Orphan and *de novo* Genes in Fungi and Animals: Identification, Origins and Functions

Ercan Seçkin[1,2] , Dominique Colinet[1] , Edoardo Sarti[2], Etienne GJ Danchin[1]

1. Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, Sophia Antipolis, France
2. Centre Inria at Université Côte d'Azur, France

* corresponding author: Ercan Seçkin, email: ercan.seckin@inrae.fr

Abstract

Genes that don't have identifiable homologs in other species have been an intriguing and interesting topic of research for many years. These so-called orphan genes were first studied in yeast and since then, they have been found in many other species. This has fostered a whole field of research aiming at tracing back their evolutionary origin and functional significance. Orphan genes represent an important part of protein-coding genes in many species. Their presence was initially mainly hypothesized to result from high divergence from a pre-existing gene, with duplications or horizontal gene transfer facilitating their accelerated evolution. More recently, their possible *de novo* emergence from non-genic regions has gained particular interest. Some orphan genes are predicted to be involved in fertility, while others are involved in specific developmental stages, in adaptation mechanisms such as freeze protection or even human disease. However, there is currently no unified resource or synthesis that brings together existing knowledge about how often prevalent orphan genes are across different species and what their roles might be. In this review, we focus on orphan genes in animals and fungi (i.e opisthokonts). We provide a detailed summary of what has been discovered over time in terms of their prevalence in genomes, their origins as well as their roles in different biological contexts.

Introduction

Orphan genes and *de novo* gene birth

The definition of orphan genes varies across studies: some describe them as genes of unknown function (Hartig et al. 2011) or as orphan receptors that do not bind known ligands (Nothacker 2008). However, we use here the more classical evolutionary biology definition, which refers to orphan genes as those with no detectable homologs in other species. Orphan genes have been first described in the *Saccharomyces cerevisiae* yeast genome

(Dujon 1996) and were predicted to represent up to 30% of protein-coding genes in eukaryotes (Tautz and Domazet-Lošo 2011). Their emergence represents an important opportunity for the acquisition of new functions during evolution, in particular by driving genus or species-specific adaptations (Fakhar et al. 2023). Orphan genes may derive from a pre-existing gene that has accumulated high divergence reaching the point of no recognizable homology. This can be facilitated by gene duplication or horizontal gene transfer events, followed by rapid evolution. Studies in several species suggest, however, that this explanation concerns only a part of existing orphan genes (Vakirlis, Carvunis, and McLysaght 2020). The other hypothesis is that these orphan genes may have emerged from non-genic regions. This phenomenon, known as *de novo* gene birth, occurs when previously non-coding and/or not transcribed DNA sequences acquire the capacity to be transcribed then translated to a functional protein (Schmitz and Bornberg-Bauer 2017; Weisman 2022). For a long time, *de novo* emergence was considered highly unlikely. Indeed, the probability for a newly emerged gene coding for a functional protein to be maintained in populations by selection is intuitively extremely low (Jacob 1977). With the explosion of genomic sequencing projects and the resulting increase in available genome data for a higher diversity of species, it was realized that *de novo* gene emergence is not as rare as initially thought and that many species- or lineage-specific genes lack recognizable homologs (Khalturin et al. 2009). Several studies took advantage of this richer set of genome data to confirm the likely existence of *de novo* emerged genes (Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016; Van Oss and Carvunis 2019). A recent review provides detailed information specifically on de *novo genes*, including the methods to identify them, their possible functions and the challenge they still pose at an evolutionary biology point of view (Li Zhao, Svetec, and Begun 2024).

Mechanisms of *de novo* gene birth

In the case of a protein-coding gene, *de novo* emergence involves two main distinct processes: (i) transcription of initially non-coding DNA and (ii) acquisition of an open reading frame (ORF) (Figure 1). The order of these events allows two main mechanisms to be distinguished (Van Oss and Carvunis 2019): "transcription first" (Figure 1A) and "ORF first" (Figure 1B).

The "transcription first" mechanism is thought to be the most prevalent (Van Oss and Carvunis 2019), as a significant number of non-genic sequences are identified as transcribed. These non-coding transcribed sequences typically lack a canonical ORF due to the presence of premature stop codons and/or non-functional splice sites. Accumulation of mutations that eliminate these stop codons and/or establish correct splicing sites can result in the acquisition of an ORF and, consequently, the emergence of a *de novo* gene that can now be translated to a protein. In this context, such intermediate sequences have been described as protogenes. Protogenes may initially produce proteins or peptides with weak or detrimental

or even no functionality, and many of them are likely to be eliminated by natural selection. However, in rare cases, a protogene can provide a slight benefit to the organism, leading to its retention and gradual refinement through the accumulation of beneficial mutations. Over time, this process can result in the fixation of the protogene and its evolution into a fully functional gene.

In the case of the "ORF first" mechanism, an open reading frame (ORF) would be present but would not be transcribed due to the absence of an expression regulatory region. When mutations lead to the acquisition of such a promoter or regulatory region, the ORF starts being transcribed and becomes a *de novo* gene. This can also be facilitated by the insertion of a transposable element and its transcriptional regulatory regions upstream of an ORF.

However, it is important to note that the distinction between "transcription first" and "ORF first" mechanisms is not always straightforward. Just as it can be difficult to definitively classify an orphan gene as *de novo* or highly diverged, the temporal sequence of transcription and ORF acquisition may not be neatly separated. For example, an ORF formed in a region of low transcription may gradually acquire regulatory features, or a *de novo* gene may later undergo rapid divergence that obscures its origin.

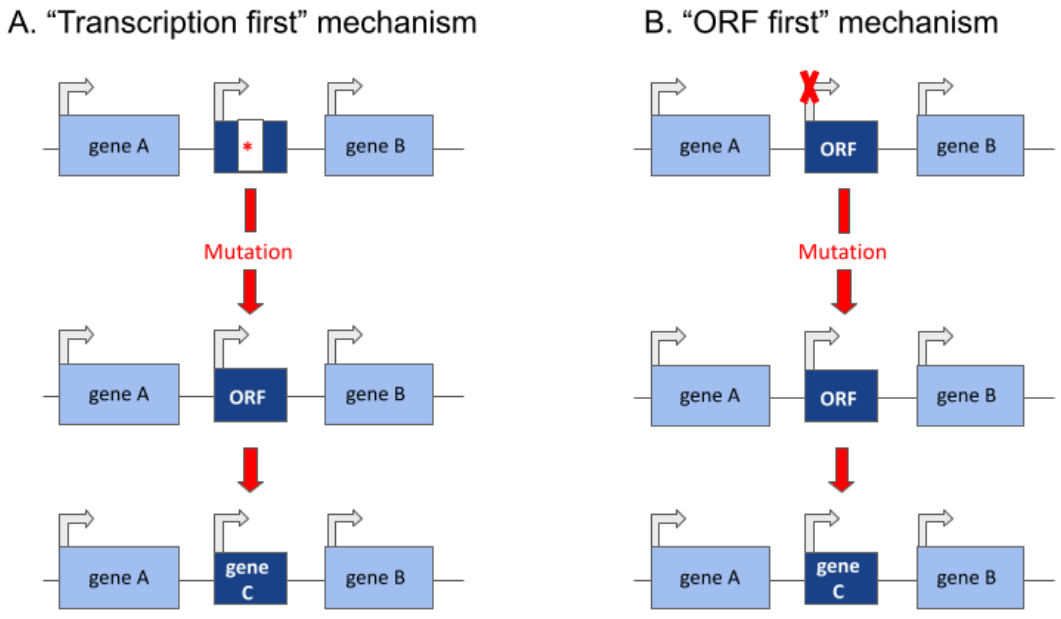

Figure 1: Emergence of a *de novo protein-coding* gene from a non-genic region

Gene C represents the *de novo* gene that emerges following one of the mechanisms described. A: In the "transcription first" mechanism, a non-genic sequence undergoes one or more mutations that eliminates premature stop codons (red asterisk), resulting in the acquisition of an ORF and the emergence of a *de novo* gene. B: In the ORF first mechanism, the acquisition of an expression regulatory region (grey arrows) allows the transcription of an existing ORF and the emergence of a *de novo* gene.

Methods to identify orphan genes and *de novo* gene birth

The most common approach to identify orphan genes is to start from a focal branch in the tree of life and search for homologs in other species using comparative genomics. One of the most widely used methods is phylostratigraphy, which involves identifying homologs for each gene in a species or clade of interest using BLAST (McGinnis and Madden 2004) or analogous similarity search tools. It should be noted here that most of these methods use protein sequences as a proxy for protein-coding genes. Then, based on these searches, groups or clusters of homologous genes are built using state-of-the-art software such as OrthoFinder (Emms and Kelly 2019), ORFan-Finder (Ekstrom and Yin 2016) or SonicParanoid (Cosentino and Iwasaki 2023). The identification of a gene exclusively within one or few closely related species enables the determination of the probable relative date of gene emergence, as well as the classification of the gene as orphan. The differences between orphan gene identification methods using comparative genomics have already been examined in detail in another review (Fakhar et al. 2023).

From an initial dataset of orphan genes, *de novo* genes can be identified by aligning the corresponding proteins to the genome of a closely related species translated in its 6 frames and looking for similarities in the corresponding non-coding regions. If the corresponding region in the related species is non-coding and mutations can be identified at specific positions that have led to the acquisition of an ORF, a *de novo* emergence event can be assumed. However, in case of high divergence, establishing reliable correspondences between genomes can be difficult. Translocations, structural changes, or incomplete assemblies can also obscure the ancestral origin of a gene. Distinguishing between *de novo* genes and highly diverged homologs is particularly challenging because highly diverged homologs no longer have detectable sequence similarity, making them appear to have arisen from non-coding regions. Conversely, *de novo* genes arise from non-coding sequences that may superficially resemble highly divergent homologs, further complicating their identification.

Nevertheless, incorporating the broader genomic context via conserved synteny analysis can help disentangle between these two possibilities. This consists in determining whether genes surrounding the candidate orphan gene in the focal species are conserved in target closely related species. In case of conservation of the surrounding genes, then the next step is to examine the homologous target locus corresponding to the candidate orphan gene. If at this target locus, another gene is present but lacks homology to the orphan gene, then we can hypothesize the orphan gene has highly diverged from the common ancestral gene. Conversely, if at this locus there is no predicted gene but partial alignment of the orphan gene with frameshifts and/or invalid splice sites, then the *de novo* gene birth hypothesis is more likely.

In recent years, new tools have been developed to facilitate the study of orphan and *de novo* genes by integrating existing methods into streamlined pipelines. One such tool is DENSE (Roginski et al. 2024), which combines comparative genomics, synteny analysis and expression data to identify candidate *de novo* genes. While such tools represent an important step towards standardising and simplifying *de novo* gene discovery, they are not yet widely used mainly because they are not easily applicable to all kinds of datasets. There is also another recent and comprehensive review on this subject where the identification of *de novo* genes is more broadly discussed (Grandchamp et al. 2025).

Known characteristics and possible functions of orphan genes

The functions of the majority of orphan genes are still unknown, as most of them lack known motifs, domains, recognizable folds and reliable protein structure predictions (Fakhar et al. 2023). However, there has been huge progress in the field and there are several clues to the functions of orphan genes in different species. These progresses are mainly achieved by combining biochemical and experimental structure analysis and also by working on the expression patterns of orphan and/or *de novo* genes in different compartments of an organism.

Orphan Genes Identified and Functionally Studied in Fungi and Animals

Since, historically, orphan genes were first described in yeast, we reviewed orphan gene cases in yeasts, then besides yeast in other fungi and finally more broadly in other opisthokonts such as animals, including human beings. Therefore, in the following sections, we will review several examples of highly divergent and *de novo* orphan genes by phylogenetic groups in chronological order to show how much these genes contribute to the genomes of the studied species, how they are identified, and what has changed over time in terms of our knowledge and the methods used to identify them.

