The Origin of Novel Genes

Richard Buggs

Alexander Bowles, Ulrike Bechtold, and Jordi Paps, "The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty," *Current Biology* 30, no. 3 (2020): 530–36.e2, doi:10.1016/j.cub.2019.11.090.

Cristina Guijarro-Clarke, Peter Holland, and Jordi Paps, "Widespread Patterns of Gene Loss in the Evolution of the Animal Kingdom," *Nature Ecology and Evolution* 4 (2020): 519–23, doi:10.1038/s41559-020-1129-2.

Jordi Paps and Peter Holland, "Reconstruction of the Ancestral Metazoan Genome Reveals an Increase in Genomic Novelty," *Nature Communications* 9, no. 1,730 (2018), doi:10.1038/s41467-018-04136-5.

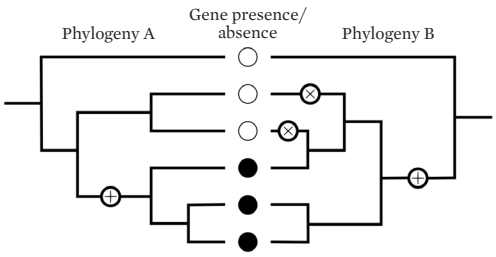
Thomas Dunwell, Jordi Paps, and Peter Holland, "Novel and Divergent Genes in the Evolution of Placental Mammals," *Proceedings of the Royal Society B: Biological Sciences* 284, no. 1,864 (2017), doi:10.1098/rspb.2017.1357.

A series of papers from Peter Holland's lab at the University of Oxford and the lab of his former postdoc Jordi Paps at the University of Bristol investigate patterns of gene presence and absence in plants and animals. These patterns are described in terms of gene gains and losses within a bifurcating phylogeny whose topology is derived from other sources. The authors make the assumption that each gene can be gained only once, but can be lost multiple times.

The four studies find that organisms with different morphologies possess different sets of genes. Given that genes provide much of the information encoding the morphology of living organisms, this finding may not seem a surprise. That novel genes do not accumulate with Darwinian gradualism in the phylogeny is perhaps more surprising. The authors describe bursts of innovation: upon the origin of placental mammals, 357 novel genes; upon the origin of the metazoan, 1,189 novel genes; upon the origin of the land plants, 1,167 novel genes; and upon the origin of the flowering plants, 2,525 novel genes.

Equally surprising is evidence that the patterns of presence and absence of many genes in these studies do not form a nested hierarchy congruent with the accepted phylogeny. Particular genes often appear in more than one clade (Figure 1). This leads the authors to infer massive gene losses and frequent horizontal gene transfer in the history of life.

Figure 1.



In Phylogeny A, one gene gain event is needed to explain the presence of a particular gene in three species, forming a neatly nested hierarchy. If Phylogeny B is assumed to be true, one gain and two loss events are needed to explain the pattern of gene presence or absence.

The unexpected nature of these findings was not lost on the authors of the studies, nor the editors of the journals that published their manuscripts. Three of the paper titles emphasize unexpected novelty and one emphasizes unexpected loss. But all four show similar patterns. More is revealed in each than a single title can convey.

In this context, the gene does not embody solely unique protein-coding sequences, nor groups of slightly different protein-coding sequences, but larger cohesive clusters that the authors term homology groups. These are sets of protein sequences found using a Markov cluster algorithm applied to a protein–protein similarity graph. According to the parameters applied by the authors, the homology groups in question are typically larger groupings than a gene family.

That is not to say that there is no detectable similarity among homology groups. In their study, Thomas Dunwell

et al. examined 87 homology groups found exclusively in nine or more of ten placental mammal species.[1] Searches were performed with low stringency based on amino-acid translations of the DNA sequences (BLASTP). Of the 87 homology groups examined, 15 exhibited detectable similarity with other homology groups in mammals and 39 with other homology groups in animals. This left 33 of the 87 homology groups with no detectable similarity to any other groups in their study. Even if all proteins that have any detectable similarity with BLASTP were joined together, the resulting sequence space would resemble an archipelago, rather than a continent.

