Cell Systems
Previews

CellPress

Between noise and function: Toward a taxonomy of the non-canonical translatome

Zachary Ardern[1,*] and Md Hassan uz-Zaman[2,*]
[1]Parasites and Microbes Programme, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK
[2]Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, USA
*Correspondence: zachary.ardern@sanger.ac.uk (Z.A.), h.uzzaman@utexas.edu (M.H.u.-Z.)
https://doi.org/10.1016/j.cels.2023.04.004

Eukaryotic genomes are pervasively translated, but the properties of translated sequences outside of canonical genes are poorly understood. A new study in *Cell Systems* reveals a large translatome that is not under significant evolutionary constraint but is still an active part of diverse cellular systems.

Canonical genes are sections of DNA that are known to be translated into proteins and are typically well conserved between species. We are learning, however, that proteins can be produced from sequences outside of known genes. For both eukaryotic and prokaryotic genomes, most of the genome is transcribed under at least some environmental conditions, a phenomenon termed "pervasive transcription."[1] Growing evidence across taxa shows that open reading frames (ORFs) within these "non-canonical" transcripts are also translated.[1–3]

New protein-coding genes have arisen at many points over evolutionary history.[1] If these are genuinely novel, they must ultimately derive from non-coding sequences, and young genes should show some evidence of the processes involved in the transition. The discovery of many non-canonical translated ORFs, hypothesized to serve as "raw materials" for new genes,[1] brings together research themes in molecular and cellular biology with issues fundamental for evolutionary biology. Key questions raised include the biophysical properties of the proteins produced and whether the new ORFs should be classified as real genes, "noise," or something else.

A paper in this issue of *Cell Systems*[4] sheds light on these questions in the budding yeast *Saccharomyces cerevisiae*. Large-scale analyses of translation and selection are combined with experimental tests of knockout phenotypes for individual ORFs. Integrating analyses of translation and selection allows detection of translated ORFs and classifying them into two groups: "conserved," which are nearly all already annotated, and "evolutionarily transient." Some of these from each category partially or fully overlap known protein-coding genes, leading to multiple categories of ORFs (Figure 1A). Translation is ascertained using data pooled from 412 ribosome profiling (riboseq) experiments analyzed with a new approach termed "iRibo." Specifically, Wacholder et al. decide whether an ORF is translated or not using a binomial test for triplet periodicity in the riboseq data, with the p value threshold chosen using shuffled controls to obtain an estimated false discovery rate of 5%. This has the benefit of being a simple and clear approach that should minimize false positives if triplet periodicity is not an artifact of factors other than genuine translation. They discover a remarkable nearly 19,000 additional translated ORFs located either in between known genes or overlapping partially or fully in antisense to them (ORFs overlapping known genes on the same strand are excluded). These ~19,000 ORFs are then the subject of subsequent analyses.

What, then, is the status of these non-canonical, non-conserved, yet translated ORFs; are they functional genes that emerged recently, or are they merely the byproducts of stochastic translation, a kind of cellular noise? Wacholder et al. argue that a large part of the translatome cannot be categorized as either but requires a third concept. Their argument proceeds in two steps. First, by scanning these sequences for signs of purifying selection—a tell-tale indicator of functionality—they report that the vast majority of these elements are not under selection either within the genus or the species. As such, they are not likely to be retained in the genome as standard functional genes or their precursors but instead constitute a "transient" subset of the translatome. Second, by leveraging expression, genetic interaction, and fitness data, the authors show that a substantial fraction of these elements display a specific phenotype, suggesting they are more than inert consequences of stochastic translation. So, it is argued that these elements are neither genes nor noise, but rather constitute a third category of genomic elements (Figure 1B).

Could protein products formed merely as a consequence of cellular noise, like stochasticity in translation, have phenotypic consequences? In this connection, Sean Eddy[5] asks us to imagine the consequences of introducing a completely novel or random genome into a cell. It seems likely that not only would the stochastic nature of transcription lead to large parts of the genome being expressed, but their products would also have a measurable physiological impact following knockout. More recently, Weisman[6] introduced the concept of a "freeloader function" in the cell, in which the deeply interconnected cell interactome offers an environment where each component can potentially be affected merely through binding with something else, even through weak, non-specific interactions. Indeed, such interactions are likely to be common even among non-genic sequences, and it is not surprising that products of stochastic translation might participate in such processes. The precise distinction between noise and function within the cell is, therefore, hard to draw.