FUNGI

Yeast

In 1995, Espinet et al. identified a series of genes involved in cell growth and they demonstrated that 11 of them, from SHE1 to SHE11, do not have any homologs outside of *Saccharomyces cerevisiae* (Espinet et al. 1995). These were the first examples of functional genes in yeast lacking homologs in other species. The term orphan was introduced by

Bernard Dujon in *S. cerevisiae* in 1996, once the yeast genome project was accomplished (Dujon 1996). A process of comparative analysis between yeast sequences and the available genome sequences of other species at that time from various databases indicated that 25% of the *S. cerevisiae* genome contained genes that had no identifiable homologs, referred to as orphan genes (Oliver et al. 1992). Later, in 2001, a study demonstrated that the SHE9 gene, which was initially called an orphan gene, had a homolog in another yeast, *Candida albicans* (Andaluz et al. 2001). The study also showed that overexpression of this gene impairs cell growth in this species. Homology research was conducted with BLAST (McGinnis and Madden 2004) and the expression levels were estimated by Northern blot analysis. Two ATGATT hexamers were identified in the promoter region and, when present in the forward orientation, this hexamer exerts a positive regulatory control in response to cell proliferation. As this study showed a homolog for SHE9 gene outside of *C. albicans*, we could no longer consider this gene as an orphan gene for *S. cerevisiae*. Moreover, when we checked the Saccharomyces Genome Database (SGD), we noticed that only SHE1, SHE2, SHE10 remain labelled as orphan genes. This shows how important depth is in a phylogenetic sampling to consider a gene orphan or not.

In 2008, an orphan yeast gene, BSC4, was identified and considered for the first time as a *de novo* emerged gene (Cai et al. 2008). Researchers performed a tBLASTN search using the protein sequence of BSC4 as a query against the genome sequences of 81 fungal species, including *S. bayanus, S. kudriavzevii, S. mikatae, S. paradoxus* and *S. cerevisiae* revealing that the BSC4 gene is unique to *S. cerevisiae*. To rule out the possibility that the homology search was problematic, they performed a genomic Southern blot with a probe designed against BSC4 and concluded that only the *S. cerevisiae* genome showed obvious hybridisation signals. Further synteny analysis indicated that the flanking genes of BSC4 have their orthologs in the same syntenic blocks of *S. bayanus, S. mikatae* and *S. paradoxus*. This also revealed that the species other than *S. cerevisiae* had multiple premature stop codons at the expected position of BSC4 gene. In the light of this evidence, they concluded that this was a case of *de novo* origin and classified the BSC4 gene as *de novo* emerged. They then carried out a series of experiments in the light of previous experiments on this gene to determine its function. Studies suggested that the expression of BSC4 is upregulated when *S. cerevisiae* enters the stationary phase. Therefore, this gene is potentially playing a role in DNA repair and contributing to the evolutionary fitness of *S. cerevisiae* in nutrient-poor environments. However in 2024, a study re-investigated its *de novo* status and suggested that this gene may be emerging from the end part of another gene, suggesting a gene fission from a precursor gene where the N-term of BSC4 aligns partially with the C-term of the precursor (Hannon Bozorgmehr 2024).

Building on the growing evidence for *de novo* gene emergence, another study in 2010 identified MDF1, a gene with a distinct regulatory function in yeast mating (D. Li et al. 2010). The study showed that the protein-coding sense gene MDF1 arose *de novo* and can

significantly suppress mating efficiency. Firstly, the authors performed a BLAST search against the UniRef90 database using PSI-BLAST and found no significant homologous ORF in the closely related species. They verified that the synteny was conserved in multiple species across fungi, then they manually aligned the intergenic region between the flanking genes in other species and verified that this region could not encode for proteins in any species other than *S. cerevisiae* due to the presence of multiple stop codons and frame-shifting indels. The function of this *de novo* gene is understood by working on an antisense gene that acts as a transcriptional repressor of the MDF1 gene by binding to its promoter. Microarray analysis showed that when the MDF1 gene was suppressed, mating success was significantly higher. By binding to a protein that is one of the determinants of yeast mating type, MDF1 suppresses yeast mating behaviour and allows rapid vegetative growth.

In 2018, a more comprehensive research was conducted on 15 different yeast species where 703 *de novo* gene candidates were identified. The existence of 85 of them was validated by proteomic data and 25 among them had evidence of translation according to mass spectrometry experiments (Vakirlis et al. 2018). The study suggested that *de novo* gene birth is a widespread phenomenon in yeast, but only a few are ultimately maintained by selection. To identify the 703 *de novo* genes, the authors first performed a multiple sequence alignment of the protein sequences for each family and constructed HMM and PSSM profiles. They then performed exhaustive similarity searches against several databases using BLASTP for singletons and PSI-BLAST for families with their own HMM or PSSM profiles. They then took the singletons or families with no hits against nr, compared the families with no hits against nr between them, and merged the families with significant similarity. To distinguish between orphan genes that highly diverged from ancestral genes and *de novo* genes, they simulated the evolution of protein families using the ROSE program (Stoye, Evers, and Meyer 1998) and then inferred the branch of origin for each family along the genus phylogeny by phylostratigraphy using a custom pipeline. They concluded, with their analysis, that if a simulated family was assigned to the root of the focal genus, it was a highly diverged gene and if not, it was a case of *de novo* gene birth. This was one of the first studies in yeast where the results were not simply obtained by a BLAST search or similar, but where the proteins were classified using a more comprehensive and detailed pipeline, including the use of HMMs.

Expanding on the evolutionary significance of *de novo* genes, another 2018 study examined their spread and fixation within *S. cerevisiae* populations, revealing key insights into their persistence under different conditions (B. Wu and Knudson 2018). The research identified 84 *de novo* genes in *S. cerevisiae* and some of them are only expressed and translated under certain conditions. To do this, the authors first performed a BLASTP search of the *S. cerevisiae* proteins against those of 20 other *Saccharomycetaceae* species. Once the orphans were identified, they excluded the genes for which they could not find the orthologous non-coding sequence in the outgroup genomes of *S. paradoxus* and *S. mikatae*. Finally, they

confirmed the expression of the genes using transcriptomic data to conclude that they had identified 84 *de novo* genes. They compared their results with three previous studies (Carvunis et al. 2012; Vakirlis et al. 2018; Lu, Leu, and Lin 2017) and found that only 33% of their *de novo* genes were shared with at least one of the other three studies. Surprisingly, there were no *de novo* genes common to all 4 studies. The authors explained this by the exclusion of overlapping ancient genes for certain studies, e-value differences for homology searches and different thresholds for required expression levels. They also suggest that one of the studies had identified noncoding regions also as homologs rather than only the protein coding genes, some genes were only expressed under certain conditions which was not taken into account by one of the other studies and indeed 10% of the newly identified *de novo* genes were only expressed under specific conditions. This highlights the fact that different studies may apply different thresholds, scoring systems, and criteria leading to differing outcomes in orphan and *de novo* gene identification. Recent efforts are proposing solutions to try to solve this issue by standardizing different annotations in a common file format which would make different analyses more comparable (Dohmen et al. 2025).

Furthermore, the researchers of this paper also compared the transcriptomic data for these *de novo* genes for the wild type and two mutants where the products of the mutants were two proteins involved in pre-mRNA splicing and nonsense-mediated mRNA decay (Gould et al. 2016; Chapman and Boeke 1991). The results showed high expression levels for 8 *de novo* genes in the case of mutants which could be regulated by the mutant proteins and therefore they concluded that these *de novo* genes are possibly involved in mRNA processing. They also used ribosome profiling data to show that 51% of their *de novo* genes were found to be translated at specific time points or conditions. They then took advantage of several microarray data from SPELL database which is a query-driven search engine for large gene expression microarray compendia (Hibbs et al. 2007). Results showed that among the 84 *de novo* genes, 87% were associated with 52 functional categories defined by SPELL. Overall, 73% of the genes were identified as involved in carbon utilization processes while 7% were involved in cell aging.

Another study, published in 2020 characterized a *de novo* gene YBR196C-A in *S. cerevisiae* (Vakirlis et al. 2020), coding for a transmembrane protein. The orphan status of the gene was verified by the absence of homologs in other *Saccharomyces* and fungal species. Syntenic studies then revealed that this gene most likely emerged *de novo* from a thymine-rich intergenic region. Expression of this gene was shown to have a beneficial impact on yeast fitness. The authors verified the bioinformatics prediction of a transmembrane localization experimentally by using EGFP-tagged visualization of the protein via confocal microscopy and membrane association assays, which revealed the presence of the protein at endoplasmic reticulum (ER) membrane. Follow-up studies in 2023 and 2024 further characterized this *de novo* gene. Constructing a reference translatome for *S. cerevisiae* and using an experimental approach to mutate ATG to AAG codon in some strains, preventing

ORF translation, Wacholder et al. revealed that YBR196C-A gene has phenotypic consequences when its translation is inhibited (Wacholder et al. 2023). There was a fitness reduction under stress conditions. The authors also highlighted the orphan status of other genes of *S. cerevisiae*, most importantly HUR1, ICS3, YPR096C and YDL204W-A. These genes are involved in DNA repair (Omidi et al. 2018), copper homeostasis (Alesso, Discola, and Monteiro 2015), regulation of a gene involved in sugar metabolism (Hajikarimlou et al. 2020) and cell fitness respectively (Houghton et al. 2024). Confirming initial analyses, Saeki et al. explicitly identified the YBR196C-A gene as encoding a beneficial emerging protein (BEP) localized at ER using overexpression profiling experiments (Saeki et al. 2023). Houghton et al. re-analyzed fitness measurements from the 2020 study and showed that ER-localized BEPs all contain transmembrane domains followed by short C-termini (Houghton et al. 2024). They also showed the pathways that this protein might be involved in and revealed that ER-localized BEPs are beneficial across more conditions than other BEPs. Given all these evolutionary and experimental studies, we can assume that YBR196C-A is one of the best-characterized *de novo* genes so far, even though the function of this gene is not yet fully understood.

Other fungi

In 2015, Kohler et al. conducted a comparative genomics analysis to elucidate the evolution of the mycorrhizal lifestyle in fungi and determined that 7-38% of the genes induced during symbiosis are orphan genes, many of which encode secreted effector-like proteins (Mycorrhizal Genomics Initiative Consortium et al. 2015). The study involved sequencing the genomes of 13 ectomycorrhizal (ECM), orchid (ORM), and ericoid (ERM) fungal species, along with 5 saprotrophic species, and comparing them with existing fungal genomes with Markov Cluster Algorithm (MCL). The gene expression of identified genes were assessed with RNA-seq. These findings suggest that the evolution of mycorrhizal symbiosis in fungi occurred through convergent evolution, leading to the emergence of distinct sets of genes that are specifically activated during mycorrhizal interactions in different fungal lineages. In contrast to most of the previously described methods to identify orphans in yeast, in this more recent study, MCL algorithm was used for the comparative genome analysis.

In 2016, another study investigated the evolution of orphan genes in the genome of *Zymoseptoria tritici*, a fungal pathogen of wheat. The authors identified 296 such genes in the *Z. tritici* genome (Plissonneau, Stürchler, and Croll 2016). Utilizing single-molecule real-time sequencing, genetic mapping, and transcriptomics, they assembled and annotated the genome of the virulent *Z. tritici* field isolate 3D7. Comparative analyses with the reference genome IPO323 of the same species using BLASTn and synteny analysis revealed significant chromosomal inversions and variations in transposable element clusters, leading to extensive chromosomal-length polymorphisms. Notably, both genomes contained large, unique sequence tracts with the 3D7 genome harboring 296 genes absent in IPO323. These

orphan genes were enriched in putative effector genes, including one highly upregulated during wheat infection. However, the paper does not state that these 296 orphan genes are missing in other fungal species or other species in general. They compared their genome only to the reference genome, which is IPO323. Therefore we cannot conclude for sure that *Z. tritici* has 296 orphan genes as there might be gene loss cases in IPO323.

