eLIFE elifesciences.org

INSIGHT

GENETIC NOVELTY

How new genes are born

Analysis of yeast, fly and human genomes suggests that sequence divergence is not the main source of orphan genes.

URMINDER SINGH AND EVE SYRKIN WURTELE

Related research article Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**:e53500. DOI: 10.7554/eLife.53500

For half a century, most scientists believed that new protein-coding genes arise as a result of mutations in existing protein-coding genes. It was considered impossible for anything as complex as a functional new protein to arise from scratch. However, every species has certain genes, known as 'orphan genes', which code for proteins that are not homologous to proteins found in any other species. What do these orphan genes do, and how are they formed?

To date the roles of hundreds of orphan genes have been characterized. Although this is just a tiny fraction of the total, it is known that most of them code for proteins that bind to conserved proteins such as transcription factors or receptors. Some of these proteins are toxins, some are involved in reproduction, some integrate into existing metabolic and regulatory networks, and some confer resistance to stress (***Carvunis et al., 2012***; ***Li et al., 2009***; ***Xiao et al., 2009***; ***Arendsee et al., 2014***; ***Belcaid et al., 2019***). However, none of them are enzymes (***Arendsee et al., 2014***). Orphan genes arise quickly, so they may provide a disruptive mechanism that allows a given species to survive changes to its environment. Thus, the study of how orphan genes arise (and fall) is central to understanding the forces that drive evolution (***Figure 1***).

One possible mechanism is the 'de novo' appearance of a gene from an intergenic region or a completely new reading frame within an existing gene (***Tautz and Domazet-Lošo, 2011***). An alternative mechanism is that the coding sequence of the orphan gene arises by rapid divergence from the coding sequence of a pre-existing gene: this would mean that an entire set of regulatory and structural elements would be available to the gene as it evolves. Now, in eLife, Nikolaos Vakirlis and Aoife McLysaght (both from Trinity College Dublin) and Anne-Ruxandra Carvunis (University of Pittsburgh) report how they have studied yeast, fly and human genes to compare the contributions of these two mechanisms to the emergence of orphan genes (***Vakirlis et al., 2020***).

Previous studies have used simulations to estimate the number of orphan genes that appear by divergence; until now, no one had relied on actual genomics data to study this phenomenon. Vakirlis et al. use a new approach to analyze orphan genes that have originated through divergence. They examine regions of the genome that correspond to each other (so-called syntenic regions) in related species to determine whether a gene exists in both regions and, if so, whether the proteins are non-homologous. If the genes have no homology, they may have originated by rapid divergence from the coding sequence of a preexisting gene.

Using this method, Vakirlis et al. infer that at most 45% of *S. cerevisiae* (yeast) orphan genes, 25% of *D. melanogaster* (fruit fly) orphan genes, and 18% of human orphan genes arose by rapid divergence, but this is an upper estimate. For example, it is possible that a new coding

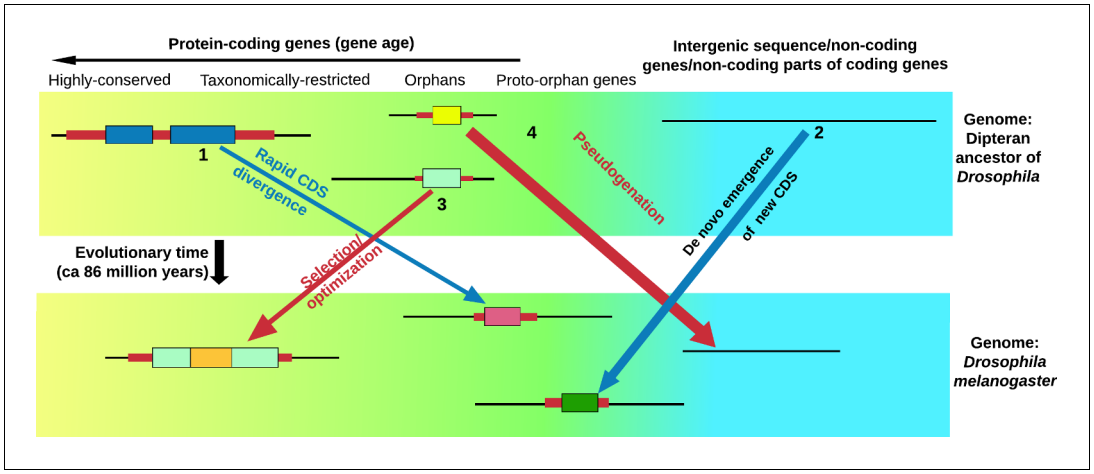

Figure 1. Life cycle of orphan genes. Every species has orphan genes that have no homologs in other species. This schematic shows the genome of the fruit fly (bottom) and the genome of an ancestor of the fruit fly (top). Each panel also shows (from left to right): genes that are highly conserved and can be traced back to prokaryotic organisms (yellow background); genes that are found in just a few related species (taxonomically restricted genes), orphan genes and potential orphan genes that are not currently expressed and are thus free from selection pressure (proto-orphan genes); and regions of the genome that do not code for proteins (blue background) (***Van Oss and Carvunis, 2019***; ***Palmieri et al., 2014***). An orphan gene can form through the rapid divergence of the coding sequence (CDS) of an existing gene (1), or arise de novo from regions of the genome that do not code for proteins (including the non-coding parts of genes that evolve to code for proteins; 2). Some orphan genes will be important for survival, and will thus be selected for and gradually optimized (3). This means that the genes in a single organism will have a gradient of ages (***Tautz and Domazet-Lošo, 2011***). Many proto-orphan genes will undergo pseudogenation (that is, they will not be retained; 4). Coding sequences (shown as thick colored bars) with detectable homology are shown in similar colors. Vakirlis et al. estimate that a minority of orphan genes have arisen by divergence of the coding sequence of existing genes.

