CellPress
REVIEWS

Review

New Gene Origin and Deep Taxon Phylogenomics: Opportunities and Challenges

Christian Rödelsperger,[1] Neel Prabh,[1] and Ralf J. Sommer[1,*]

The immense morphological and phenotypic diversity within eukaryotes coincides with large-scale differences in genic repertoires, including the presence of thousands of new genes in every genome. New genes arise through duplication and divergence of existing coding sequences or *de novo* from noncoding sequences. These processes together cause individual genomes to contain up to one-third of orphan genes without any detectable homology in other lineages. Recently, deep taxon phylogenomics, the genome comparisons of extremely closely related species, provided novel insight into the evolutionary dynamics of such rapidly evolving genes. This review focuses on deep taxon phylogenomics and its importance in studying the evolution of new genes and discusses challenges and opportunities.

Highlights

Deeply sampled phylogenomic data sets have been established in yeasts, vertebrates, insects, nematodes, and plants.

These data form the basis to study the age, origin, and evolutionary dynamics of new genes.

New genes can arise ancestrally from protein-coding genes by duplication and divergence, or alternatively *de novo* from previously noncoding parts of the genome.

Ribosome-profiling data shows that translation alone is not sufficient to define protein-coding genes.

Despite abundant evidence of *de novo* genes, their evolutionary stability must be considered a rare event.

New Genes Can Arise from Ancestrally Coding and Noncoding Regions

Genetic studies in multiple species have demonstrated that biological innovation is frequently coupled with the evolution of new genes [1–5]. But how new genes arise in the first place remains controversial and is becoming a major research topic in evolutionary biology. Controversy results from the fact that there are multiple potential mechanisms of origin of new genes. Proper definitions of 'new', 'taxonomically-restricted', 'orphan', and '*de novo*' genes are therefore crucial, even if some of these terms can overlap (see Figure I in Box 1). New genes are genes that emerged recently in a given lineage (see Figure I in Box 1). In principle, each genome can be divided into regions that encode for proteins and regions that do not. Therefore, the primary distinction between different types of new genes is whether they originate from ancestrally **protein-coding sequences** (see Glossary) or noncoding genomic sequences [6,7]. While duplication of protein-coding genes is certainly one of the major forces generating new genes, the origin of **orphan genes**, which lack traceable homologs in other lineages, is much less understood. It was previously considered that duplication-divergence would be the only mechanism to produce orphan genes out of previously protein-coding material [8], but multiple studies have shown that divergence alone, exaptation of transposable elements, strand-switching, and resurrection of **pseudogenes** can create orphan genes without the need for duplication [9–12] In prokaryotes, new genes can also arise frequently by horizontal gene transfer (HGT) [13]. HGT can be considered as a special case of gene duplication, where a copy of a foreign gene is integrated into the host genome. In contrast, HGT in eukaryotic genomes can only explain a meager fraction of new genes despite numerous well-documented reports [14–16].

***De novo* genes** make up the most controversial class of new genes. Even though the idea of gene emergence through *de novo* creation from noncoding sequence was raised in the first half of the 20th century [17], duplication was considered the only viable mechanism for gene birth until the end of the last decade. The chief proponent for the exclusivity of the duplication mechanism, Susumu Ohno, claimed in his renowned book *Evolution by Gene Duplication* that due to the relentless pressure of natural selection, only a redundant cistron is unfettered to emerge as a new gene locus [18]. The stance of stalwarts such as Ohno and Francois Jacob against the *de novo* gene origin [19], coupled with the enormous difficulty of finding evidence supporting such mechanisms, had stifled investigation into gene origin apart from duplication. It took decades until the sheer abundance of orphan genes (Box 1) in eukaryotic genomes and the first evidence of *de novo* genes enabled researchers to seriously challenge this view [7,8,20,21]. Recently, however, *de novo* gene birth has been intensively discussed [8,22–25]. In this review, we focus on the importance of **deep taxon phylogenomics** in studying new gene origin and their evolutionary dynamics and we will discuss associated opportunities and challenges.

[1]Department for Integrative Evolutionary Biology, Max Planck Institute for Developmental Biology, Max Planck Ring 9, 72076 Tübingen, Germany

