The Enigma of Taxonomically Restricted Essential Genes of Unknown Function

Change Laura Tan, Division of Biological Sciences, University of Missouri-Columbia[1]
and Paul A. Nelson, Biola University[2]

1. 105 Tucker Hall, Columbia MO 65201, USA, ct.columbia.mo.65203@gmail.com
2. 13800 Biola Avenue, La Mirada CA 90639, USA, paul.alfredp@gmail.com

ABSTRACT: The advent of automated DNA sequencing in the mid-1990s enabled the development shortly thereafter of transposon mutagenesis screens of entire genomes, to determine what genes (and their protein or functional RNA products) were essential for cell viability under laboratory conditions. Two decades later, these experiments, as well as systematic or targeted gene deletion experiments, have consistently returned an unexpected result: many essential genes are taxonomically restricted in their distribution, not universally, or even widely, shared. Moreover, when first annotated, these genes are most often classified as "unknown function." Evident already in the initial transposon mutagenesis screens of *Mycoplasma genitalium* (Hutchison *et al.* 1999), where approximately one third of the essential genes were listed as "unknown function," the same "unique, essential, and unknown function" (UEU) signal has been found in all three domains of life. We have begun a project to collect and characterize these UEU sequences, to assess their impact on our understanding of theories of cell function and evolution.

BACKGROUND: From its very beginning, whole-genome sequencing has revealed the presence of taxon-specific essential genes of unknown function. In 1995, with the publication of the first complete genome, *Mycoplasma genitalium*, Fraser *et al.* (1995) noted that a significant fraction (96/470) of the predicted ORFs showed no homology with archived sequences. Four years later, using *M. genitalium* as their model system, Hutchison *et al.* (1999) found via transposon mutagenesis that an estimated 1/3 of their essential gene set were "proteins of unknown function." They observed that the "presence of so many genes of unknown function among the essential genes of the simplest known cell suggests that all the basic molecular mechanisms underlying cellular life may not yet have been described."

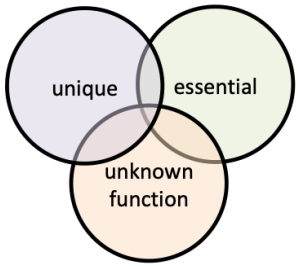

Fig. 1. The intersection of uniqueness, essentiality, and unknown function for any candidate ORF promises to yield biological findings of great interest.

Thus the intrinsic biological interest of these "unique, essential, unknown function" (UEU) genes, and their (protein or functional RNA) products, was immediately clear – not only for molecular and cellular biology, but for theories of evolution as well. We have begun a systematic compilation of UEU sequences, now found in a wide range of species, with two main questions in mind:

1. What might UEU genes and their products tell us about the adequacy or completeness of our descriptions of cell function?

2. What might UEU sequences tell us about current theories of evolution? In particular, do we have causally sufficient models for transitions between what might be called the "different operating systems" (namely, UEU complexes) for species?

The second question gains its importance from the highly counterintuitive nature of UEU genes. ***Essentiality*** and ***phylogenetic conservation*** tend to go hand-in-hand analytically within evolutionary theory. By contrast, ***uniqueness*** (i.e., taxonomic restriction to the ranks of genera and species), especially when linked with functional essentiality or criticality, is unexpected. As a first step in gaining a global perspective on possible UEU sequences, therefore, we surveyed the distribution of orphan genes in 317 representative species (with whole-genome data), drawn from all three domains of life.

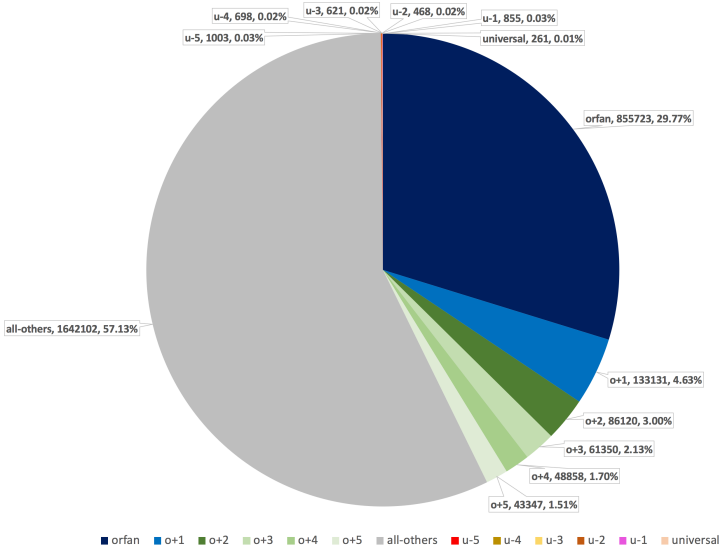

Fig. 2. Distribution of universally conserved and orphan genes.

DISTRIBUTION OF UNIVERSAL VS. ORPHAN (POSSIBLE UEU) GENES

Method: Binary phylogenetic profiles recording the presence or absence of gene orthologs within a representative set of 317 organisms (144 Eukarya, 142 Bacteria and 31 Archaea) were obtained from Lecompte via Nevers (Nevers *et al.*, 2019). Numbers of genes with 0 to 5, or 311 to 316 orthologs in the 317 organisms (except self) were calculated for each of the species and then summed. Genes with no orthologs are species-specific and labeled as **orfans** (Fig. 2) while genes with orthologs in all species are designated as **universal.** Genes with 1 to 5 orthologs are referred to as o+1 to o+5. Genes with orthologs in all species, less 1 to 5, are referred to as u-1 to u-5. Genes with orthologs in 6 to 310 species are collected in the "all-others" group. For each group, the number and percentile of genes belonging to that group are shown. Perhaps surprisingly, orfans and near-orfans (o+1 to o+5) outnumber the small fraction of universal and near-universal genes (u-1 to u-5).

DISCUSSION

Until the discovery of orphan genes in the late 1990s, few biologists imagined the extent to which species-unique genes occurred within Earth's biota. Our preliminary survey of 317 representative species shows that UEU sequences may be ubiquitous. Space limitations on this poster prevent us from providing examples (drawn from the current literature) of specific UEUs, but the enormous dark blue "slice of the sequence universe pie" (Fig. 2) contains many surprises we would be happy to discuss. The next step in our analysis will be a systematic catalog of UEU sequences, within the ongoing bioinformatics projects we have designated ORFanID and ORFanBase (Gunasekera *et al.* 2018).

REFERENCES

Fraser, C. *et al.* 1995. The Minimal Gene Complement of *Mycoplasma genitalium*. *Science* 270:397-403.
Gunasekera, R. *et al.* 2018. ORFanID: A Web-Based Computational Algorithm for Taxonomically Restricted Genes. ISMB 2018, Chicago IL.
Hutchison, C. *et al.* 1999. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science* 286:2165-2169.
Nevers, Y., Kress, A., Defosset, A., Ripp, R., Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. 2019. OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Research* 47:D411-D418.

ACKNOWLEDGMENT: We are very grateful for the thoughtful assistance of Andrew Jones with computational aspects of this project.