De Novo Gene Emergence: Summary, Classification, and Challenges of Current Methods

Anna Grandchamp[1], Margaux Aubel[2], Lars A. Eicholt[2], Paul Roginski[3], Victor Luria[4, 5, 6], Amir Karger[7], Elias Dohmen[2]

[1]Aix Marseille University, INSERM, TAGC, UMR_S1090, Marseille, France

[2]Institute for Evolution and Biodiversity, University of Münster, Münster, 48149, Germany

[3]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

[4]Department of Neuroscience, Yale School of Medicine, New Haven, 06510, CT, USA

[5]Department of Systems Biology, Harvard Medical School, Boston, 02115, MA, USA

[6]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, 02115, MA, USA

[7]IT-Research Computing, Harvard Medical School, Boston, 02115, MA, USA

*Corresponding author: anna.grandchamp@inserm.fr

Abstract

A novel mechanism of *de novo* gene origination from non-genic sequences was first proposed in the early 2000s. Subsequent studies have since provided evidence of *de novo* gene emergence across all domains of life, revealing its occurrence to be more frequent than initially anticipated. While studies mainly agree on the general concept of *de novo* emergence from non-genic DNA, the exact methods and definitions for detecting *de novo* genes differ significantly.

Here, we provide a comprehensive step-by-step description of the most commonly used methods for *de novo* gene detection. In addition, we address the limitations of nomenclature and detection methods and clarify some complex concepts that are sometimes misused.

This review is accompanied by the publication of a *de novo* gene annotation format to standardise the reporting of methodology, enable reproducibility and improve the comparability of datasets.

1 Introduction

Throughout evolution, genes can arise by 'recycling the old', emerging from pre-existing genetic material through mechanisms such as duplication (Ohno, 1970), exon shuffling (Gilbert, 1978), horizontal gene transfer (Griffith, 1928; Freeman, 1951), retrotransposition (Baltimore, 1970; Temin et al., 1970; Coffin and Fan, 2016) and gene fusion (Mitelman et al., 2007; Nowell and Hungerford, 2004).

However, it is now well documented that new genes can also emerge *de novo*, through a series of mutations in the non-coding genome (Begun et al., 2006; Tautz and Domazet-Lošo, 2011; Begun et al., 2007; Toll-Riera et al., 2009; Rancurel et al., 2009; Tautz and Domazet-Lošo, 2011; Heinen et al., 2009; Neme and Tautz, 2013; Zhao et al., 2014; Xia et al., 2025). Several mechanisms of *de novo* gene emergence have been identified, including overprinting (Keese and Gibbs, 1992; Pavesi, 2006; Rogozin et al., 2002; Delaye et al., 2008), exonisation (Schmitz and Brosius, 2011; Sorek, 2007; Schmitz and Brosius, 2011; Cai et al., 2008), gene antisense emergence (Thomas et al., 2023; Ardern et al., 2020), emergence from scratch in intergenic regions (Schlötterer, 2015; Iyengar and Bornberg-Bauer, 2023; McLysaght and Guerzoni, 2015; Papadopoulos et al., 2021; Heames et al., 2020; Lombardo et al., 2023), and genomic reshuffling through transposable element (TE) insertion (Schlötterer, 2015; Iyengar and Bornberg-Bauer, 2023; McLysaght and Guerzoni, 2015; Papadopoulos et al., 2021; Heames et al., 2020; Lombardo et al., 2023). Despite their different origins, these mechanisms share a common feature: the *de novo* gene or its encoded protein lack detectable similarity to any other known gene or protein (McLysaght and Hurst, 2016).

One of the major challenges in *de novo* gene research is to accurately determine whether a gene truly emerged *de novo* or has arisen through other mechanisms (Tautz and Domazet-Lošo, 2011; Casola, 2018). For example, after a duplication event, the duplicated gene copy can evolve rapidly and its sequence can undergo significant rearrangement (Innan and Kondrashov, 2010) so that it is misidentified as originating *de novo*. The work of (Casola, 2018) shed light on inaccuracies in the validation of *de novo* gene emergence, and was followed by significant advances in the precision of detection and the design of pipelines for confirming *de novo* origins. As methods for *de novo* gene detection and validation have become more sophisticated, proper annotation of the methodology has become essential (Weisman et al., 2022; Moyers and Zhang, 2016, 2017).