It is considered obvious among biologists that natural groups of proteins exist and can be found using clustering algorithms. It is worth pausing to note this granularity in the protein-space of life. In the 1850s, Charles Darwin considered it obvious that the morphological variation of life was continuous: "all the parts and organs of many independent beings" are "linked together by graduated steps."[2] The largest of these four studies included more than nine million protein-coding sequences from 208 genomes, spanning eukaryotic life from yeast to humans to ash trees.[3] Such a sample might be expected to show continuous variation. Instead, these nine million sequences clustered into 661,545 homology groups.

It is far from clear how these homology groups might be linked in graduated steps. The evolution of novel genes is a subject with a substantial literature all its own, which has recently shifted from the view that all new genes begin as duplicates of pre-existing genes to a view that many genes evolve de novo from noncoding sequences.[4] The mechanisms underlying this process are not well understood.

The central question tackled by the teams of Holland and Paps is not the mechanism by which each individual homology group originates, but the patterns with which they appear and disappear in the history of life. Rather than emerging gradually, a few at a time, the evidence presented in these four papers suggests the occurrence of punctuated bursts. At every major phylogenetic node that was examined, the appearance of hundreds, and in some cases thousands, of novel homology groups was detected.

Evolution by bursts is, of course, not expected if natural selection is the main driver. "[N]atural selection acts only by taking advantage of slight successive variations," Darwin remarked; "she can never take a great and sudden leap, but must advance by the short and sure, though slow steps."[5] The findings presented in these papers suggest otherwise. It seems that the evolution of life is characterized by leaps involving large numbers of novel homology groups.

In future studies encompassing more species, it may well be the case that such bursts of novelty appear smaller. The largest of the four studies by Holland, Paps, et al. includes only 208 species. The addition of further species with sister-group relationships to major clades will undoubtedly ameliorate the size of evolutionary transitions. This would introduce additional nodes to which the origin of novel homology groups may be mapped.

If a sister lineage to all extant angiosperms, *Amborella*, had been omitted from the study by Alexander Bowles et al., the number of novel homology groups at their base may well have exceeded 3,000.[6] The inclusion of *Amborella* allowed this transition to be graduated into two nodes, one with 713 novel homology groups and another with 2,525. If two orders of flowering plants, Nymphaeales and Austrobaileyales, and a further group, Magnoliids, had also been included, three further nodes would have been introduced, allowing for further graduation. Assuming current phylogenies are correct, the addition of gradation would stop at this point. There are no other known extant groups with sister relationships to the other angiosperms. Each node would still have, on average, hundreds of novel homology groups. The leaps would become slightly smaller, but there would still not be anything resembling short steps.

The fossil record depicts the appearance of the first angiosperms as a sudden event, with no clear progenitors. This was known, in part, to Darwin, who famously complained to the director of Kew Gardens in 1879 that the origin of the dicotyledonous angiosperms was an "abominable mystery."[7] The mystery has since deepened to include all other angiosperms.[8] It is not until the Cretaceous period that angiosperms first appear in the fossil record, and by the Late Cretaceous many examples can be found that closely resemble modern taxa.[9] Taken at face value, the fossil record does not appear to allow sufficient time for the accumulation of angiosperm-specific homology groups. Bowles et al. suggest that whole-genome duplication at the base of the angiosperms could account for their origin, but this would require remarkably rapid divergence of identical duplicates into new homology groups. Although this is a more credible explanation than thousands of de novo genes, the mystery endures.

The nodes at the origin of the angiosperms are certainly striking in terms of the total number of novel genes that seem to have appeared in a short space of time. But Bowles et al. regard the node at the origin of all land plants as even more significant. At this node, 103 genes originate that are preserved in all descendent lineages—with the possible exception of a single species. It is not unreasonable to speculate that these genes are essential to being a land plant. To what extent they all had to be in place before land-based plant life became viable is an open question.

These studies by the teams of Holland and Paps are not alone in finding bursts of novel genes in the history of life. In a paper published earlier this year, Zhang et al. conducted an analysis of plants similar to Bowles et al., with better sampling of charophytes and bryophytes.[10] Despite using different gene clustering methods and a smaller set

of species, they found gene gains at key nodes on similar orders of magnitude.