Continuing the exploration of orphan genes in fungal pathogens, a 2020 study on *Fusarium graminearum* identified an orphan protein that actively modulates host immunity (Jiang et al. 2020). The authors used BLASTp for protein homology search and also tBLASTn to search against genomes, firstly to two closely related *Fusarium* species and if they were orphan, they were compared also against nr. They identified a total of 971 (~7,3% of all protein coding genes) orphan genes. The authors then focused on one of these orphan genes which was predicted to encode a protein with a signal peptide for secretion, Osp24. According to protein interaction assays, this protein, which is unique to *F. graminearum*, appears to facilitate infection by targeting TaSnRK1α, a key regulator of the plant's immune response. The researchers demonstrated that the orphan protein interacts with TaSnRK1α by targeting it for degradation through the proteasome pathway, thereby weakening the plant's immune defenses.

Also in 2020, other researchers investigated the emergence of new gene families in another fungal genus, *Amanita*, focusing on their association with the evolution of ectomycorrhizal (ECM) symbiosis and the study identified 109 gene families unique to ECM *Amanita* species, absent in closely related asymbiotic species (Y.-W. Wang et al. 2020). These unique gene families were found to be under strong purifying selection and upregulated during symbiosis, suggesting their functional relevance to the mutualistic association. Among the unique gene families, the most upregulated gene in symbiotic cultures encodes a 1-aminocyclopropane-1-carboxylate deaminase, an enzyme capable of downregulating the synthesis of the plant hormone ethylene, a common negative regulator of plant-microbial mutualisms. Furthermore, the homology search and synteny showed 2 gene families of these orphan gene families are candidate *de novo* gene families, with so far no known function.

Later on, in late 2022 and 2023, Wang et al. focused on understanding the lineage-specific genes in the fungi *Neurospora crassa* and revealed that there are 670 lineage-specific orphan genes (Zheng Wang et al. 2022). Following that, they also demonstrated that gene duplication, relocation, and regional rearrangement drive the formation of these genes (Zheng Wang et al. 2023). They employed a phylostratigraphic approach and then BLAST search against FungiDB to identify lineage-specific gene clusters. Then, the expression of these genes were verified via transcriptomic data. By analyzing synteny and clustering patterns, they identified that 78% of lineage-specific gene clusters are located near telomeric regions, which contain extensive non-coding DNA and duplicated genes. These

regions, termed “rummage regions”, allow for rapid recombination and mutation, creating a favorable environment for new genes to arise and evolve. Using transcriptomics from 68 experimental data points, the researchers identified that these genes are often involved in peripheral regulatory functions, though they play critical roles under specific conditions. The study highlighted *mas-1*, a lineage-specific orphan gene likely derived from a *lysophospholipase* precursor, which contributes to cell wall integrity and antifungal resistance.

Aside from their roles in adaptation and symbiosis, orphan genes have also proven useful as molecular markers for species identification. A 2022 study developed an approach to distinguish *Aspergillus* species using orphan genes (Zhong Wang et al. 2022). The researchers developed a multiplex PCR method to identify *Aspergillus cristatus* and *Aspergillus chevalieri* in Liupao tea using species-specific orphan genes In this study, six fungal strains were isolated from Liupao tea and identified as *A. cristatus, A. chevalieri,* and *A. pseudoglaucus*. According to this study, traditional ITS sequencing proved insufficient to distinguish closely related species due to high sequence conservation. To overcome this, the researchers used comparative genomics to identify orphan genes unique to each species and designed species-specific primers for multiplex PCR. This approach enabled rapid and accurate identification of *A. cristatus* and *A. chevalieri* in both Liupao and Fu brick teas, highlighting the utility of orphan genes in distinguishing closely related species.

ANIMALS

Drosophila and other arthropods

In 2000, a study of the model fly species *Drosophila melanogaster*, nematode species *Caenorhabditis elegans* as well as humans showed that about 30% of *D. melanogaster* genes had no identifiable homologs and were therefore considered orphans according to BLASTP results (Rubin et al. 2000). Then, in 2003, another study followed up to investigate whether there was a change in the proportion of predicted orphan genes over time in *Drosophila* and compared about 14,000 predicted proteins of the *Drosophila* proteome with other insects using BLASTP (Domazet-Loso and Tautz 2003). The authors compared the different results obtained with different e-values varying from 10E-100 to 10 and, as expected, the number of sequences with no homologs is very small at the highest e-values due to many insignificant random matches. The results for more stringent lower e-values, the ones preferred by many studies, 10E-3 to 10E-5, showed that there were still 26%-29% of *D. melanogaster* genes that had no identifiable homologs. The results thus indicated that there was no significant change in the proportion of orphans, despite the growth of the database and improvements in annotation over time. To be sure that this e-value range was the best choice, they compared the different homologs obtained at different e-values and concluded that at lower

cutoffs the proportion of false positives increased and at higher cutoffs true orphans were increasingly lost, thus confirming that the 10E-3 to 10E-5 range was the best balance between sensitivity and selectivity. This e-value range is still the most used in most of the studies. The authors then carried out a comparative analysis of expressed genes only between *D. melanogaster* and *D. yakuba* and the results showed 8.4% and 19.7% of orphan genes for *D. melanogaster* were expressed for the embryonic and adult stages respectively. Compared to the whole-genome analysis, these values were significantly lower. The study suggested that this could be due to incorrect annotation at the genomic level, or that orphans are likely to be expressed at lower levels than non-orphan genes. Incorrect annotations can be problematic because they may lead to the misidentification of genes, causing some genuine orphan genes to be overlooked or misclassified. This can result in an underestimation of their prevalence and functional significance. Also, it is important to note that some genes might be expressed only at certain stages of life. Finally, the researchers concluded that *D. melanogaster* contains an important number of orphan genes even in the light of new data and the selection of e-value is important, with the preferred range being between 10E-3 and 10E-5.

While these early studies focused on the proportion of orphan genes in the genome, subsequent research shifted toward understanding their biological significance, particularly in reproduction. In 2006, a study described 5 *de novo* genes expressed in the testes and implicated in male production in *D. melanogaster* under selective pressure (Levine et al. 2006). First, the authors identified orphan genes by BLASTN against the genomes of two other *Drosophila* species and kept only those that had complete cDNA sequences according to the flybase database and/or those that were experimentally confirmed. They then applied syntenic approaches and kept only 5 genes that had high quality syntenic alignments of the flanking regions of the *de novo* gene in *D. melanogaster* compared to *D. yakuba, D. erecta* and *D. ananassae*. Southern blot analysis of these 5 genes confirmed the computational prediction and they concluded that there were 5 *de novo* genes in *D. melanogaster* that met all their stringent criteria, and therefore there were probably many more. RT-PCR data from RNA isolated from whole adult female and male reproductive tissues showed that all five genes were expressed in the testes and they demonstrated that 4 of their 5 *de novo* genes are X-linked. In 2007, a follow-up study showed that *D. yakuba* and/or *D. erecta* also have 7 additional *de novo* genes involved in male reproduction (Begun et al. 2007). They analysed the *D. yakuba* testis-derived cDNA library and followed a similar procedure to the previous study that identified *D. melanogaster de novo* genes. They concluded that *de novo* gene birth is an important phenomenon for male reproduction in *Drosophila* species. A subsequent study conducted in 2014 provided further evidence that a greater number of testis-expressed *de novo* genes are involved in male reproduction in *D. melanogaster* by examining different populations of this species (L. Zhao et al. 2014). An Illumina paired-end RNA sequencing approach was employed to characterise the testis transcriptome of six previously sequenced *D. melanogaster* strains. The resulting analysis revealed that there are

a total of 142 expressed *de novo* genes in the testis even under the very strict filtering criteria.

While most *de novo* gene studies in *Drosophila* have focused on male reproductive functions, one study identified a *de novo* gene involved in female reproduction, expanding the known functional repertoire of orphan genes in this species. Similar approaches to those employed in recent studies were used, including BLAST for homology search, synteny to detect non-coding regions of the *de novo* gene in closely related species, and expression levels in different tissues for the identified gene (Lombardo et al. 2023).

Whereas previous studies examined species-specific *de novo* genes in *Drosophila*, later research expanded the scope to investigate orphan genes across multiple species within the genus, providing insights into broader evolutionary trends. In 2020, another group of researchers who had been investigating orphan genes and *de novo* gene birth in *Drosophila* demonstrated that across 12 *Drosophila* species, there are 6,297 orphan genes, with between 8.7% and 39.2% of them resulting from *de novo* gene birth (Heames, Schmitz, and Bornberg-Bauer 2020). To identify them, the authors first clustered all sequences of the 12 *Drosophila* species and 3 outgroup species by BLASTP and then they compared the clusters to the NCBI non-redundant (nr) database. Furthermore, a phylostratigraphic method was employed to ascertain the gene gain timing scenarios, while syntenic approaches were utilised to detect instances of *de novo* gene birth within the *Drosophila* clade. Here, it is important to underline that the study was not describing species-specific orphan genes like the previous ones but it was revealing orphan genes at the whole *Drosophila* genus level.

Beyond identifying orphan genes, researchers have also sought to understand their structural properties and evolutionary stability. One such study focused on the structural characterization of the Goddard protein, a *de novo* gene involved in *Drosophila* male fertility (Lange et al. 2021). To achieve this, the researchers employed a combination of modelling, NMR and circular dichroism approaches, which revealed that the protein in question contains a central α-helix, while the remaining portions are predominantly disordered. The researchers demonstrated that this structure is a novel one by comparing the obtained structure to the PDB database. Furthermore, they proposed that this structure has been preserved by the organism over millions of years, as evidenced by its conservation across diverse *Drosophila* species (but absence from the rest of species). To substantiate this hypothesis, they reconstructed the ancestral sequence of the node shared by five *Drosophila* species that express this protein and utilized the structure that they described for each of them to infer an ancestral structure. Additionally, they demonstrated that this protein localizes to elongating sperm axonemes and that its absence impairs the individualization of elongated spermatids. Nonetheless, in 2024, a preprint study discussed the *de novo* status of the Goddard protein and suggested that it is closely related to the N-term of another protein

(Hannon Bozorgmehr 2024). Therefore, this might not constitute a case of *de novo* gene birth but rather divergence from a pre-existing gene.

Expanding on individual cases like Goddard, recent large-scale analyses have examined the structural evolution of *de novo* proteins in *Drosophila*, offering insights into their folding and functional constraints. In 2024, a study identified 555 *de novo* proteins in *D. melanogaster* by using homology and synteny approaches similar to other studies (Peng and Zhao 2024). Furthermore, they employed AlphaFold2, ESMFold and RoseTTAFold to predict structures, and demonstrated that the majority of these structures are either partially folded or unstructured, as indicated by pLDDT scores for confidence from each of the three tools. However, they also described several well-folded structures. It is noteworthy that the ancestral sequence reconstruction indicated that these well-folded *de novo* proteins were already well-folded at the time of their origin. Furthermore, a comparison with the PDB database revealed that most of these well-folded *de novo* proteins adopt existing folds, despite the low sequence identity between the sequences responsible for their construction. However, it must be highlighted that these structure prediction methods depend on multiple sequence alignments or they are trained with homologous proteins. Therefore, limitations are expected for the prediction of orphan protein structures which, by definition, lack homologs.

Overall, in *Drosophila*, numerous studies have explored orphan and *de novo* genes, but their functional characterization has been largely restricted to genes associated with reproduction or sex determination. While many orphan genes have been identified, functional validation remains a challenge, emphasizing the need for further studies beyond reproductive traits.

In 2013, Wissler et al. conducted a large-scale comparative genomic analysis to investigate the mechanisms and dynamics of orphan gene emergence in insect genomes, with a particular focus on ants (*Formicidae*) (Wissler et al. 2013). The study revealed that orphan genes make up a substantial fraction of insect genomes, ranging from 10% to over 30% depending on the species analyzed. A key finding was that *de novo* gene birth appears to be the predominant mechanism in *Formicidae*: *de novo* origin accounted for 43.5% to 61.2% of species-specific orphan genes, far exceeding divergence after gene duplication (6.4% to 9.9%) and other mechanisms. The distribution of orphan genes appeared to be largely random across the genome, suggesting widespread and independent emergence events. Notably, several orphan genes exhibited specific expression profiles across tissues or developmental stages, supporting their potential role in lineage-specific traits and ecological adaptations.