In the field of *de novo* gene research, the mechanisms and definitions of *de novo* emergence remain a pivotal yet variable factor in identifying such genes. Across studies, authors have incorporated diverse evolutionary stages and criteria (Keeling et al., 2019; Weisman, 2022), such as varying thresholds for how much of a gene must have originated *de novo* (McLysaght and Hurst, 2016), and differing standards to establish the absence of homology (Casola, 2018; Vakirlis et al., 2020; Weisman et al., 2022). Although this conceptual diversity has enriched the field, it has also introduced ambiguities that challenge the consistency and comparability of results (Schmitz et al., 2018) (Dohmen et al., 2025). At this stage, maintaining an openness to exploring various methodologies remains critical, but addressing these semantic and conceptual divergences is equally important to advance the field and improve the integration of findings across studies.

In this review, we outline the key steps that currently allow for accurate discrimination between *de novo* genes and genes arising from other mechanisms. We also highlight the main methodological differences between studies and address the challenges and controversies that remain with current approaches. As a consequence of the differences in methods and approaches identified here, we have developed an annotation format to standardise the reporting of the methodology used, and allow for easy comparison between datasets (Dohmen et al., 2025).

2 Tools and Techniques in the Computational Detection of *De Novo* Genes

2.1 Choice of Candidate Genes

The initial step in the identification of *de novo* genes or proto-genes is the selection of candidate genes from a given species, population or individual. Unless a subset of genes has already been identified as candidate *de novo* genes, often, the entire genome or transcriptome is screened to distinguish *de novo* genes from others. Importantly, in the present article, the definition of *de novo* genes assumes the presence of a transcribed ORF, even though the definition of a gene does not always require a coding status (Orgogozo et al., 2016; Li and Liu, 2019). Two distinct approaches are commonly employed in the identification of *de novo* genes: the first involves the assessment of annotated genes within an annotated genome, while the second entails the evaluation of ORFs extracted from a transcriptome, sometimes accompanied by the validation of translation.

2.1.1 Candidate Genes from Annotated Genomes

The identification of potential *de novo* genes in an annotated genome consists in determining which annotated genes correspond to genes that have potentially emerged *de novo* in a specified taxonomic group. Annotated genomes can be obtained from public databases, such as NCBI (Schoch et al., 2020), or they can be obtained through genome assembly from DNA-seq data. In the latter case, it is necessary to annotate the genomes. In the specific context of *de novo* gene detection, a combination of homology-based approaches (Eddy, 2009; Söding, 2005) with *ab initio* approaches (Scalzitti et al., 2020; Baker et al., 2023; Wang et al., 2004) is encouraged, given that the latter relies on algorithms that recognize various genic properties within a genome even without gene homology (Figure 1 a, Table 1).