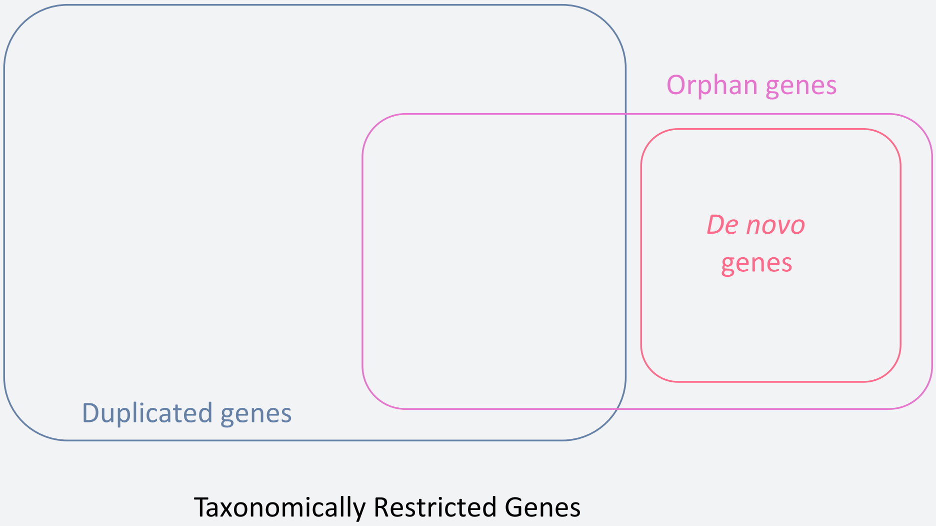
*Correspondence:
ralf.sommer@tuebingen.mpg.de

https://doi.org/10.1016/j.tig.2019.08.007

Box 1. What Are New Genes?

New genes are genes that, based on sequence comparison, appear to have emerged recently in a given lineage. By definition, they are taxonomically restricted or even species-specific.

Importantly, gene duplication events can also be taxonomically restricted and thus, we also consider products of duplication events as new genes. Since taxonomically restricted genes can be the result of gene loss, a comprehensive analysis of the phylogenetic context (e.g., **phylostratigraphic analysis**) is needed to distinguish losses from the emergence of new genes. New genes that lack detectable sequence homology (e.g., BLASTP e-value < 0.001) in other taxa are commonly referred to as orphan genes. The definition of orphan genes is always context-dependent and, in sparsely sampled taxonomic clades, orphan genes make up to one-third of a gene set in a given genome [8]. Conversely, some orphan genes may be relatively old and the lack of detected homology may be a result of strong divergence. Such firm divergence can create entirely new sequences and such genes may act as the conduits of evolutionary innovation and arguably should be treated as new genes. Whether orphan genes are just a product of extensive divergence from ancestrally protein-coding genes or arose *de novo* from noncoding sequences is one of the most actively studied questions in evolutionary biology. While none of these different gene classes is absolutely identical to the other (Figure I), genes in all of these categories tend to show similar properties, such as spatiotemporally restricted expression, the high propensity of being lost, and relaxed evolutionary constraint; we consider all the different types of taxonomically restricted genes (duplicated genes, orphan genes, and *de novo* genes) as new genes.



Trends in Genetics

Figure I. Different Classes of New Genes.

New genes arose recently in an evolutionary lineage and are consequently taxonomically restricted. Out of all taxonomically restricted genes, we consider duplicated genes, orphan genes, and *de novo* genes as new genes.

Glossary

Deep taxon phylogenomics: genome comparisons of closely related species with a well-defined molecular phylogeny.
***De novo* genes**: genes that emerged from an ancestrally noncoding sequence (intergenic, intronic, or noncoding RNAs).
Orphan genes: a gene without detectable protein homology outside a predefined taxonomic clade.
Pervasive translation: the idea that large parts of the genome are transcribed and translated at a background level.
Phylogenomics: genomic analysis in a given phylogenetic framework.
Phylostratigraphic analysis: an approach to determine the age of a given sequence by tracing the founding member in a species tree based on BLAST searches in extant taxa.
Protein-coding sequence: exonic sequences that are translated into peptides.
Proto-genes: products of **pervasive translation** with some genic properties but without any biological function.
Pseudogenes: genes that have lost functions due to mutations in their coding sequence (e.g., stop codons, frame shifts).

Deep Taxon Phylogenomics Is Indispensable for Understanding New Gene Origin

Since the first sequencing projects, eukaryotic genomes were known to contain substantial fractions of orphan genes, but whether orphan genes are the result of strong divergence or *de novo* origin is still debated. Historically, two complementary approaches have been used to identify the origin of orphan genes: bioinformatic methods for remote homology detection [26,27] and **phylogenomic** approaches, including phylostratigraphy [28–32]. While both approaches are widely used, it is important to note that they have distinct capabilities to investigate different classes of new genes. Methods like PSI-blast [26] or HHsearch [27] can detect distantly related proteins, which are missed by BLASTP, thus leading to initial classification as orphan genes. Such approaches work best for large orphan gene families, where many sequences are available to construct generalized sequence profiles. However, methods for remote homology detection are better suited to identify extensive divergence rather than *de novo* origin, as orthologous noncoding nucleotide sequences get scrambled beyond recognition at deeper evolutionary time scales.