In 2021, a group of researchers were interested in orphan genes in another insect, the diamondback moth (*Plutella xylostella)* and they demonstrated two functional orphan genes through a combination of RNA interference (RNAi) and gene expression analyses (T. Li et al.

2021). RNAi silencing of these two genes led to reduced sperm count and decreased motility, significantly impairing male fertility. Further analysis indicated also that these genes are highly expressed in the testes, with one of them in particular showing expression patterns consistent with late-stage spermatogenesis. These findings suggested that these genes contribute to male reproductive success and are likely under strong selection pressures due to their roles in sperm function, highlighting the importance of orphan genes in species-specific reproductive adaptations in *P. xylostella*. Another study in 2024 described another orphan gene in the same species which enhances the male reproductive success (Q. Zhao et al. 2024). The authors demonstrated that this orphan gene called *lushu* encodes a sperm protein and through CRISPR/Cas9-generated mutants lacking this gene, they found out that males exhibited reduced fertility, with lower sperm viability and motility. Expression analysis also showed that *lushu* is highly active in the testes, suggesting a role during sperm maturation. This gene's location on the Z chromosome and its high prevalence in different *P. xylostella* populations suggest it may be under strong selective pressure, likely evolving to meet reproductive demands specific to this species, similar to *Drosophila*.

Nematoda

In 2015, Mayer et al. investigated the role of an orphan gene named dauerless in the *Pristionchus pacificus* necromenic and predatory nematodes, specifically its regulation of dauer development and intraspecific competition (Mayer et al. 2015). The dauer stage is a stress-resistant, non-feeding larval stage in nematodes that allows survival under harsh environmental conditions such as overcrowding or starvation where the metabolism and development are in pause. The study revealed that the *dauerless* gene influences the dauer formation process. The researchers showed that copy number variation (CNV) in the dauerless gene plays a crucial role in regulating the nematode's ability to enter or bypass the dauer stage by several experiments and RNA-seq data. Nematodes with higher copy numbers of the dauerless gene were more likely to suppress dauer formation, which in turn gave them a competitive advantage in environments where resources were limited. This study highlights how CNV in an orphan gene can drive intraspecific competition and influence survival strategies in nematodes.

Following this finding, a study in 2016 described the retroviral origins of an orphan gene, F58H7.5, in *Caenorhabditis elegans* (Kapulkin 2016). While the gene's orphan status was confirmed through direct homology searches, which demonstrated the absence of detectable homologs in other species, the author conducted a comprehensive investigation into its retroviral origins. The study traced the gene back to a potential retroviral insertion, thereby suggesting that exogenous viral elements may have contributed to its emergence within the nematode lineage. Supporting evidence was provided for this hypothesis by identifying sequence similarities between the orphan gene and known retroviral elements, focusing on structural motifs that are typically associated with viral proteins. Furthermore, the integration site of the gene was investigated, demonstrating that the surrounding

genomic region exhibited hallmarks of retroviral insertions, including long terminal repeats (LTRs) and flanking sequences commonly associated with viral integration events. These findings provide compelling evidence for the gene's retroviral origin, elucidating the manner in which viral genetic material was likely co-opted and repurposed for functional use in *C. elegans*. Overall, this constitutes a case of lineage-specific horizontal acquisition of a retroviral element eventually leading to the emergence of an orphan gene lacking homology in other nematodes.

In 2019, another study on *C.elegans* identified 893 orphan genes specific to this species, demonstrating that 4.4% of its protein-coding genes lack homologs in other species (Zhang et al. 2019). Among these, the researchers determined that six genes originated *de novo*. To identify orphan genes, a BLASTP search against closely-related species was performed, which was followed by a BLAST search of coding sequences (CDS) to locate possible non-coding regions in closely related species to be able to identify *de novo* gene candidates. In the identified non-coding regions, the authors searched for the presence of alternative start and stop codons and verified synteny to confirm these candidates as *de novo* genes. Then, similar to previous studies, they verified the expression of these genes via transcriptomic and translation via proteomic data. This multi-step approach allowed them to characterize these genes as recent additions unique to the *C. elegans* lineage, highlighting the potential for *de novo* gene birth in driving species-specific adaptations. The authors found that the expression levels of *de novo* genes are predominantly very low in restricted developmental stages and tissues, but 50% of the identified *de novo* genes showed detectable expression in the dauer stage. Moreover, the study revealed that an important part of these genes were expressed in gonads in adult tissues, which suggest a role in reproduction.

In the same year, Lightfoot et al. uncovered a self-recognition mechanism in *P. pacificus* that prevents cannibalism among its offspring (Lightfoot et al. 2019). The study identified an orphan gene encoding a small peptide, SELF-1, which allows *P. pacificus* to recognize its progeny and avoid consuming them. Through behavioral assays, the researchers demonstrated that *P. pacificus* selectively avoided predation on its own larvae while attacking unrelated larvae, implicating SELF-1 in self-recognition. SELF-1, a 63-amino acid peptide located on the larval surface, has a hypervariable C-terminal region crucial for its function; even a single amino acid change in this region disrupts recognition, leading to cannibalistic behavior. When examining homologs in other nematodes, the team identified SELF-1 as a taxon-restricted orphan gene, suggesting that it either evolved rapidly within *P. pacificus* or emerged *de novo*, providing a unique evolutionary adaptation to enhance survival strategies in competitive environments. This study represents one of the first explorations of orphan genes in behavioral adaptations, with SELF-1 as an example of a gene driving intraspecific recognition.

Later in 2019, another study investigated the whole set of orphan genes in the *Pristionchus* genus (Prabh and Rödelsperger 2019). The authors revealed that approximately 10% of all genes in *Pristionchus* species lack homologs in any other species, while 70% have homologs within *Pristionchus* species, using comparative genomics and phylostratigraphy. Among these, they identified 29 high-confidence species-specific orphan genes in *P. pacificus*, two of which were shown to have emerged *de novo*. To identify these *de novo* genes, the researchers employed tools such as CYNENATOR (Rödelsperger and Dieterich 2010) for synteny analysis and Exonerate (Slater and Birney 2005) for mapping orphan proteins to the genomes of closely related species. Even though they did not provide functional insights, the authors hypothesized that these species-specific genes may contribute to this nematode's ability to thrive in specific environmental niches. Again, it is important to note that the study identified species-specific orphan genes as well as genus-specific ones therefore this must be taken into account when comparing to other studies.

In 2021, Rödelsperger et al. expanded on their research on *P. pacificus*, demonstrating that sperm cells are a source of genomic novelty and rapid evolution in this species, similar to patterns observed in *Drosophila* (Rödelsperger et al. 2021). This study utilized spatially-resolved transcriptome data to map gene expression across distinct anatomical regions in adult nematodes, revealing that sperm cells exhibited particularly high levels of novel gene activity and rapid gene evolution. The authors suggested that many of these novel genes correspond to highly diverged or *de novo* orphan genes identified in their previous research, proposing that sperm-specific regions could drive evolutionary innovation in nematodes by fostering the emergence of new, adaptive genes. Moving on in 2022, Prabh et Rödelsperger also analyzed gene turnover rates in *P. pacificus* to explore the evolutionary dynamics of *de novo* genes compared to duplicated genes (Prabh and Rödelsperger 2022). By sequencing six diverse strains, the study investigated how different origins of genes—*de novo* formation versus duplication—affect their evolutionary persistence and rates of loss. The researchers found that *de novo* genes, aligning with a rapid turnover hypothesis, experience faster rates of both gain and loss. The study highlighted that *de novo* genes remain under weak evolutionary constraints and tend to disappear or evolve rapidly, especially in young age classes. In contrast, duplicated genes showed greater stability and longer retention across evolutionary time scales. These findings suggest that *de novo* genes contribute to genomic innovation, albeit with high rates of attrition, emphasizing the role of gene turnover in shaping *P. pacificus* adaptability and diversity over time.

In 2022, a new study on *C. elegans* uncovered intraspecific *de novo* gene birth by analyzing presence–absence variants (PAVs), a novel approach for identifying genes that are specific to certain strains but absent in others (B. Y. Lee, Kim, and Lee 2022). This study represents a shift from traditional interspecies comparisons to intraspecies analyses, allowing the researchers to capture recently emerged genes within the *C. elegans* lineage. Using long-read sequencing and Iso-Seq technology, the authors sequenced the genomes and

transcriptomes of two strains, *CB4856* and *PD1074*, and identified 46 species-specific genes unique to these strains, many of which are likely *de novo* genes. By employing BLAST and LiftOver (Genovese et al. 2024) for precise gene localization, they confirmed that these genes were either newly formed or lost in the other strains.

Humans and other vertebrates

The studies in model species such as yeast *Drosophila*, and *C. elegans* demonstrated that their genomes comprise a substantial number of orphan genes, which perform a variety of functions. This led researchers to become interested in such genes also in humans. In 2010, a study demonstrated that an orphan gene, according to the evolutionary biology definition of this review, which emerged *de novo*, is associated with human brain functions (C.-Y. Li et al. 2010). The expression of this gene in the brain was confirmed by RT-PCR analysis in multiple tissues, and its orphan status was verified through homology searches against the nr and uniref databases. Subsequently, syntenic genome alignments confirmed that this is a human-specific orphan gene that emerged *de novo*. Furthermore, the study demonstrated that this gene is overexpressed in the brains of individuals with Alzheimer's disease (AD), once again through RT-PCR analysis on 18 healthy brains and 20 AD brains. This identified gene constituted the inaugural example of a *de novo* gene in humans, exhibiting substantial evidence for a function in the brain.

While the 2010 study identified a single *de novo* gene associated with human brain functions, researchers soon expanded their scope to identify *de novo* genes on a genome-wide scale. In 2011, a group of researchers sought to determine the total number of *de novo* genes in humans. They identified 60 such genes (D.-D. Wu, Irwin, and Zhang 2011). To identify them, they searched all human proteins against the sequences of other primates and identified 584 genes specific to humans, i.e. orphan genes. They excluded the ones that did not have start or stop codons in humans and then they performed BLAST analysis against chimpanzee and orangutan genomes with the remaining 352 orphan genes. Then, they identified the ones that had potentially translatable open reading frames and if these regions were disrupted in chimpanzee or orangutan (presence of stop codons, frameshift indels, bad start codons) via a custom pipeline. Finally, they described 60 *de novo* genes, including the *de novo* gene from the 2010 study of the brain. Moreover, the expression levels of these genes in humans, as determined by RNA-seq data on diverse tissues, indicated that the majority of these genes exhibit elevated expression in the cerebral cortex and testes. This observation suggests that these genes may contribute to traits that are exclusive to the human species. However, the *de novo* status of the orphan gene from the 2010 study was contradicted later on in 2024, where the researchers hypothesized that this gene may not have completely emerged *de novo* but diverged from an old pseudogene so highly that we cannot identify a homolog (Hannon Bozorgmehr 2024). Therefore,

although the gene probably emerged from non-coding DNA, the process might be more complex and involve a 'revived' former pseudogene.

