conditions required for expression of many genes. Integration of information from sequence features, conservation, and transcriptomic, translatomic, and proteomic analyses, will most likely provide the best strategy for obtaining the most complete picture of the coding potential of prokaryotic and eukaryotic organisms [23].

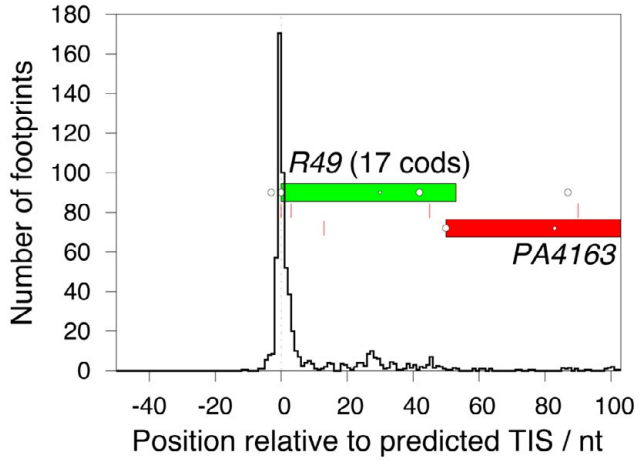

Figure 1: RIBO-seq evidence of expression of the newly identified gene R49, encoding a 17aa-long peptide 5' of annotated gene PA4163. Three rows of symbols represent potential start codons (circlets) and stop codons (vertical red bar) in the three reading frames of the strand encoding the represented gene, with the top row representing the frame of the gene. Bigger circlets indicate the position of canonical start codons ATG or GTG, and smaller circlets indicate alternative start codons TTG, CTG, or ATT.

Gene set	Total	Conserved
Annotated	3,531,708	3,239,662
Annotated ≤300 nt	390,233	235,413
New	889,837	182,294
New ≤300 nt	740,695	97,836

Table 1: Total number and conserved genes identified among annotated or newly-predicted genes in 1000 bacterial genomes [12].

Acknowledgment

This work is supported by NIH Grant 5R01GM87485-2.

References

1. Su M, Ling Y, Yu J, Wu J, Xiao J (2013) Small proteins: untapped area of potential biological importance. Front Genet 4: 286.
2. Storz G, Wolf YI, Ramamurthi KS (2014) Small proteins can no longer be ignored. Annu Rev Biochem 83: 753-777.
3. Gassel M, Möllenkamp T, Puppe W, Altendorf K (1999) The KdpF subunit is part of the K(+)-translocating Kdp complex of Escherichia coli and is responsible for stabilization of the complex in vitro. J Biol Chem 274: 37901-37907.
4. Barry C, Fichant G, Kalogeropoulos A, Quentin Y (1996) A computer filtering method to drive out tiny genes from the yeast genome. Yeast 12: 1163-1178.
5. Lazazzera BA, Solomon JM, Grossman AD (1997) An exported peptide functions intracellularly to contribute to cell density signaling in B. Subtilis Cell 89: 917-925.
6. Herbert E, Uhler M (1982) Biosynthesis of polyprotein precursors to regulatory peptides. Cell 30: 1-2.
7. Krieger DT (1983) Brain peptides: what, where, and why? Science 222: 975-985.
8. Calvo SE, Pagliarini DJ, Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci USA 106: 7507-7512.
9. Ye Y, Liang Y, Yu Q, Hu L, Li H, et al. (2015) Analysis of human upstream open reading frames and impact on gene expression. Hum Genet 134: 605-612.
10. Johnstone TG, Bazzini AA, Giraldez AJ (2016) Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J 35: 706-723.
11. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673-679.
12. Oden S, Brocchieri L (2015) Quantitative frame analysis and the annotation of GC-rich (and other) prokaryotic genomes. An application to Anaeromyxobacter dehalogenans. Bioinformatics 31: 3254-3261.
13. Wade JT, Grainger DC (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol 12: 647-653.
14. Smircich P, Eastman G, Bispo S, Duhagon MA, Guerra-Slompo EP, et al. (2015) Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi. BMC Genomics 16: 443.
15. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-223.
16. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789-802.
17. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, et al. (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res 22: 2208-2218.

18. Lee S, Liu B, Huang SX, Shen B, Qian SB (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci USA 109: E2424-E2432.

19. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell 147: 1295-1308.

20. Shalgi R, Hurt JA, Krykbaeva I, Taipale M, Lindquist S, et al. (2013) Widespread regulation of translation by elongation pausing in heat shock. Mol Cell 49: 439-452.

21. Thoreen CC, Chantranupong L, Keys HR, Wang T, Gray NS, et al. (2012) A unifying model for mTORC1-mediated regulation of mRNA translation. Nature 485: 109-113.

22. Crappé J, Ndah E, Koch A, Steyaert S, Gawron D, et al. (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res 43: e29.

23. Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, et al. (2011) Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res 21: 634-641.