Phylogenomic approaches allow study of the ages and origins of new genes; in particular, they enable the characterization of orphan and taxonomically restricted genes [28,32]. However, the power of such approaches lies with their phylogenetic resolution (Figure 1A). While genomes of different insect orders are too divergent for studying the evolution and origin of new genes, comparisons between closely related members of the same genus, such as *Drosophila*, can be used to reliably measure evolutionary constraints as well as to make statements about gene origin [28,33–35]. Studies in the last decade began to use closely related species that are separated only for short time periods (Figure 1A). Such deep taxon phylogenomics can capture recent evolutionary origins and dynamics. In principle, the use of a robust phylogenomic framework allows the identification of enabling mutations (i.e., mutations that allow the transition from noncoding to coding sequences), which is considered to be the strongest indicator of a *de novo* origin. Indeed, some recent studies have elucidated the mechanisms of origin and evolutionary dynamics of new genes (i.e., studies in primates [11,20,36–38], yeast [29,39,40], mouse [41,42] and *Drosophila* [34,35,43] spearheaded this research). Such established model organisms for genome evolution can rely on well curated genomic resources, which are important to distinguish true orphan genes from annotation artifacts. In addition, recent studies also employed deep taxon sampling in nonclassical model organisms, such as *Drosophila pseudoobscura* [28], the yeast genus *Lachancea* [29,44], rice [30,45], and the nematode *Pristionchus pacificus* [9,31], the latter of which will be discussed in more detail in the following section.

Finally, it is important to note that many aspects of genome architecture can have a strong influence on the frequency and patterns of new gene origin [46]. Specifically, differences in genomic composition (i.e., fraction of coding sequence, intron size, and transposable elements) can lead to largely different contributions of origin mechanisms. For example, the frequency of *de novo* origin would be expected to be much higher in a mammalian genome with 3% coding sequences as opposed to a nematode genome with 20–30% coding sequences. Therefore, it is very likely that in order to reveal the full complement of mechanisms of origin of new genes, genomic analysis must be coupled with both broad taxonomic coverage and deep taxon phylogenomics.

Deep Taxon Phylogenomics of *Pristionchus* and the Origin and Evolution of New Genes in Nematodes

Fungi, insects, mammals, and plants have traditionally provided important insight into genome evolution and the occurrence of new genes. Recently, deep taxon phylogenomics were applied to study genome evolution and the emergence of new genes in nematodes of the genus *Pristionchus*. The hermaphroditic species *P. pacificus* is a satellite model to *Caenorhabditis elegans* with an established functional toolkit [47–49]. This species is a soil nematode that is reliably found in association with scarab beetles in world-wide samplings, which enabled the systematic search for related species with similar associations [50–54]. As a result, a collection of more than 40 *Pristionchus* species is currently available, some of which form sterile F1 hybrids, indicating that these species are extremely closely related. A phylogenetic framework of the *Pristionchus* genus was established by transcriptome sequencing of all species and revealed striking patterns of gene loss [55].

Initial sequencing of the *P. pacificus* genome classified roughly one third of all genes as orphan genes when *P. pacificus* was compared with *C. elegans* [56–58]. These two nematodes belong to distinct families, the Diplogastridae and Rhabditidae, respectively, and their protein sequence divergence is five times larger than the divergence between human and mouse [56]. Given that until recently the *P. pacificus* genome was the only diplogastrid genome available, the high number of orphan genes is likely due to sparse taxon sampling. Nevertheless, as initial gene annotations were completely based on automated gene predictions with the guidance of limited transcriptome and