sequence might have arisen de novo within an existing gene, rather than the existing coding sequence having been modified beyond recognition.

But how can a protein sequence continue to be selected for as it rapidly diverges? Vakirlis et al. suggest that divergence might occur by a process of partial pseudogenation: the existing gene becomes non-functional, and then, with no selection pressure to retain the old protein, it diverges to form an orphan gene.

Many orphan genes may not have been identified yet, because they do not have homologs in other species, and have few recognizable sequence features. This means that up to 80% of orphan genes can be missed when a new genome is annotated (***Seetharam et al., 2019***). The approach detailed by Vakirlis, Carvunis and McLysaght evaluates specifically those annotated orphan genes for which a similar gene exists in a related species (which is ~50% of them; ***Arendsee et al., 2019***). As high-quality genomes from more species become available, and as more orphan genes are annotated, the approach will provide yet deeper insights into the origin of these genes.

One of the many open questions in this field deals with genes of 'mixed age'. Some such genes have incorporated 'chunks' of orphans into their coding sequences. A gene that has done this is (somewhat arbitrarily) considered to be the age of its most ancient segment, but we know little about the mechanism of this process or its significance. Another question involves the unique strategies and rates of evolution of each gene (***Revell et al., 2018***). How might the abundance and mechanisms of orphan gene origin vary among species? And how do different environments affect the emergence of orphan genes?

Urminder Singh is in the Department of Genetics, Developmental and Cell Biology, Iowa State University, Ames, United States
https://orcid.org/0000-0003-3703-0820
Eve Syrkin Wurtele is in the Department of Genetics, Developmental and Cell Biology, Iowa State University, Ames, United States
mash@iastate.edu
https://orcid.org/0000-0003-1552-9495

Competing interests: The authors declare that no competing interests exist.

Published 19 February 2020

References

Arendsee ZW, Li L, Wurtele ES. 2014. Coming of age: orphan genes in plants. *Trends in Plant Science* **19**: 698–708. DOI: https://doi.org/10.1016/j.tplants.2014.07.003, PMID: 25151064

Arendsee Z, Li J, Singh U, Seetharam A, Dorman K, Wurtele ES. 2019. Phylostratr: a framework for phylostratigraphy. *Bioinformatics* **35**:3617–3627. DOI: https://doi.org/10.1093/bioinformatics/btz171, PMID: 30873536

Belcaid M, Casaburi G, McAnulty SJ, Schmidbaur H, Suria AM, Moriano-Gutierrez S, Pankey MS, Oakley TH, Kremer N, Koch EJ, Collins AJ, Nguyen H, Lek S, Goncharenko-Foster I, Minx P, Sodergren E, Weinstock G, Rokhsar DS, McFall-Ngai M, Simakov O, et al. 2019. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. *PNAS* **116**:3030–3035. DOI: https://doi.org/10.1073/pnas.1817322116, PMID: 30635418

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M. 2012. Proto-genes and de novo gene birth. *Nature* **487**:370–374. DOI: https://doi.org/10.1038/nature11184, PMID: 22722833

Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES. 2009. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *The Plant Journal* **58**:485–498. DOI: https://doi.org/10.1111/j.1365-313X.2009.03793.x, PMID: 19154206

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. *eLife* **3**:e01311. DOI: https://doi.org/10.7554/eLife.01311, PMID: 24554240

Revell LJ, González-Valenzuela LE, Alfonso A, Castellanos-García LA, Guarnizo CE, Crawford AJ. 2018. Comparing evolutionary rates between trees, clades and traits. *Methods in Ecology and Evolution* **9**: 994–1005. DOI: https://doi.org/10.1111/2041-210X.12977

Seetharam AS, Singh U, Li J, Bhandary P, Arendsee Z, Wurtele ES. 2019. Maximizing prediction of orphan genes in assembled genomes. *bioRxiv*. DOI: https://doi.org/10.1101/2019.12.17.880294

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**: 692–702. DOI: https://doi.org/10.1038/nrg3053, PMID: 21878963

Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**:e53500. DOI: https://doi.org/10.7554/eLife.53500

Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLOS Genetics* **15**:e1008160. DOI: https://doi.org/10.1371/journal.pgen.1008160, PMID: 31120894

Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S. 2009. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLOS ONE* **4**: e4603. DOI: https://doi.org/10.1371/journal.pone.0004603, PMID: 19240804