Beyond their potential roles in brain development, some orphan genes have been shown to be implicated in disease processes. One notable example is PBOV1, a *de novo* gene linked to cancer progression. In 2013, a study revealed the presence of this gene, with tumor-specific expression particularly in prostate and breast cancers (Samusik et al. 2013). To identify PBOV1 as a *de novo* gene, the authors performed a comparative genomic analysis using MULTIZ multiple genome alignments available from the UCSC Genome Browser to compare the PBOV1 protein-coding sequence (CDS) across 34 genomes of placental mammals. This comparative alignment allowed them to map homologous regions and identify frame-shift mutations and stop codons that disrupt the ORF in non-human species. They then assessed the alignment between human PBOV1 and other mammalian genomes by calculating the fraction of the human CDS that could be aligned to each species. In placental mammal species such as Laurasiatheria and Glires, mutations, such as the loss of the ATG start codon and a 12-base-pair frame-shift deletion, rendered the sequences incapable of producing a similar protein. The genomic analysis showed that while over 99% of the human PBOV1 sequence could be aligned with primate genomes, in non-hominid primates, an early stop codon restricted the protein similarity to 80% of its length. However, this stop codon was mutated in the common ancestor of hominids, restoring the open reading frame and allowing the gene to evolve into a functional protein in humans. Then, similar to other studies, RT-PCR analysis on different tissues revealed that this *de novo* gene is expressed in important part of the cancer types; including breast cancer, cervical, ovary and endometrial cancer, lung cancer, nonHodgkin lymphomas, meningioma and seminoma. Using publicly available microarray datasets, the researchers also found that high levels of PBOV1 expression in breast cancer and glioma samples were significantly associated with positive clinical outcomes. Interestingly, PBOV1 expression was observed in primary but not recurrent high-grade gliomas, suggesting a negative selection against PBOV1-expressing cancer cells.

In 2015, another study revealed 634 human *de novo* genes using BLAST for homology search and synteny for the verification of the *de novo* status (Ruiz-Orera et al. 2015). The analysis of the patterns of tissue expression in assembled transcripts demonstrated that the majority of these genes were expressed in the testis. Conversely, only a few were expressed in the brain, liver, and heart. Consequently, the researchers concluded that *de novo* genes were twice as likely to exhibit testis-restricted expression compared to the rest of the genes in humans.

Then, in 2016, Guerzoni et al. (Guerzoni and McLysaght 2016) investigated the *de novo* emergence of genes in the primate lineage, revealing a slow but consistent rate of new gene formation over evolutionary time. The study utilized similar methods to previous ones to identify *de novo* gene candidates across multiple primate genomes, particularly great apes

such as humans, chimpanzees, orangutans and gorillas. By examining coding and non-coding regions for sequence homology and structural alignments, the authors identified genes with no clear ancestral counterparts in closely related species, establishing their *de novo* origin. One of the key findings was that some *de novo* genes had experienced incomplete lineage sorting (ILS). For instance, in some cases the *de novo* gene was present in humans and gorillas, while in chimpanzees, this is the ancestral non-coding regions that was retained at the same locus. This ILS phenomenon was notably present in genes that showed tissue-specific expression in humans, particularly the brain, suggesting an adaptive role in traits unique to primates. Such instances of ILS suggest *de novo* genes may initially have a neutral effect on fitness and experience a long period of polymorphism prior to fixation. This paper was another example of high impacts of methodology used to identify *de novo* genes. The researchers compared their results with those of Ruiz-Orera et al. (Ruiz-Orera et al. 2015) but found no overlap in the *de novo* gene candidate lists. This is largely explained by filtering-out of intronless genes in the former study, while such genes constitute nearly half of the cases in the new study. The other half is mainly explained as regions not annotated as genes in the version of the databases used in the more recent study.

Moving on, in 2022, Vakirlis et al. (Vakirlis et al. 2022) described the *de novo* birth of functional microproteins in humans. The study focused on microproteins, which are small proteins originating from small open reading frames (sORFs) and are known to have significant fitness effects. To trace their evolutionary origins, the authors performed a comparative analysis across 99 vertebrate species. They reconstructed phylogenetic trees and ancestral sequences to determine when each sORF emerged. If an ancestor lacking an intact ORF was found to precede those with an intact ORF, the ORF was classified as having originated *de novo*. Expression of the *de novo* sORFs was then confirmed using transcriptomic data. Ultimately, the study identified 155 *de novo* microproteins, of which 44 had significant fitness effects, indicating a role in human biological functions. Notably, two of these microproteins likely emerged after the human-chimpanzee split, highlighting their role in human-specific traits and evolution.

In 2023, a group of researchers identified 74 *de novo* genes with long non-coding RNA (lncRNA) origins that play unique roles in human brain development (An et al. 2023). The study concentrated on the evolutionary transition of lncRNAs into protein-coding genes through mechanisms such as RNA splicing and nuclear export. By employing comparative genomics and experimental verification (mass spectrometry and RNA-seq) in human cortical organoids and transgenic mice, the researchers demonstrated that 45 of these genes are human-specific, whereas the remainder are hominoid-specific, having evolved subsequent to the divergence from rhesus macaques. The *de novo* genes were found to contribute to key human brain traits, including cortical development and brain size expansion, thereby emphasizing their potential role in shaping human-specific cognitive abilities. Later on in 2024, a study from Leushkin and Kaessmann contradicted and critically re-evaluated the

findings (Leushkin and Kaessmann 2024). The re-analysis, utilizing various genomic resources and extensive ribosome profiling data, revealed that SMIM45 is, in fact, a mis-annotated part of an ancient and longer vertebrate gene starting just upstream. The authors also identified issues with the remaining loci, indicating that most do not correspond to hominoid-specific *de novo* genes. This study underlined again the necessity for rigorous validation in orphan and *de novo* gene research to accurately determine the origins and evolutionary significance of these genes.

In 2024, another study conducted a comprehensive analysis to identify and characterize human orphan genes across multiple tissues and diseases (Singh et al. 2024). Using extensive RNA-seq data, a self-built pipeline and phylostratigraphy, the researchers discovered thousands of highly expressed transcripts that did not correspond to any previously annotated genes. Approximately 80% of these transcripts contained ORFs with the potential to encode proteins unique to humans. The authors validated these findings using independent strand-specific and single-cell RNA-Seq datasets which confirmed the expression of these novel transcripts. Further differential expression analysis revealed that many of these orphan genes are dynamically regulated, exhibiting selective accumulation in specific tissues, cell types, developmental stages, tumors, and in response to conditions such as COVID-19. In addition, survival analysis indicated that hundreds of these novel transcripts overlapped with deleterious genomic variants, and thousands showed significant associations with disease-specific patient survival, suggesting their potential as diagnostic biomarkers or therapeutic targets.

Lastly, in a recent study in 2024, an investigation was conducted into the evolution of ORFs derived from a single gene, which are separated by a transcriptional silencer. The study demonstrated that one of these ORFs has emerged *de novo* and is likely to play a role in human brain development, as it is one of the identified *de novo* genes in the previous study (Delihas 2024). The *non-de novo* ORF has ancient origins, dating back approximately 462 million years, and is present across different species. The absence of homology has been verified, and the synteny with mouse has shown that at the same position, mouse only has the *non-de novo* ORF. The study also suggested that the transcriptional silencer in between them likely regulates the *de novo* ORF, which provides important evidence of a possible function.

Besides humans, orphan genes have also been studied in other mammals such as mice and other vertebrates such as teleost fish.

In 2022, Petržilek et al. examined the *de novo* emergence, existence, and eventual loss of the gene D6Ertd527e in murine rodents, shedding light on the high turnover rate of *de novo* genes within this lineage (Petrzilek et al. 2022). The researchers used CRISPR-Cas9 gene editing to delete the D6Ertd527e gene in *Mus musculus* to assess its functional role,

specifically targeting the gene's coding regions to produce knock-out models. This deletion resulted in fertile mice with smaller litter. They also conducted RNA-seq across multiple murine species to analyze gene expression, focusing on D6Ertd527e's presence in oocytes and other reproductive tissues. These transcriptomic analyses revealed species-specific expression patterns, suggesting variability in the gene's adaptive significance. Visualization of RNA-seq data helped to map and confirm expression differences between *M. musculus* and other rodents. This approach illustrated how *de novo* genes, although potentially adaptive, can be short-lived under shifting evolutionary pressures, demonstrating D6Ertd527e's emergence and gradual loss within specific rodent lineages.

In 2014, antifreeze glycoprotein genes (AFGPs) in codfishes were studied by Zhuang and it was revealed that codfish AFGPs are orphans and likely have originated from non-coding DNA according to synteny (Zhuang 2014). Then in 2018, another study examined this origin and evolutionary pathway of AFGPs, particularly in the Atlantic rod codfish *Gadus morhua* (Baalsrud et al. 2018). The authors found that AFGPs likely emerged around 13–18 million years ago from non-coding DNA—a remarkable example of *de novo* gene birth. This development coincided with the onset of freezing temperatures in the Northern Hemisphere, supporting the hypothesis that AFGPs provided a survival advantage under extreme conditions. The study employed whole-genome sequencing and comparative genomic analysis using BLAST to trace the origins and distribution of AFGP genes, identifying these genes' presence in multiple codfish lineages and variations in copy numbers across species. They noted a concentration of antifreeze functionality in the sequences, likely evolving from short repetitive tripeptide sequences found in non-coding regions that were repurposed into functional protein sequences for ice-binding. Furthermore, in species exposed to more severe freezing, codfishes show higher copy numbers of AFGP genes, indicating copy number variation as an adaptation to environmental demands. Later on in 2019, another study focused this time on another codfish family, Arctic cod (*Gadidae*) (Zhuang et al. 2019). The researchers found that a short sequence of non-coding DNA underwent repeated duplications, forming a tripeptide repeat sequence (threonine-alanine-alanine) that could bind ice crystals in the blood. Additional events followed: a single nucleotide deletion allowed for proper protein processing and secretion, and a translocation or insertion event provided the transcriptional signals necessary for gene expression regulation. However, another study suggested that this antifreeze orphan gene may not be a *de novo* case but a highly diverged gene from an apolipoprotein homolog (Leushkin and Kaessmann 2024).

Discussion & Conclusion

The study of orphan and *de novo* genes represent a critical area of evolutionary and functional genomics, providing insights into lineage and species-specific adaptations and

biological innovation. They have been consistently identified across a range of animal and fungal species, though the estimated numbers of these genes vary significantly between studies. Model organisms, such as *Drosophila*, *S. cerevisiae*, and humans, are more extensively studied, allowing a clearer understanding of both the prevalence and functional roles of orphan genes within these species. While research in non-model organisms has been more limited, these studies also provide valuable insights into the evolution and potential functions of orphan genes across diverse lineages. Therefore, we know more about orphan genes in model species but it does not mean that they are absent in other species or they do not have important functions, they are just less studied. A summary of orphan genes with known possible functions can be found in Table 1.

In examining various species, it is evident that the number of orphan genes and their representation among protein-coding genes varies significantly. In some species, such as *S. cerevisiae* and *Drosophila*, orphan genes can make up as much as 30% of the protein-coding genes. In contrast, this percentage is lower in species like the fungus *F. graminearum*, the nematodes *P. pacificus*, *C. elegans* or humans in which orphan genes comprise around 4-15% of protein-coding genes. These differences may arise from biological factors, including evolutionary pressures and unique genomic features of each species, as well as methodological variations between studies. Although most studies use similar approaches to identify orphan and *de novo* genes—homology search with comparative genomic tools, phylostratigraphy, alignment on closely related species and syntenic verification to classify *de novo* genes—the specific tools and parameters used can vary considerably. Different studies may apply different thresholds, scoring systems, and criteria leading to differing outcomes in orphan and *de novo* gene identification. Early studies in yeast and *Drosophila* primarily relied on straightforward but likely too simplistic BLAST homology searches with specific e-values against public databases. However, more recent research increasingly incorporates comprehensive pipelines, employing advanced comparative genomic tools such as OrthoFinder (Emms and Kelly 2019), ORFan-Finder (Ekstrom and Yin 2016), OrthoMCL (L. Li, Stoeckert, and Roos 2003), and HMMER (Finn, Clements, and Eddy 2011) to systematically cluster and regroup homologous sequences. Therefore differences can be observed even for the same species with different approaches. Also, it is obvious that with time there were higher-quality annotated genomes available for more and more species, which explains the contradiction to orphan status of some genes in several species. Thus, the relative abundance of orphan genes within a species' genome likely reflects not only inherent biological characteristics but also the diversity of research approaches and criteria used to identify orphan genes. Furthermore, it is also important to note that the orphan aspect varies between studies. Some studies focus on species-specific orphan genes or even on species-specific orphan genes as markers while others on genus-specific ones. This highlights the need for caution when comparing orphan gene counts across studies, as variations in scope can impact results.