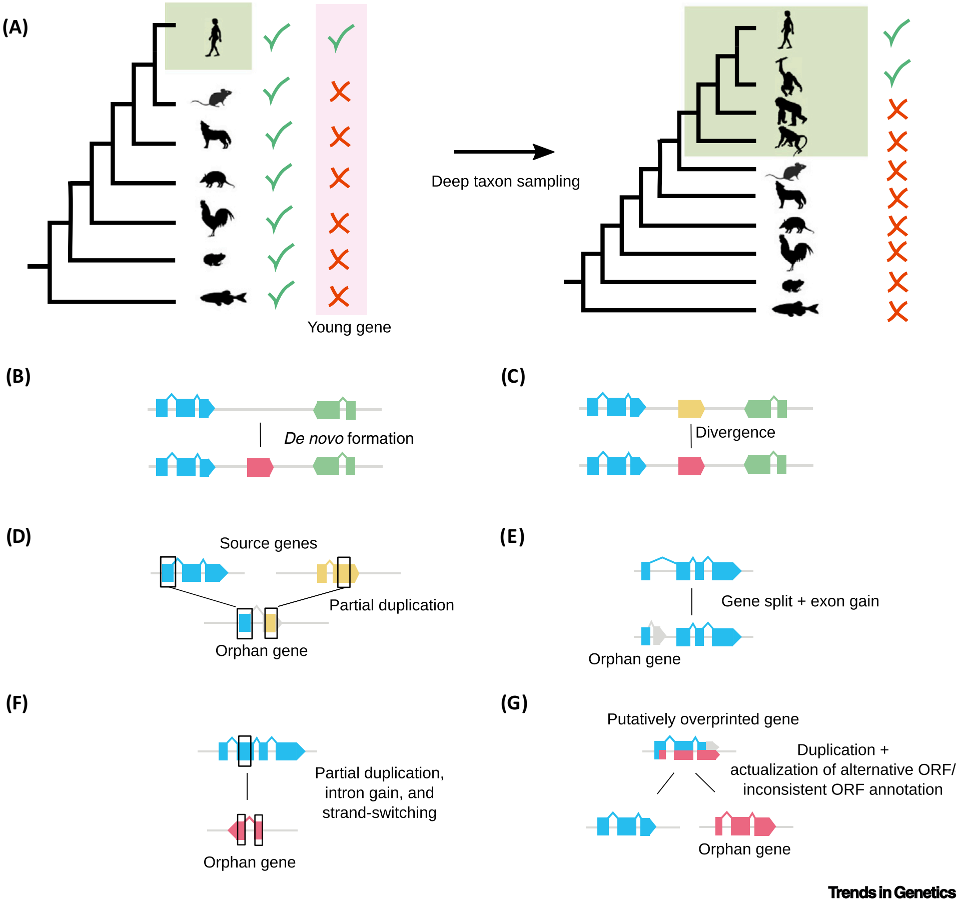


Figure 1. Deep Taxon Sampling and the Origin of Orphan Genes.
(A) Deep taxon sampling facilitates the investigation of new gene formation in closely related genomes. Silhouette images of animals were taken from PhyloPic (www.phylopic.org) and are available for reuse under the Public Domain Dedication 1.0 license. (B–F) The diversity of orphan gene origin as identified in *Pristionchus* nematodes. (B) The schematic genomic locus shows a candidate orphan gene (red) with two neighboring syntenic anchor genes (blue, green). Closer investigation of the orthologous region in sister taxa may eventually reveal either *de novo* formation (B) or divergence (C) as the mechanism of origin. (D) Partial exonic duplications from different source genes can result in a chimeric orphan gene. Loss of detectable homology is caused by the small size of duplicated fragments combined with moderate divergence. (E) Gene split with subsequent gain of a completely new exon can result in a new orphan gene. In this case indels in the duplicated exon induce frameshift causing the loss of protein homology. (F) Partial duplication, intron gain, and transcription from the opposite strand generate an orphan gene. (G) Orphan genes can arise by duplication and actualization of an alternative open reading frame (ORF) in a putatively overprinted gene.

proteome data [56,57], it was important to investigate if *P. pacificus* orphan genes are real. Indeed, RNA-seq data as well as comparative and population genomic data revealed that the vast majority of orphan genes is expressed under standard laboratory conditions and is under evolutionary constraint [59].

Utilizing the high phylogenetic resolution of *Pristionchus* nematodes, a more detailed phylogenomic analysis was performed by sequencing the genomes of eight species with a ladder-like phylogeny and two outgroups to determine the age of new genes and to study their evolutionary dynamics [31]. This analysis revealed a number of genome-wide features. First, deep taxon sampling and the selection of species with a ladder-like phylogeny allowed for the classification of new genes into age classes and to contrast their evolutionary dynamics at various time scales. Second, genes of new age classes are localized at the chromosomal periphery, whereas they are rare in the chromosomal centers. This observation also reflects general trends of nematode chromosome architecture, such as higher recombination rate and more genetic diversity at the chromosome arms as well as distinct epigenetic profiles between chromosome arms and centers [58,60]. Third, many new genes show weaker expression as opposed to old genes, thus confirming previous findings. This suggests that expression either increases or becomes broader over time [31] and is consistent with the 'out of testis' hypothesis, which states that either sexual selection or special cellular environments may create favorable conditions for the expression of gene-like sequences [25]. Interestingly, a parallel study of *P. pacificus* epigenetic profiles found evidence that new genes arise in the vicinity of enhancers of older genes [60]. This is consistent with previous finding of abundant transcriptional activity around *cis*-regulatory regions [61,62] and suggests that enhancers may function as promoters for new genes [60]. Fourth, new genes have a higher propensity of being lost than older and conserved genes. And finally, new genes exhibit only weak evolutionary constraints.

A more detailed study of orphan origins applied phylostratigraphy and additional quality filters to define a high confidence set of 29 species-specific genes, for which the origin could be traced based on manual inspection of syntenic regions in closely related genomes [9]. This revealed diverse divergence mechanisms, including chimeric origin, alternative reading frame usage, and gene splitting with subsequent gain of *de novo* exons as well as cases of complete *de novo* origin (Figure 1B–F). In addition, this study pointed out that technical problems such as annotation artifacts and heuristic failure of homology searches inflate the number of species-specific orphan genes. Together, this work and related studies of gene duplication established *P. pacificus* [63,64] as the primary system to study the evolution of new genes in nematodes. These studies also confirmed many known trends from vertebrates, insects, and plants, including the either low or spatiotemporally restricted expression of new genes [42,65,66] and the inverse relationship between evolutionary rate and age [28,30,67,68]. Importantly however, the availability of chromosome-scale assemblies revealed differences in the chromosomal distribution of new genes across phyla. For example, while novel genes cluster near the centromeric regions in rice genomes, nematodes have holocentric chromosomes and new genes preferentially cluster at the chromosome arms [30,31,58]. In *Drosophila*, new genes seem to be enriched on certain sex chromosomal areas [28], whereas in humans an association with DNA replication timing has been reported [69]. These differences highlight the need to study the evolution and origin of new genes across different phyla.