Assessing the fitness effects of ORF deletion is a reasonable starting point, as it seems unlikely that deleting products of cellular noise would be detrimental. Using an existing dataset of yeast gene interactions,[7] Wacholder et al. found that nearly all the 84 transient ORFs tested showed a deleterious fitness effect when

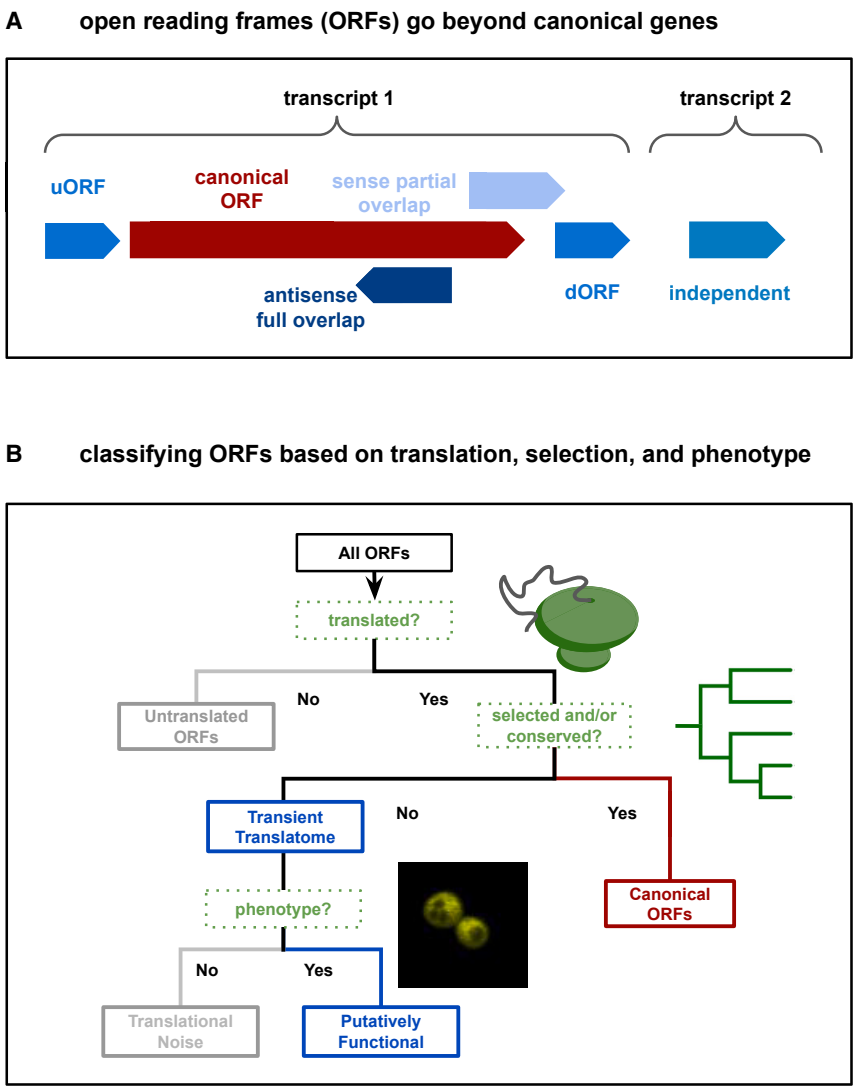

Figure 1. Classifying the translatome

(A) Classes of open reading frame (ORF) in terms of their genomic location with respect to canonical ORFs, i.e., annotated protein-coding genes. Non-canonical ORFs are shown in different shades of blue. uORF, upstream ORF; dORF, downstream ORF. Both sense and antisense overlapping ORFs can overlap either in full—embedded within the canonical ORF—or partially. Sense overlaps were excluded from the set of translated ORFs used by Wacholder et al. due to limitations of ribosome profiling data.

(B) Translated ORFs were characterized according to whether they are evolutionarily conserved and/or under purifying selection and whether they have evidence of phenotype.

The microscopy image is adapted from Wacholder et al.

another gene, inferred to be its interacting partner, was also deleted. The negative fitness effect of a double knockout was more pronounced than expected given the independent effects of two single knockouts. However, in Wacholder et al.'s single gene deletion screen, only a small fraction of highly translated genes (8 out of 49 ORFs tested) showed a deleterious fitness effect. Taken together, this suggests that the fitness effects of most translated ORFs may only be visible when their interacting partner gene is also deleted.