Another important consideration is the difference between highly divergent orphan genes and *de novo* genes. Most studies to date have suggested that only a small fraction of orphan genes arise *de novo*. However, in 2020, Vakirlis et al. provided important insights into the origin of orphan genes, challenging the assumption that high sequence divergence from ancestral genes is the primary cause of their orphan status (Vakirlis, Carvunis, and McLysaght 2020). They re-analyzed orphan gene datasets from previous studies spanning multiple taxonomic groups, including yeast, flies, humans, and other vertebrates. Using a synteny-based pipeline developed in-house, they demonstrated that for most orphan genes, there is no clear evidence that they emerged by accumulating high divergence from pre-existing gene sequences, but rather from previously non-coding regions. Such findings highlighted the need for a revised perspective in orphan gene research, encouraging methodologies that are based on examining non-coding regions and transcriptional changes, rather than focusing solely on lack of homologs and sequence divergence. As a result, this study highlighted that *de novo* gene emergence may be more common than previously thought. However, it also suggested that there are limitations in using synteny to determine an ancestor due to genome rearrangements and other evolutionary events. Besides, a study from 2024 suggested other possibilities on the emergence of four of the most known *de novo* genes in different model species, including yeast, *Drosophila* and humans (Hannon Bozorgmehr 2024). Indeed, this study suggested these four genes emerged by re-arrangement and tinkering of previously-existing genes. Hence, understanding the origin and mechanisms of emergence of orphan genes is still a difficult task to accomplish. It depends on methods, genome and predicted proteome quality as well as all the criteria used.

Despite methodological challenges, the functional significance of orphan genes has been demonstrated across diverse species. In humans, *de novo* genes such as PBOV1 and SMIM45 have been linked to cancer progression and brain development, respectively, highlighting their roles in physiological and disease contexts. In fungi, orphan genes like Osp24 in *F. graminearum* mediate host-pathogen interactions by modulating plant immune responses, while lineage-specific genes in ectomycorrhizal fungi are crucial for symbiosis with plant hosts. Similarly, in nematodes, orphan genes such as dauerless and SELF-1 regulate key survival strategies, including dauer development and self-recognition to prevent cannibalism. In codfishes, *de novo* antifreeze glycoproteins provide a survival advantage under freezing conditions, illustrating how environmental pressures can drive functional innovation. These examples demonstrate that orphan and *de novo* genes often evolve to fulfill specialized functions that address unique ecological, developmental, or reproductive challenges faced by their host organisms. This functional versatility underscores the significance of orphan genes as a rich source of evolutionary novelty, shaping specific traits and adaptations.

The study of orphan and *de novo* genes faces challenges, including methodological inconsistencies and difficulties in functional validation. Ensuring accurate identification of these genes remains crucial, therefore there is still a need to define a reference methodology for that. However, advances in sequencing technologies, computational tools, and experimental techniques offer promising solutions to these challenges. By integrating these approaches and fostering interdisciplinary collaboration, future research can deepen our understanding of gene evolution and uncover applications in fields like biomedicine and agriculture.

This review has summarized the progress in understanding the prevalence, origins, and roles of orphan genes, particularly in well-studied model organisms like *Drosophila*, yeast, and humans but also in non-model organisms. Expanding research in non-model organisms highlights that these genes are neither rare nor insignificant in other lineages.

Moving on, paleogenomics will certainly offer a promising way to understand the origins of orphan and *de novo* genes. By comparing modern genomes with those of extinct species, we can identify ancestral homologs and distinguish true *de novo* emergence from cases of high divergence or gene loss. While its application is limited for now, advances in ancient DNA analysis could enhance our understanding of lineage-specific genes. Also, international projects like ERGA and the Darwin Tree of Life are expected to greatly increase the number of high-quality genome assemblies. These efforts will improve comparative analyses across different groups of organisms and help us identify genes in previously underrepresented groups. Also, advances in environmental genomics and metagenomics can show us lineage-specific genes in uncultivated or cryptic organisms, helping us to understand more about gene emergence and diversity in natural populations.

As we look ahead, the study of orphan and *de novo* genes will undoubtedly continue to redefine our understanding of genomic innovation, illuminating the remarkable capacity of life to generate novelty from previously considered ‘junk’ genetic material. This knowledge holds the potential to address key scientific and societal challenges in the years ahead.

Gene or Gene Set	Species/Genus	Orphan status	Possible Function	Reference
SHE genes	*S. cerevisiae*	Orphan	Cell growth (partial)	Espinet et al. (1995)
BSC4	*S. cerevisiae*	*de novo*	DNA repair in stationary phase	Cai et al. (2008)
MDF1	*S. cerevisiae*	*de novo*	Suppression of mating	Li et al. (2010)

YBR196C-A	*S. cerevisiae*	*de novo*	transmembrane protein in ER, involved in fitness.	Vakirlis et al. (2020), Wacholder et al. (2023), Saeki et al. (2023), Houghton et al. (2024)
HUR1	*S. cerevisiae*	Orphan	Involved in DNA repair	Omidi et al. (2018), Wacholder et al. (2023)
ICS3	*S. cerevisiae*	Orphan	Involved in copper homeostasis	Alesso et al. (2015), Wacholder et al. (2023)
YPR096C	*S. cerevisiae*	Orphan	cell fitness (regulates a gene involved in sugar metabolism)	Hajikarimlou et al. (2020), Wacholder et al. (2023)
YDL204W-A	*S. cerevisiae*	Orphan	Cell fitness	Wacholder et al. (2023), Houghton et al. (2024)
Symbiosis-induced genes	ECM fungi	Likely mixed	Symbiosis establishment	Kohler et al. (2015)
296 unique genes	*Z. tritici*	Orphan	Infection-related	Plissonneau et al. (2016)
Osp24	*F. graminearum*	Orphan	Suppression of wheat immunity	Jiang et al. (2020)
Lineage-specific genes	*N. crassa*	Lineage-specific orphans, some likely *de novo*	Reproduction, cell wall integrity	Wang et al. (2022, 2023)
5 de novo testis genes	*D. melanogaster*	*de novo*	Male fertility	Levine et al. (2006)
7 de novo testis genes	*D. yakuba/erecta*	*de novo*	Male fertility	Begun et al. (2007)
142 de novo testis genes	*D. melanogaster*	*de novo*	Male fertility	Zhao et al. (2014)
Female reproductive tract de novo gene	*D. melanogaster*	*de novo*	Female reproduction	Lombardo et al. (2023)
Goddard protein	*D. melanogaster*	*de novo*	Sperm individualization	Lange et al. (2021)
555 de novo proteins	*D. melanogaster*	*de novo*	Mostly implied in fertility	Peng & Zhao (2024)
Tssor-3 and Tssor-4	*P. xylostella*	Orphan	Sperm count, fertility	Li et al. (2021)
Iushu	*P. xylostella*	Orphan	Sperm maturation, motility	Zhao et al. (2024)

PBOV1	Human	*de novo*	Tumor-specific expression	Samusik et al. (2013)
De novo lncRNA-derived genes	Human	*de novo* (debated)	Brain development (human-specific traits)	An et al. (2023)
Thousands of orphan genes	Human	Orphan	Tissue-specific regulation; potential disease links	Singh et al. (2024)
de novo ORF of SMIM45	Human	*de novo*	Brain development	Delihas (2024)
AFGPs	Codfishes (Gadidae)	*de novo* (debated)	Freeze protection	Baalsrud et al. (2018)
D6Ertd527e	Murid rodents	*de novo*	Oocyte expression	Petržilek et al. (2022)
dauerless	*P. pacificus*	Orphan	Dauer development	Mayer et al. (2015)
SELF-1	*P. pacificus*	Orphan	Self-recognition, cannibalism prevention	Lightfoot et al. (2019)
29 species-specific orphans	*P. pacificus*	Orphan	Niche adaptation	Prabh & Rödelsperger (2019)
46 de novo genes	*C. elegans*	*de novo*	Involved in dauer stage and reproduction	Lee et al. (2022)

Table 1: Examples of orphan and de novo genes with possible known functions

References

Alesso, C.A., K.F. Discola, and G. Monteiro. 2015. “The Gene ICS3 from the Yeast Saccharomyces Cerevisiae Is Involved in Copper Homeostasis Dependent on Extracellular pH.” *Fungal Genetics and Biology* 82 (September):43–50. https://doi.org/10.1016/j.fgb.2015.06.007.

An, Ni A., Jie Zhang, Fan Mo, Xuke Luan, Lu Tian, Qing Sunny Shen, Xiangshang Li, et al. 2023. “De Novo Genes with an lncRNA Origin Encode Unique Human Brain Developmental Functionality.” *Nature Ecology & Evolution* 7 (2): 264–78. https://doi.org/10.1038/s41559-022-01925-6.

Andaluz, Encarnación, Juan-José R. Coque, Rosario Cueva, and Germán Larriba. 2001. “Sequencing of a 4.3 Kbp Region of Chromosome 2 of *Candida Albicans* Reveals the Presence of Homologues of *SHE9* from *Saccharomyces Cerevisiae* and of Bacterial Phosphatidylinositol-phospholipase C.” *Yeast* 18 (8): 711–21. https://doi.org/10.1002/yea.716.

Baalsrud, Helle Tessand, Ole Kristian Tørresen, Monica Hongrø Solbakken, Walter Salzburger, Reinhold Hanel, Kjetill S Jakobsen, and Sissel Jentoft. 2018. “De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole

Genome Sequence Data." *Molecular Biology and Evolution* 35 (3): 593–606. https://doi.org/10.1093/molbev/msx311.
Begun, David J, Heather A Lindfors, Andrew D Kern, and Corbin D Jones. 2007. "Evidence for *de Novo* Evolution of Testis-Expressed Genes in the *Drosophila Yakuba* / *Drosophila Erecta* Clade." *Genetics* 176 (2): 1131–37. https://doi.org/10.1534/genetics.106.069245.
Cai, Jing, Ruoping Zhao, Huifeng Jiang, and Wen Wang. 2008. "*De Novo* Origination of a New Protein-Coding Gene in *Saccharomyces Cerevisiae*." *Genetics* 179 (1): 487–96. https://doi.org/10.1534/genetics.107.084491.
Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, et al. 2012. "Proto-Genes and de Novo Gene Birth." *Nature* 487 (7407): 370–74. https://doi.org/10.1038/nature11184.
Chapman, Karen B., and Jef D. Boeke. 1991. "Isolation and Characterization of the Gene Encoding Yeast Debranching Enzyme." *Cell* 65 (3): 483–92. https://doi.org/10.1016/0092-8674(91)90466-C.
Cosentino, Salvatore, and Wataru Iwasaki. 2023. "SonicParanoid2: Fast, Accurate, and Comprehensive Orthology Inference with Machine Learning and Language Models." Preprint. Bioinformatics. https://doi.org/10.1101/2023.05.14.540736.
Delihas, Nicholas. 2024. "Evolution of a Human-Specific De Novo Open Reading Frame and Its Linked Transcriptional Silencer." *International Journal of Molecular Sciences* 25 (7): 3924. https://doi.org/10.3390/ijms25073924.
Dohmen, Elias, Margaux Aubel, Lars A. Eicholt, Paul Roginski, Victor Luria, Amir Karger, and Anna Grandchamp. 2025. "DeNoFo: A File Format and Toolkit for Standardised, Comparable de Novo Gene Annotation." Bioinformatics. https://doi.org/10.1101/2025.03.31.644673.
Domazet-Loso, Tomislav, and Diethard Tautz. 2003. "An Evolutionary Analysis of Orphan Genes in *Drosophila*." *Genome Research* 13 (10): 2213–19. https://doi.org/10.1101/gr.1311003.
Dujon, Bernard. 1996. "The Yeast Genome Project: What Did We Learn?" *Trends in Genetics* 12 (7): 263–70. https://doi.org/10.1016/0168-9525(96)10027-5.
Ekstrom, Alex, and Yanbin Yin. 2016. "ORFanFinder: Automated Identification of Taxonomically Restricted Orphan Genes." *Bioinformatics* 32 (13): 2053–55. https://doi.org/10.1093/bioinformatics/btw122.
Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y.
Espinet, Carme, Maria Angeles De La Torre, Marti Aldea, and Enrique Herrero. 1995. "An Efficient Method to Isolate Yeast Genes Causing Overexpression-mediated Growth Arrest." *Yeast* 11 (1): 25–32. https://doi.org/10.1002/yea.320110104.
Fakhar, Ali Zeeshan, Jinbao Liu, Karolina M. Pajerowska-Mukhtar, and M. Shahid Mukhtar. 2023. "The Lost and Found: Unraveling the Functions of Orphan Genes." *Journal of Developmental Biology* 11 (2): 27. https://doi.org/10.3390/jdb11020027.
Finn, R. D., J. Clements, and S. R. Eddy. 2011. "HMMER Web Server: Interactive Sequence Similarity Searching." *Nucleic Acids Research* 39 (suppl): W29–37. https://doi.org/10.1093/nar/gkr367.
Genovese, Giulio, Nicole B Rockweiler, Bryan R Gorman, Tim B Bigdeli, Michelle T Pato, Carlos N Pato, Kiku Ichihara, and Steven A McCarroll. 2024. "BCFtools/Liftover: An Accurate and Comprehensive Tool to Convert Genetic Variants across Genome Assemblies." Edited by Christina Kendziorski. *Bioinformatics* 40 (2): btae038. https://doi.org/10.1093/bioinformatics/btae038.
Gould, Genevieve M., Joseph M. Paggi, Yuchun Guo, David V. Phizicky, Boris Zinshteyn, Eric T. Wang, Wendy V. Gilbert, David K. Gifford, and Christopher B. Burge. 2016. "Identification of New Branch Points and Unconventional Introns in *Saccharomyces Cerevisiae*." *RNA* 22 (10): 1522–34. https://doi.org/10.1261/rna.057216.116.