What Does it Take to Be a Gene: From Intraspecies Characterization to Pervasive Translation

Despite the power of deep taxon phylogenomics, it is not likely to provide a full understanding of the evolutionary dynamics of all new gene classes. Important challenges remain for multiple reasons. First, even at considerably high interspecies phylogenetic resolution, the detection of molecular fossils, such as enabling mutations, is often hampered by the rapid divergence of noncoding sequences. Importantly, apart from identifying the enabling mutations in a sister species, an additional noncoding sequence from an outgroup species is required to reliably infer that the enabling mutation is actually the derived state. With only one sequence from a sister species, a putative enabling mutation could actually represent a pseudogenization event of an ancestral gene of unknown origin. Given these difficulties, conclusive evidence for *de novo* origin is still quite rare [9,29,70]. The resulting large number of species-specific orphan genes, for which origin cannot be analyzed at the species level, either hints at rapid divergence of corresponding noncoding sequences in the sister species [e.g., complex structural variations, open reading frame (ORF) switching] or indicates that the standard approaches for homology detection (e.g., BLAST) are inadequate to find traces of homology at the species level. Therefore, to achieve an optimal phylogenetic resolution, the interspecies comparison has to be complemented with intraspecies studies to compare genomes of multiple diverging populations. Currently, corresponding resources have been developed in several species, with the most comprehensive dataset available in humans [35,71].

Second, besides phylogenetic resolution, additional conceptual problems exist. One important hurdle in the determination of the relative contribution of the two gene origin mechanisms is the ascertainment that a candidate gene is really protein-coding. As new genes tend to be expressed either weakly or in a very restricted manner [60,63], direct evidence of transcription and translation is often limited [29,59]. Similarly, indirect evidence based on evolutionary constraints tends to retain low statistical power for individual genes at the intraspecies level [23,59]. In addition, recent studies employing ribosome profiling proposed that large numbers of translated ORFs seem to be either nonfunctional or serve regulatory roles in the expression of downstream coding sequences, while their actual peptide sequences hold little significance [72–75]. Thus, the ability to distinguish all *bona fide* protein-coding genes remains elusive, even at a very narrow phylogenetic distance; measures of evolutionary constraint are not powerful enough and direct evidence of translation is either unavailable or inconclusive. Together with accumulating evidence that the lifespan of new genes may be associated with their mechanism of origin [8,28,76,77], the relative contribution between *de novo* and divergence will also depend on the definition of a gene in a given study [78]. If putative products of pervasive transcription and translation with annotated ORFs are already considered as gene-like sequences, more genes can be classified as *de novo* genes, in contrast to a more conservative definition that only considers genes that survived long enough to have gained verifiable exon–intron boundaries [9] (Figure 2).

In summary, we argue that systematic approaches and individual case studies of intra- and cross-species phylogenomic data sets are necessary to elucidate gene origin mechanisms and to quantify their contribution. Cross-species analyses at maximal phylogenetic resolution are needed to study evolutionary constrained and most likely functional new genes. At the same time, intraspecies studies are well suited to characterize the raw material, out of which new genes are born. Finally, this has to be complemented by comprehensive transcriptomic data across life stages and tissues to test hypotheses about their regulation and developmental importance [79,80].

Linking Mechanism of Origin and Evolutionary Stability

Numerous new genes have been associated with phenotypic innovations, but most of them are taxonomically restricted duplicates [1,4] or orphan genes of unknown origin [2,3]. This puts into doubt the importance of *de novo* genes for phenotypic evolution. In addition, the fast turnover and limited constraint acting on gene-like sequences raises the questions of how quickly new genes acquire functions at an organismic level and which types of new genes live long enough to be integrated into the biology of their hosts. We hypothesize that the high numbers of reported **proto-genes** and apparently nonfunctional peptides (as a result of pervasive transcription and translation), is just a byproduct of the cells' inability to regulate the transcriptional and translational machinery tightly enough to only express those genes that are absolutely necessary at a given point in time. This exposes noncoding sequences to a low level of basal expression and thus offers a playground out of which novel functions may evolve. The idea that random polypeptides might display many unfavorable properties, such as toxicity due to aggregation, has inspired the preadaptation hypothesis [73,77]. This theory states that initial purging of weakly expressed but strongly deleterious peptides shifts the raw material towards more gene-like properties. In combination with loss of nonfunctional peptides due to genetic drift, this leaves only a small pool of survivors [76] beyond a certain age. Even if these *de novo* genes have no particular function as of yet, they get better adjusted to the cellular environment due to their long exposure and may start interacting with other components of the cellular network. It is important to note here that new duplicates, unless they are immediately strongly selected against due to dosage imbalance [81], have at least a couple of advantages over *de novo* genes. First, they are less likely to exhibit toxicity as they derive from sequences that evolved over sufficiently long periods to remove their toxicity. Further, sequences arising from duplications can easily be longer than newly evolved ORFs, are more likely to have an optimized amino acid composition and codon usage, and have higher chances of containing secondary structures or functional motifs that readily allow them to interact with the existing cellular networks. Once the integration of a new gene into any cellular network provides a fitness advantage to the organism, selection will act to preserve such a gene, leading to long-term survival. Thus, initial persistence of gene-like sequences may be the key to acquiring