How do we explain, in evolutionary terms, this portrait of the transient translatome showing widespread phenotypic effects without corresponding purifying selection? Two options seem to be relevant. First, selection might be acting on these translation products in a somewhat sequence-independent way. It has been argued that newly emerged ORFs with beneficial fitness effects are enriched in transmembrane domains.[8] This suggests that adaptive benefits of proteins may often be mediated by "higher level" physicochemical properties compatible with substantial change in amino acid sequences. This pool of proteins may participate in a wide range of cellular roles facilitated by properties such as the presence of transmembrane domains. However, this specific hypothesis might not apply to other organisms, as *de novo* emerged proteins in humans do not seem to display similar domain-forming propensities.[9] The second possibility, suggested by the authors, is that the transient translatome is a consequence of translation functions unrelated to standard gene expression, such as regulation of neighboring genes or RNA decay. In this view, neither the sequence nor the general physicochemical properties of the translated products are under selection. If this is accurate, the fact that a large fraction of the translatome still participates in gene interaction networks would suggest that such capacities are easily accessible to expressed sequences and can manifest even without being shaped by selection—e.g., through weak, non-specific interactions with various cellular components. This, in turn, could either mean that genomic regions "sampled" by translation are enriched in such phenotypically relevant properties or that these properties are common among all possible sequences. Both of these factors could be relevant in explaining subsets of the transient translatome.

To better understand the "functions" of the transient translatome, future studies could explore the general phenotypic consequences of translation,[10] for example, by expressing sequences encoding random peptides in the cell, ideally in different genomic contexts. A related theme is to consider the regulation of the transient translatome. Cellular roles of protein products are determined not only by their particular physico-chemical properties but likely also by the ORFs' regulatory contexts. With the growing availability of protein-structure models, the biophysical properties of protein products are increasingly accessible. Further analysis of selection will also be useful; for instance, it might be expected that members of the transient translatome may be under positive rather than purifying selection, and delving into evolutionary

analyses further should be informative. This raises questions for cellular biology regarding the potential bioenergetic and fitness burden on the cell due to producing so many non-canonical protein products.

In summary, Wacholder et al. show that there are many unannotated ORFs that are translated but are neither canonical "genes" under strong purifying selection nor simply translational "noise" with no interesting phenotypic impact. Some of them may be newly emerging genes or the consequences of non-genic translation with impacts on the cellular environment. In the latter view, these elements serve as a large pool of raw materials for the evolution of new genes, potentially awaiting conservation through selection when their phenotypic consequences coincide with adaptive benefits for the cell. Taken together, these results not only highlight the underappreciated complexity of the cellular environment but also illuminate the process by which new genes emerge from non-genic sequences in the genome.

ACKNOWLEDGMENTS

Work conducted at the Wellcome Sanger Institute was supported by the Wellcome Trust, grant [220540/Z/20/A].

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Van Oss, S.B., and Carvunis, A.-R. (2019). De novo gene birth. PLoS Genet. *15*, e1008160.
2. Smith, C., Canestrari, J.G., Wang, A.J., Champion, M.M., Derbyshire, K.M., Gray, T.A., and Wade, J.T. (2022). Pervasive translation in Mycobacterium tuberculosis. Elife *11*, e73980. https://doi.org/10.7554/eLife.73980.
3. Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S., and Neuhaus, K. (2020). Evidence for Numerous Embedded Antisense Overlapping Genes in Diverse E. coli Strains. bioRxiv. https://doi.org/10.1101/2020.11.18.388249.
4. Wacholder, A., Parikh, S.B., Coelho, N.C., Acar, O., Houghton, C., Chou, L., and Carvunis, A.-R. (2023). A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Systems *14*, 363–381.
5. Eddy, S.R. (2013). The ENCODE project: missteps overshadowing a success. Curr. Biol. *23*, R259–R261.
6. Weisman, C.M. (2022). The Origins and Functions of De Novo Genes: Against All Odds? J. Mol. Evol. *90*, 244–257.
7. Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420. https://doi.org/10.1126/science.aaf1420.
8. Vakirlis, N., Acar, O., Hsu, B., Coelho, N.C., Van Oss, S.B., Wacholder, A., Medetgul-Ernar, K., Bowman, R.W., 2nd, Hines, C.P., Iannotta, J., et al. (2021). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat. Commun. *12*, 200.
9. Vakirlis, N., Vance, Z., Duggan, K.M., and McLysaght, A. (2022). De novo birth of functional microproteins in the human lineage. Cell Rep. *41*, 111808.
10. Luthra, I., Chen, X.E., Jensen, C., Rafi, A.M., Salaudeen, A.L., and de Boer, C.G. (2022). Biochemical activity is the default DNA state in eukaryotes. bioRxiv. https://doi.org/10.1101/2022.12.16.520785.