Grandchamp, Anna, Margaux Aubel, Lars Eicholt, Paul Roginski, Victor Luria, Amir Karger, and Elias Dohmen. 2025. “De Novo Gene Emergence: Summary, Classification, and Challenges of Current Methods.” Bioinformatics. https://doi.org/10.32942/X2DP88.

Guerzoni, Daniele, and Aoife McLysaght. 2016. “De Novo Genes Arise at a Slow but Steady Rate along the Primate Lineage and Have Been Subject to Incomplete Lineage Sorting.” *Genome Biology and Evolution* 8 (4): 1222–32. https://doi.org/10.1093/gbe/evw074.

Hajikarimlou, Maryam, Houman Moteshareie, Katayoun Omidi, Mohsen Hooshyar, Sarah Shaikho, Tom Kazmirchuk, Daniel Burnside, et al. 2020. “Sensitivity of Yeast to Lithium Chloride Connects the Activity of YTA6 and YPR096C to Translation of Structured mRNAs.” Edited by Arthur J. Lustig. *PLOS ONE* 15 (7): e0235033. https://doi.org/10.1371/journal.pone.0235033.

Hannon Bozorgmehr, Joseph. 2024. “Four Classic ‘de Novo’ Genes All Have Plausible Homologs and Likely Evolved from Retro-Duplicated or Pseudogenic Sequences.” *Molecular Genetics and Genomics* 299 (1): 6. https://doi.org/10.1007/s00438-023-02090-6.

Hartig, Monika B., Arcangela Iuso, Tobias Haack, Tomasz Kmiec, Elzbieta Jurkiewicz, Katharina Heim, Sigrun Roeber, et al. 2011. “Absence of an Orphan Mitochondrial Protein, C19orf12, Causes a Distinct Clinical Subtype of Neurodegeneration with Brain Iron Accumulation.” *The American Journal of Human Genetics* 89 (4): 543–50. https://doi.org/10.1016/j.ajhg.2011.09.007.

Heames, Brennen, Jonathan Schmitz, and Erich Bornberg-Bauer. 2020. “A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila.” *Journal of Molecular Evolution* 88 (4): 382–98. https://doi.org/10.1007/s00239-020-09939-z.

Hibbs, Matthew A., David C. Hess, Chad L. Myers, Curtis Huttenhower, Kai Li, and Olga G. Troyanskaya. 2007. “Exploring the Functional Landscape of Gene Expression: Directed Search of Large Microarray Compendia.” *Bioinformatics* 23 (20): 2692–99. https://doi.org/10.1093/bioinformatics/btm403.

Houghton, Carly J., Nelson Castilho Coelho, Annette Chiang, Stefanie Hedayati, Saurin B. Parikh, Nejla Ozbaki-Yagan, Aaron Wacholder, et al. 2024. “Cellular Integration of Beneficial *de Novo* Proteins via Ancient Pathways.” Evolutionary Biology. https://doi.org/10.1101/2024.08.28.610198.

Jacob, François. 1977. “Evolution and Tinkering.” *Science* 196 (4295): 1161–66. https://doi.org/10.1126/science.860134.

Jiang, Cong, Ruonan Hei, Yang Yang, Shijie Zhang, Qinhu Wang, Wei Wang, Qiang Zhang, et al. 2020. “An Orphan Protein of Fusarium Graminearum Modulates Host Immunity by Mediating Proteasomal Degradation of TaSnRK1α.” *Nature Communications* 11 (1): 4382. https://doi.org/10.1038/s41467-020-18240-y.

Kapulkin, Wadim J. 2016. “Retroviral Origins of the *Caenorhabditis Elegans* Orphan Gene F58H7.5.” https://doi.org/10.1101/073510.

Khalturin, Konstantin, Georg Hemmrich, Sebastian Fraune, René Augustin, and Thomas C.G. Bosch. 2009. “More than Just Orphans: Are Taxonomically-Restricted Genes Important in Evolution?” *Trends in Genetics* 25 (9): 404–13. https://doi.org/10.1016/j.tig.2009.07.006.

Lange, Andreas, Prajal H. Patel, Brennen Heames, Adam M. Damry, Thorsten Saenger, Colin J. Jackson, Geoffrey D. Findlay, and Erich Bornberg-Bauer. 2021. “Structural and Functional Characterization of a Putative de Novo Gene in Drosophila.” *Nature Communications* 12 (1): 1667. https://doi.org/10.1038/s41467-021-21667-6.

Lee, Bo Yun, Jun Kim, and Junho Lee. 2022. “Intraspecific *de Novo* Gene Birth Revealed by Presence–Absence Variant Genes in *Caenorhabditis Elegans*.” *NAR Genomics and Bioinformatics* 4 (2): lqac031. https://doi.org/10.1093/nargab/lqac031.

Leushkin, Evgeny, and Henrik Kaessmann. 2024. “Identification of Old Coding Regions Disproves the Hominoid de Novo Status of Genes.” *Nature Ecology & Evolution* 8 (10): 1826–30. https://doi.org/10.1038/s41559-024-02513-6.

Levine, Mia T., Corbin D. Jones, Andrew D. Kern, Heather A. Lindfors, and David J. Begun.

2006. “Novel Genes Derived from Noncoding DNA in *Drosophila Melanogaster* Are Frequently X-Linked and Exhibit Testis-Biased Expression.” *Proceedings of the National Academy of Sciences* 103 (26): 9935–39. https://doi.org/10.1073/pnas.0509809103.

Li, Chuan-Yun, Yong Zhang, Zhanbo Wang, Yan Zhang, Chunmei Cao, Ping-Wu Zhang, Shu-Juan Lu, et al. 2010. “A Human-Specific De Novo Protein-Coding Gene Associated with Human Brain Functions.” Edited by Philip E. Bourne. *PLoS Computational Biology* 6 (3): e1000734. https://doi.org/10.1371/journal.pcbi.1000734.

Li, Dan, Yang Dong, Yu Jiang, Huifeng Jiang, Jing Cai, and Wen Wang. 2010. “A de Novo Originated Gene Depresses Budding Yeast Mating Pathway and Is Repressed by the Protein Encoded by Its Antisense Strand.” *Cell Research* 20 (4): 408–20. https://doi.org/10.1038/cr.2010.31.

Li, Li, Christian J. Stoeckert, and David S. Roos. 2003. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” *Genome Research* 13 (9): 2178–89. https://doi.org/10.1101/gr.1224503.

Li, Tian-pu, Li-wen Zhang, Ya-qing Li, Min-sheng You, and Qian Zhao. 2021. “Functional Analysis of the Orphan Genes Tssor-3 and Tssor-4 in Male Plutella Xylostella.” *Journal of Integrative Agriculture* 20 (7): 1880–88. https://doi.org/10.1016/S2095-3119(21)63655-9.

Lightfoot, James W., Martin Wilecki, Christian Rödelsperger, Eduardo Moreno, Vladislav Susoy, Hanh Witte, and Ralf J. Sommer. 2019. “Small Peptide–Mediated Self-Recognition Prevents Cannibalism in Predatory Nematodes.” *Science* 364 (6435): 86–89. https://doi.org/10.1126/science.aav9856.

Lombardo, Kaelina D, Hayley K Sheehy, Julie M Cridland, and David J Begun. 2023. “Identifying Candidate de Novo Genes Expressed in the Somatic Female Reproductive Tract of *Drosophila Melanogaster*.” Edited by S Macdonald. *G3: Genes, Genomes, Genetics* 13 (8): jkad122. https://doi.org/10.1093/g3journal/jkad122.

Lu, Tzu-Chiao, Jun-Yi Leu, and Wen-Chang Lin. 2017. “A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces Sensu Stricto Yeasts.” *Molecular Biology and Evolution* 34 (11): 2823–38. https://doi.org/10.1093/molbev/msx210.

Mayer, Melanie G., Christian Rödelsperger, Hanh Witte, Metta Riebesell, and Ralf J. Sommer. 2015. “The Orphan Gene Dauerless Regulates Dauer Development and Intraspecific Competition in Nematodes by Copy Number Variation.” Edited by Stuart K. Kim. *PLOS Genetics* 11 (6): e1005146. https://doi.org/10.1371/journal.pgen.1005146.

McGinnis, S., and T. L. Madden. 2004. “BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools.” *Nucleic Acids Research* 32 (Web Server): W20–25. https://doi.org/10.1093/nar/gkh435.

McLysaght, Aoife, and Laurence D. Hurst. 2016. “Open Questions in the Study of de Novo Genes: What, How and Why.” *Nature Reviews Genetics* 17 (9): 567–78. https://doi.org/10.1038/nrg.2016.78.

Mycorrhizal Genomics Initiative Consortium, Annegret Kohler, Alan Kuo, Laszlo G Nagy, Emmanuelle Morin, Kerrie W Barry, Francois Buscot, et al. 2015. “Convergent Losses of Decay Mechanisms and Rapid Turnover of Symbiosis Genes in Mycorrhizal Mutualists.” *Nature Genetics* 47 (4): 410–15. https://doi.org/10.1038/ng.3223.

Nothacker, Hans-Peter. 2008. “Orphan Receptors.” In *Encyclopedia of Molecular Pharmacology*, edited by Stefan Offermanns and Walter Rosenthal, 914–17. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-38918-7_224.

Oliver, S. G., Q. J. M. Van Der Aart, M. L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, et al. 1992. “The Complete DNA Sequence of Yeast Chromosome III.” *Nature* 357 (6373): 38–46. https://doi.org/10.1038/357038a0.

Omidi, Katayoun, Matthew Jessulat, Mohsen Hooshyar, Daniel Burnside, Andrew

Schoenrock, Tom Kazmirchuk, Maryam Hajikarimlou, et al. 2018. “Uncharacterized ORF HUR1 Influences the Efficiency of Non-Homologous End-Joining Repair in Saccharomyces Cerevisiae.” *Gene* 639 (January):128–36. https://doi.org/10.1016/j.gene.2017.10.003.