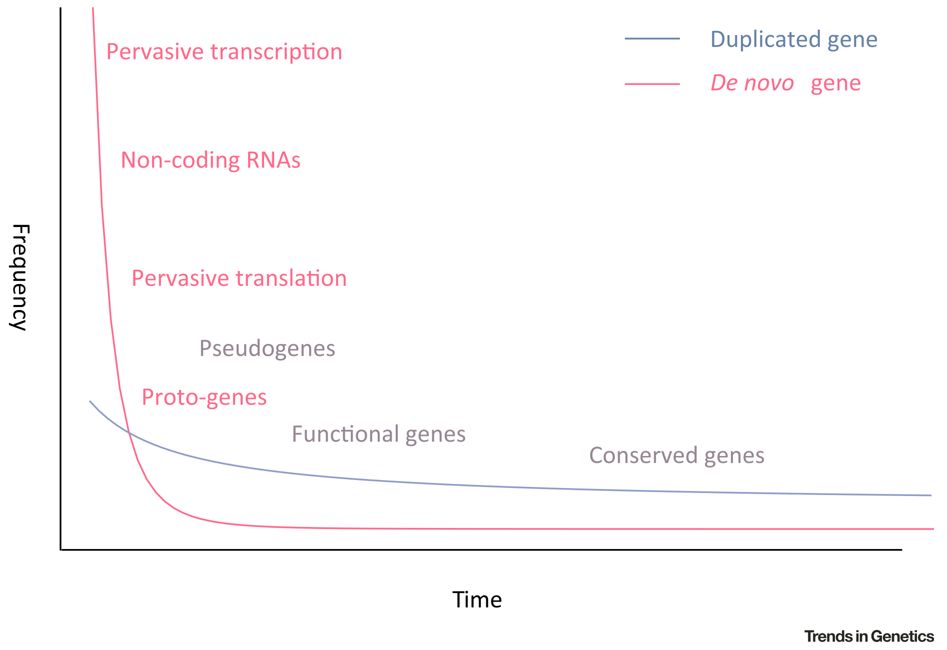


Figure 2. Hypothetical Model of Birth Rate and Lifetime of New Genes.
The lifespans of new genes may depend on their mechanism of origin. Thus, in this model, the relative contribution of different mechanisms will depend on the definition of a gene and the evolutionary time scale of the comparison.

biological functions and evolutionary stability. The fundamental differences in starting conditions between different classes of new genes may explain why *de novo* genes are frequently generated but quickly lost, while at the same time duplicated genes dominate the population of new genes at larger evolutionary distances [28,76]. At smaller time scales, it was proposed that *de novo* gene birth may be more prevalent than gene duplication [39]. This initial observation from yeast was further supported by the finding that the amount of novel ORFs in rodent genomes outnumbers species-specific duplications by at least an order of magnitude [42,75]. The proposed model (Figure 2) would indicate that quantifying the contribution of duplication versus *de novo* formation to new gene emergence makes sense only if the variable lifespans of different classes of new genes are also considered. Thus, to fully understand the dynamics between gene emergence and loss, systematic studies that measure the contribution of different origin mechanisms to new gene emergence across various time scales are needed. Consequently, this requires deep intra- and interspecies taxon phylogenomics.

Concluding Remarks and Future Perspectives

The question of how molecular innovations facilitate phenotypic novelties is central to evolutionary biology and understanding how new genes are formed and retained is the key to answering it. With the availability of deeply sampled phylogenomic data sets, it will soon be possible to construct models of genome evolution describing how new genes are formed and why they are retained. One of the most striking conceptual implications from recent studies is that the notion of a gene as a discrete molecular entity needs to be shifted to a continuum of sequence types with various degrees of biological activities. This spectrum on one end begins with products of pervasive transcription and translation, moves to very recent *de novo* genes or proto-genes, includes pseudogenes and long noncoding RNAs, and finally ends with what we traditionally consider a functional gene (Figure 2). We hypothesize that persistence of gene-like sequences is crucial for acquiring biological functions and that different classes of new genes are not equally well equipped to survive this early phase. Thus, future studies at the interspecies and intraspecies level (see Outstanding Questions) are required to group the origins and dynamics of new genes into a general framework that can be linked to phenotypic evolution.