All four studies under review found massive gene losses for phylogenetic nodes at the base of the major groups of living organisms. This suggests that major evolutionary transitions do not occur solely by means of tinkering with existing genes. Instead, it seems that vast numbers of existing genes are jettisoned and replaced by entirely different ones. Such processes would represent a radical overhaul in the genetic composition of organisms. How this might be accomplished is another mystery.

Losses are inferred by the authors when homology groups are present in more than one major group, but these groups are less closely related than they are to other groups. If the starting phylogenies are topologically correct, the homology group does not fit neatly into a nested hierarchy of similarity—or as Darwin himself put it, "the grand fact of the natural subordination of organic beings in groups under groups."[11]

It could be that the authors have overestimated the rates of gene loss because, although widely accepted, their starting phylogenies are wrong. It would be interesting to examine whether phylogenies built on the presence or absence of homology groups differed from these accepted phylogenies. In this case, homology group gains and losses mapped to major nodes might be reduced. Such reductions would likely be small, given the multiple contrasting patterns of gain and loss shown by homology groups.

The authors suggest that horizontal gene transfer could explain the incongruent patterns of gene presence or absence that give rise to some of the apparent losses. Bowles et al. found that 323 homology groups were present in fungal and land plant genomes, but absent from all other taxa.[12] Instead of being lost in the lineages between fungi and land plants, the genes could simply have jumped. This may turn out to be a more elegant solution to the problem.

The incongruence between patterns in the absence or presence of homology groups and widely accepted phylogenies raises a broader issue. A single phylogeny is clearly an inadequate model for the history of life, but there is no obvious replacement. This question is wide open.

Richard Buggs is a Senior Research Leader at Royal Botanic Gardens, Kew, and Professor of Evolutionary Genomics at Queen Mary University of London.

1. Thomas Dunwell, Jordi Paps, and Peter Holland, "Novel and Divergent Genes in the Evolution of Placental Mammals," *Proceedings of the Royal Society B: Biological Sciences* 284, no. 1,864 (2017), doi:10.1098/rspb.2017.1357.
2. Charles Darwin, *On the Origin of Species by Means of Natural Selection: Or, The Preservation of Favoured Races in the Struggle for Life*, 6th ed. (London: John Murray, 1872), 156.
3. Alexander Bowles, Ulrike Bechtold, and Jordi Paps, "The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty," *Current Biology* 30, no. 3 (2020): 530–36.e2, doi:10.1016/j.cub.2019.11.090.
4. See, for example: Anne-Ruxandra Carvunis et al., "Proto-Genes and *De Novo* Gene Birth," *Nature* 487 (2012): 370–74, doi:10.1038/nature11184; Aoife McLysaght and Daniele Guerzoni, "New Genes from Non-Coding Sequence: The Role of De Novo Protein-Coding Genes in Eukaryotic Evolutionary Innovation," *Philosophical Transactions of the Royal Society B: Biological Sciences* 370, no. 1,678 (2015), doi:10.1098/rstb.2014.0332; Nikolaos Vakirlis, Anne-Ruxandra Carvunis, and Aoife McLysaght, "Synteny-Based Analyses Indicate that Sequence Divergence Is Not the Main Source of Orphan Genes," *eLife* 9 (2020), doi:10.7554/eLife.53500.sa2.
5. Darwin, *On the Origin of Species*, 156.
6. Bowles, Bechtold, and Paps, "The Origin of Land Plants."
7. Francis Darwin and Albert Seward, eds., *More Letters of Charles Darwin. A Record of His Work in a Series of Hitherto Unpublished Letters* (London: John Murray, 1903), 20–21.
8. Richard Buggs, "The Deepening of Darwin's Abominable Mystery," *Nature Ecology & Evolution* 1, no. 0169 (2017), doi:10.1038/s41559-017-0169.
9. Patrick Herendeen et al., "Palaeobotanical Redux: Revisiting the Age of the Angiosperms," *Nature Plants* 3, no. 17,015 (2017), doi:10.1038/nplants.2017.15.
10. Jian Zhang et al., "The Hornwort Genome and Early Land Plant Evolution," *Nature Plants* 6 (2020): 107–18, doi:10.1038/s41477-019-0588-4.
11. Darwin, *On the Origin of Species*, 364.
12. Bowles, Bechtold, and Paps, "The Origin of Land Plants."

Published on September 28, 2020

https://inference-review.com/article/the-origin-of-novel-genes