Peng, Junhui, and Li Zhao. 2024. “The Origin and Structural Evolution of de Novo Genes in Drosophila.” *Nature Communications* 15 (1): 810. https://doi.org/10.1038/s41467-024-45028-1.

Petrzilek, Jan, Josef Pasulka, Radek Malik, Filip Horvat, Shubhangini Kataruka, Helena Fulka, and Petr Svoboda. 2022. “De Novo Emergence, Existence, and Demise of a Protein-Coding Gene in Murids.” *BMC Biology* 20 (1): 272. https://doi.org/10.1186/s12915-022-01470-5.

Plissonneau, Clémence, Alessandra Stürchler, and Daniel Croll. 2016. “The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat.” Edited by John W. Taylor. *mBio* 7 (5): e01231-16. https://doi.org/10.1128/mBio.01231-16.

Prabh, Neel, and Christian Rödelsperger. 2019. “*De Novo* , Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in *Pristionchus* Nematodes.” *G3 Genes|Genomes|Genetics* 9 (7): 2277–86. https://doi.org/10.1534/g3.119.400326.

———. 2022. “Multiple *Pristionchus Pacificus* Genomes Reveal Distinct Evolutionary Dynamics between de Novo Candidates and Duplicated Genes.” *Genome Research* 32 (7): 1315–27. https://doi.org/10.1101/gr.276431.121.

Rödelsperger, Christian, and Christoph Dieterich. 2010. “CYNTENATOR: Progressive Gene Order Alignment of 17 Vertebrate Genomes.” Edited by Sridhar Hannenhalli. *PLoS ONE* 5 (1): e8861. https://doi.org/10.1371/journal.pone.0008861.

Rödelsperger, Christian, Annabel Ebbing, Devansh Raj Sharma, Misako Okumura, Ralf J Sommer, and Hendrik C Korswagen. 2021. “Spatial Transcriptomics of Nematodes Identifies Sperm Cells as a Source of Genomic Novelty and Rapid Evolution.” Edited by Ilya Ruvinsky. *Molecular Biology and Evolution* 38 (1): 229–43. https://doi.org/10.1093/molbev/msaa207.

Roginski, Paul, Anna Grandchamp, Chloé Quignot, and Anne Lopes. 2024. “DE Novo Emerged Gene SEarch in Eukaryotes with DENSE.” Preprint. Genomics. https://doi.org/10.1101/2024.01.30.578014.

Rubin, Gerald M., Mark D. Yandell, Jennifer R. Wortman, George L. Gabor, Miklos, Catherine R. Nelson, Iswar K. Hariharan, et al. 2000. “Comparative Genomics of the Eukaryotes.” *Science* 287 (5461): 2204–15. https://doi.org/10.1126/science.287.5461.2204.

Ruiz-Orera, Jorge, Jessica Hernandez-Rodriguez, Cristina Chiva, Eduard Sabidó, Ivanela Kondova, Ronald Bontrop, Tomàs Marqués-Bonet, and M.Mar Albà. 2015. “Origins of De Novo Genes in Human and Chimpanzee.” Edited by James Noonan. *PLOS Genetics* 11 (12): e1005721. https://doi.org/10.1371/journal.pgen.1005721.

Saeki, Nozomu, Chie Yamamoto, Yuichi Eguchi, Takayuki Sekito, Shuji Shigenobu, Mami Yoshimura, Yoko Yashiroda, Charles Boone, and Hisao Moriya. 2023. “Overexpression Profiling Reveals Cellular Requirements in the Context of Genetic Backgrounds and Environments.” Edited by Joseph Schacherer. *PLOS Genetics* 19 (4): e1010732. https://doi.org/10.1371/journal.pgen.1010732.

Samusik, Nikolay, Larisa Krukovskaya, Irina Meln, Evgeny Shilov, and Andrey P. Kozlov. 2013. “PBOV1 Is a Human De Novo Gene with Tumor-Specific Expression That Is Associated with a Positive Clinical Outcome of Cancer.” Edited by Ludmila Prokunina-Olsson. *PLoS ONE* 8 (2): e56162. https://doi.org/10.1371/journal.pone.0056162.

Schmitz, Jonathan F, and Erich Bornberg-Bauer. 2017. “Fact or Fiction: Updates on How Protein-Coding Genes Might Emerge de Novo from Previously Non-Coding DNA.” *F1000Research* 6 (January):57. https://doi.org/10.12688/f1000research.10079.1.

Singh, Urminder, Jeffrey A. Haltom, Joseph W. Guarnieri, Jing Li, Arun Seetharam, Afshin Beheshti, Bruce Aronow, and Eve Syrkin Wurtele. 2024. “A Pan-Tissue, Pan-Disease Compendium of Human Orphan Genes.” Evolutionary Biology.

https://doi.org/10.1101/2024.02.21.581488.
Slater, Guy St C, and Ewan Birney. 2005. “Automated Generation of Heuristics for Biological Sequence Comparison.” *BMC Bioinformatics* 6 (1): 31. https://doi.org/10.1186/1471-2105-6-31.
Stoye, J, D Evers, and F Meyer. 1998. “Rose: Generating Sequence Families.” *Bioinformatics* 14 (2): 157–63. https://doi.org/10.1093/bioinformatics/14.2.157.
Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. “The Evolutionary Origin of Orphan Genes.” *Nature Reviews Genetics* 12 (10): 692–702. https://doi.org/10.1038/nrg3053.
Vakirlis, Nikolaos, Omer Acar, Brian Hsu, Nelson Castilho Coelho, S. Branden Van Oss, Aaron Wacholder, Kate Medetgul-Ernar, et al. 2020. “De Novo Emergence of Adaptive Membrane Proteins from Thymine-Rich Genomic Sequences.” *Nature Communications* 11 (1): 781. https://doi.org/10.1038/s41467-020-14500-z.
Vakirlis, Nikolaos, Anne-Ruxandra Carvunis, and Aoife McLysaght. 2020. “Synteny-Based Analyses Indicate That Sequence Divergence Is Not the Main Source of Orphan Genes.” *eLife* 9 (February):e53500. https://doi.org/10.7554/eLife.53500.
Vakirlis, Nikolaos, Alex S Hebert, Dana A Opulente, Guillaume Achaz, Chris Todd Hittinger, Gilles Fischer, Joshua J Coon, and Ingrid Lafontaine. 2018. “A Molecular Portrait of De Novo Genes in Yeasts.” *Molecular Biology and Evolution* 35 (3): 631–45. https://doi.org/10.1093/molbev/msx315.
Vakirlis, Nikolaos, Zoe Vance, Kate M. Duggan, and Aoife McLysaght. 2022. “De Novo Birth of Functional Microproteins in the Human Lineage.” *Cell Reports* 41 (12): 111808. https://doi.org/10.1016/j.celrep.2022.111808.
Van Oss, Stephen Branden, and Anne-Ruxandra Carvunis. 2019. “De Novo Gene Birth.” *PLOS Genetics* 15 (5): e1008160. https://doi.org/10.1371/journal.pgen.1008160.
Wacholder, Aaron, Saurin Bipin Parikh, Nelson Castilho Coelho, Omer Acar, Carly Houghton, Lin Chou, and Anne-Ruxandra Carvunis. 2023. “A Vast Evolutionarily Transient Translatome Contributes to Phenotype and Fitness.” *Cell Systems* 14 (5): 363-381.e8. https://doi.org/10.1016/j.cels.2023.04.002.
Wang, Yen-Wen, Jaqueline Hess, Jason C Slot, and Anne Pringle. 2020. “De Novo Gene Birth, Horizontal Gene Transfer, and Gene Duplication as Sources of New Gene Families Associated with the Origin of Symbiosis in *Amanita*.” Edited by Li-Jun Ma. *Genome Biology and Evolution* 12 (11): 2168–82. https://doi.org/10.1093/gbe/evaa193.
Wang, Zheng, Yaning Wang, Takao Kasuga, Yen-Wen Wang, Francesc Lopez-Giraldez, Yang Zhang, Zhang Zhang, et al. 2022. “Lineage-Specific Genes Are Clustered with Allorecognition Loci and Respond to G × E Factors Regulating the Switch from Asexual to Sexual Reproduction in *Neurospora*.” Evolutionary Biology. https://doi.org/10.1101/2022.06.10.495464.
Wang, Zheng, Yen-Wen Wang, Takao Kasuga, Hayley Hassler, Francesc Lopez-Giraldez, Caihong Dong, Oded Yarden, and Jeffrey P. Townsend. 2023. “Origins of Lineage-specific Elements via Gene Duplication, Relocation, and Regional Rearrangement in *Neurospora Crassa*.” *Molecular Ecology*, October, mec.17168. https://doi.org/10.1111/mec.17168.
Wang, Zhong, Qifang Jin, Qin Li, Xingchang Ou, Shi Li, Zhonghua Liu, and Jian’an Huang. 2022. “Multiplex PCR Identification of Aspergillus Cristatus and Aspergillus Chevalieri in Liupao Tea Based on Orphan Genes.” *Foods* 11 (15): 2217. https://doi.org/10.3390/foods11152217.
Weisman, Caroline M. 2022. “The Origins and Functions of De Novo Genes: Against All Odds?” *Journal of Molecular Evolution* 90 (3–4): 244–57. https://doi.org/10.1007/s00239-022-10055-3.
Wissler, Lothar, Jürgen Gadau, Daniel F. Simola, Martin Helmkampf, and Erich Bornberg-Bauer. 2013. “Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes.” *Genome Biology and Evolution* 5 (2): 439–55. https://doi.org/10.1093/gbe/evt009.
Wu, Baojun, and Alicia Knudson. 2018. “Tracing the *De Novo* Origin of Protein-Coding

Genes in Yeast.” Edited by John W. Taylor. *mBio* 9 (4): e01024-18. https://doi.org/10.1128/mBio.01024-18.
Wu, Dong-Dong, David M. Irwin, and Ya-Ping Zhang. 2011. “De Novo Origin of Human Protein-Coding Genes.” Edited by David J. Begun. *PLoS Genetics* 7 (11): e1002379. https://doi.org/10.1371/journal.pgen.1002379.
Zhang, Wenyu, Yuanxiao Gao, Manyuan Long, and Bairong Shen. 2019. “Origination and Evolution of Orphan Genes and de Novo Genes in the Genome of Caenorhabditis Elegans.” *Science China Life Sciences* 62 (4): 579–93. https://doi.org/10.1007/s11427-019-9482-0.
Zhao, L., P. Saelao, C. D. Jones, and D. J. Begun. 2014. “Origin and Spread of de Novo Genes in Drosophila Melanogaster Populations.” *Science* 343 (6172): 769–72. https://doi.org/10.1126/science.1248286.
Zhao, Li, Nicolas Svetec, and David J. Begun. 2024. “De Novo Genes.” *Annual Review of Genetics* 58 (1): 211–32. https://doi.org/10.1146/annurev-genet-111523-102413.
Zhao, Qian, Yahong Zheng, Yiying Li, Lingping Shi, Jing Zhang, Dongna Ma, and Minsheng You. 2024. “An Orphan Gene Enhances Male Reproductive Success in *Plutella Xylostella*.” Edited by Grace Yuh Chwen Lee. *Molecular Biology and Evolution* 41 (7): msae142. https://doi.org/10.1093/molbev/msae142.
Zhuang, Xuan. 2014. “Creating Sense from Non-Sense DNA: De Novo Genesis and Evolutionary History of Antifreeze Glycoprotein Gene in Northern Cod Fishes (Gadidae).” University of Illinois at Urbana-Champaign.
Zhuang, Xuan, Chun Yang, Katherine R. Murphy, and C.-H. Christina Cheng. 2019. “Molecular Mechanism and History of Non-Sense to Sense Evolution of Antifreeze Glycoprotein Gene in Northern Gadids.” *Proceedings of the National Academy of Sciences* 116 (10): 4400–4405. https://doi.org/10.1073/pnas.1817138116.