Outstanding Questions

Before studying mechanisms of gene origin, how is a gene defined in the first place? Neither transcriptional and translational evidence nor evolutionary conservation is enough to fully capture the set of functional genes. Integrative approaches combining functional, population, and comparative genomic data are needed to optimize gene annotation protocols.

How is homology shown for a gene that evolves so fast that the homology cannot be detected? While various divergence mechanisms have been comprehensively characterized, we still lack computational tools to reliably detect and quantify them.

How does origin affect life time for different classes of young genes? Once we can quantify the contribution of different mechanisms of origin, comparisons at multiple time scales allow us to test which classes are evolutionary most stable.

How do processes generating young genes differ between phyla? While phylogenomic data sets have been established in multiple phyla, there is no systematic study to compare processes of young gene formation within a common framework.

Acknowledgments

We would like to thank Tess Renahan and Dr Michael Werner for critically reading the manuscript and language corrections. We would further like to thank all members of the Sommer lab for general discussions.

References

1. Santos, M.E. *et al.* (2017) Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. *Science* 358, 386–390
2. Lightfoot, J.W. *et al.* (2019) Small peptide-mediated self-recognition prevents cannibalism in predatory nematodes. *Science* 364, 86–89
3. Mayer, M.G. *et al.* (2015) The orphan gene dauerless regulates dauer development and intraspecific competition in nematodes by copy number variation. *PLoS Genet.* 11, e1005146
4. Ragsdale, E.J. *et al.* (2013) A developmental switch coupled to the evolution of plasticity acts through a sulfatase. *Cell* 155, 922–933
5. Parker, B.J. and Brisson, J.A. (2019) A laterally transferred viral gene modifies aphid wing plasticity. *Curr. Biol.* 29, 2098–2103
6. Schmid, K.J. and Aquadro, C.F. (2001) The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* 159, 589–598
7. Heinen, T.J.A.J. *et al.* (2009) Emergence of a new gene from an intergenic region. *Curr. Biol.* 19, 1527–1531
8. Tautz, D. and Domazet-Lošo, T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702
9. Prabh, N. and Rödelsperger, C. (2019) De novo, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes. *G3 (Bethesda)* 9, 2277–2286
10. Raes, J. and Van de Peer, Y. (2005) Functional divergence of proteins through frameshift mutations. *Trends Genet.* 21, 428–432
11. Toll-Riera, M. *et al.* (2009) Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* 26, 603–612
12. Duret, L. *et al.* (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312, 1653–1655
13. Cortez, D. *et al.* (2009) A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol.* 10, R65
14. Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9, 605–618
15. Zhaxybayeva, O. and Doolittle, W.F. (2011) Lateral gene transfer. *Curr. Biol.* 21, R242–R246
16. Rödelsperger, C. (2018) Comparative genomics of gene loss and gain in *Caenorhabditis* and other nematodes. *Methods Mol. Biol.* 1704, 419–432
17. Stephens, S.G. (1951) Possible significance of duplication in evolution. *Adv. Genet.* 247–265
18. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer)
19. Jacob, F. (1977) Evolution and tinkering. *Science* 196, 1161–1166
20. Knowles, D.G. and McLysaght, A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res.* 19, 1752–1759
21. Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet.* 12, 263–270
22. Schlötterer, C. (2015) Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* 31, 215–219
23. McLysaght, A. and Hurst, L.D. (2016) Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17, 567–578
24. Schmitz, J.F. and Bornberg-Bauer, E. (2017) Fact or fiction: updates on how protein-coding genes might emerge from previously non-coding DNA. *F1000Res.* 6, 57
25. Van Oss, S.B. and Carvunis, A-R. (2019) De novo gene birth. *PLoS Genet.* 15, e1008160
26. Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
27. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960
28. Palmieri, N. *et al.* (2014) The life cycle of *Drosophila* orphan genes. *Elife* 3, e01311
29. Vakirlis, N. *et al.* (2018) A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* 35, 631–645
30. Stein, J.C. *et al.* (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* 50, 285–296
31. Prabh, N. *et al.* (2018) Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus* nematodes. *Genome Res.* 28, 1664–1674
32. Domazet-Lošo, T. *et al.* (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23, 533–539
33. Wiegmann, B.M. *et al.* (2009) Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol.* 7, 34
34. Levine, M.T. *et al.* (2006) Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9935–9939
35. Zhao, L. *et al.* (2014) Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343, 769–772
36. Toll-Riera, M. *et al.* (2009) Evolution of primate orphan proteins. *Biochem. Soc. Trans.* 37, 778–782
37. Xie, C. *et al.* (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* 8, e1002942
38. Chen, J.-Y. *et al.* (2015) Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral LncRNAs in primates. *PLoS Genet.* 11, e1005391
39. Carvunis, A.-R. *et al.* (2012) Proto-genes and de novo gene birth. *Nature* 487, 370–374
40. Cai, J. *et al.* (2008) De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179, 487–496

41. Murphy, D.N. and McLysaght, A. (2012) De novo origin of protein-coding genes in murine rodents. *PLoS One* 7, e48650
42. Pegueroles, C. *et al.* (2013) Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Mol. Biol. Evol.* 30, 1830–1842
43. Zhou, Q. *et al.* (2008) On the origin of new genes in *Drosophila*. *Genome Res.* 18, 1446–1455
44. Vakirlis, N. *et al.* (2016) Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.* 26, 918–932
45. Zhang, L. *et al.* (2019) Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* 3, 679–690
46. Lynch, M. (2006) *The Origins of Genome Architecture* (Sinauer)
47. Sommer, R.J. *et al.* (1996) Morphological, genetic and molecular description of *Pristionchus pacificus* sp.n. (Nematoda: Neodiplogastridae). *Fundam. Appl. Nematol.* 19, 511–521
48. Witte, H. *et al.* (2015) Gene inactivation using the CRISPR/Cas9 system in the nematode *Pristionchus pacificus*. *Dev. Genes Evol.* 225, 55–62
49. Sommer, R.J. (2015) *Pristionchus pacificus: A Nematode Model for Comparative and Evolutionary Biology* (BRILL)
50. Herrmann, M. *et al.* (2007) The nematode *Pristionchus pacificus* (Nematoda: Diplogastridae) is associated with the oriental beetle *Exomala orientalis* (Coleoptera: Scarabaeidae) in Japan. *Zool. Sci.* 24, 883–889
51. Kanzaki, N. *et al.* (2012) Description of three *Pristionchus* species (Nematoda: Diplogastridae) from Japan that form a cryptic species complex with the model organism *P. pacificus*. *Zool. Sci.* 29, 403–417
52. Kanzaki, N. *et al.* (2018) Samplings of millipedes in Japan and scarab beetles in Hong Kong result in five new species of *Pristionchus* (Nematoda: Diplogastridae). *J. Nematol.* 50, 587–610
53. Yoshida, K. *et al.* (2018) Two new species of *Pristionchus* (Nematoda: Diplogastridae) from Taiwan and the definition of the pacificus species-complex sensu stricto. *J. Nematol.* 50, 355–368
54. Herrmann, M. *et al.* (2019) Two new species of *Pristionchus* (Nematoda: Diplogastridae) include the gonochoristic sister species of *P. fissidentatus*. *J. Nematol.* 51, 1–14
55. Rödelsperger, C. *et al.* (2018) Phylotranscriptomics of *Pristionchus* Nematodes reveals parallel gene loss in six hermaphroditic lineages. *Curr. Biol.* 28, 3123–3127
56. Dieterich, C. *et al.* (2008) The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genet.* 40, 1193–1198
57. Borchert, N. *et al.* (2010) Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. *Genome Res.* 20, 837–846
58. Rödelsperger, C. *et al.* (2017) Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism *Pristionchus pacificus*. *Cell Rep.* 21, 834–844
59. Prabh, N. and Rödelsperger, C. (2016) Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics* 17, 226
60. Werner, M.S. *et al.* (2018) Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res.* 28, 1675–1687
61. Andersson, R. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461
62. Neme, R. and Tautz, D. (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* 14, 117
63. Baskaran, P. and Rödelsperger, C. (2015) Microevolution of duplications and deletions and their impact on gene expression in the nematode *Pristionchus pacificus*. *PLoS One* 10, e0131136
64. Baskaran, P. *et al.* (2015) Ancient gene duplications have shaped developmental stage-specific expression in *Pristionchus pacificus*. *BMC Evol. Biol.* 15, 185
65. Donoghue, M.T. *et al.* (2011) Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* 11, 47
66. Lemos, B. *et al.* (2005) Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol. Biol. Evol.* 22, 1345–1354
67. Albà, M.M. and Castresana, J. (2005) Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* 22, 598–606
68. Cai, J.J. and Petrov, D.A. (2010) Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* 2, 393–409
69. Juan, D. *et al.* (2014) Late-replicating CNVs as a source of new genes. *Biol. Open* 3, 231
70. Zhang, W. *et al.* (2019) Origination and evolution of orphan genes and de novo genes in the genome of *Caenorhabditis elegans*. *Sci. China Life Sci.* 62, 579–593
71. 1000 Genomes Project Consortium *et al.*. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073
72. Ruiz-Orera, J. and Mar Albà, M. (2019) Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet.* 35, 186–198
73. Wilson, B.A. and Masel, J. (2011) Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3, 1245–1252
74. Ruiz-Orera, J. *et al.* (2014) Long non-coding RNAs as a source of new peptides. *Elife* 3, e03523
75. Ruiz-Orera, J. *et al.* (2018) Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* 2, 890–896
76. Schmitz, J.F. *et al.* (2018) Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* 2, 1626–1632
77. Wilson, B.A. *et al.* (2017) Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* 1, 146
78. Gerstein, M.B. *et al.* (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681
79. Domazet Lošo, T. and Tautz, D. (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468, 815–818
80. Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326
81. Schuster-Böckler, B. (2010) Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One* 5, e9474