2.1.2 Candidate Genes from Transcriptomes

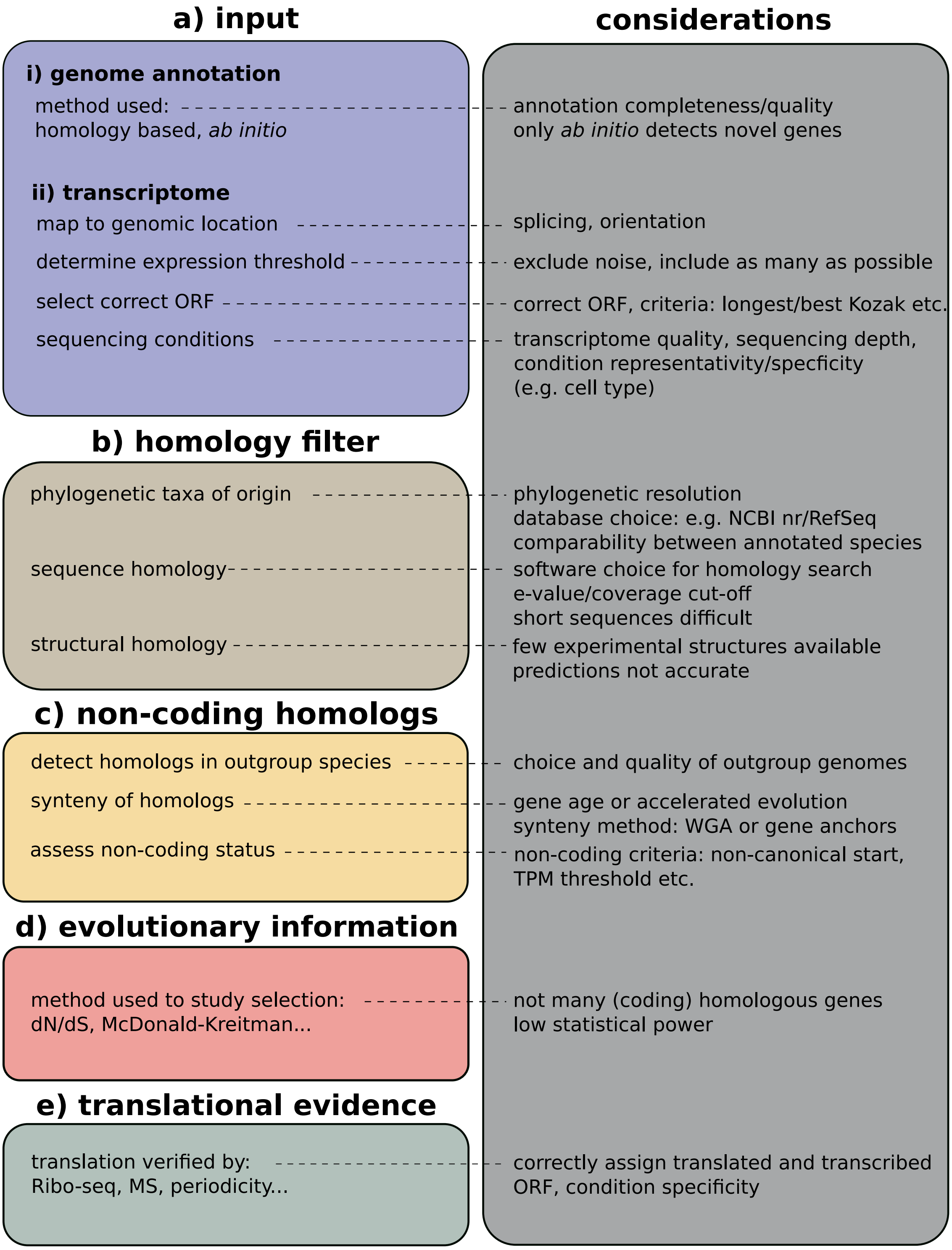
Another option for the detection of candidate *de novo* genes is to analyse transcripts from one or multiple transcriptomes. This approach involves more initial steps described below, but it likely allows for the detection of *de novo* genes in their early stages of emergence, such as proto-genes (Carvunis et al., 2012) or *de novo* open reading frames (ORFs) (Grandchamp et al., 2023b). The steps described in the following assume that the transcriptome has already been assembled based on a reference genome, using reference-based algorithms (Raghavan et al., 2022; Kovaka et al., 2019). If a transcriptome has been assembled *de novo*, the primary deviation from the described method resides in the identification of genomic locations of the ORFs. If the reference genome does not correspond to the assembled transcriptome or if no reference genome exists for the query species, the genomic location of the ORFs may lack precision.

Selection of transcripts based on genomic location

De novo genes can be located in various genomic regions, including intergenic spaces, introns, overlapping existing genes in a different frame or antisense orientation, within UTRs, or other non-genic location. Depending on the investigated *de novo* emergence mechanism(s), certain transcripts (or ORFs) may be excluded from the analysis. Utilising tools such as BEDtools (Quinlan and Hall, 2010) facilitates the determination of the genomic overlap of the transcripts, and the choice of which transcripts will be retained as candidates for further analyses. This step can also be conducted using ORFs instead of transcripts, after ORF detection in transcripts.

Detection of ORFs in a transcriptome.

After filtering transcripts based on their genomic location, the selected spliced transcripts are scanned for ORFs. Various software tools are available for extracting ORFs from a transcriptome, with one



a) input
considerations
i) genome annotation
method used:
homology based, *ab initio*
annotation completeness/quality
only *ab initio* detects novel genes
ii) transcriptome
map to genomic location
splicing, orientation
determine expression threshold
exclude noise, include as many as possible
select correct ORF
correct ORF, criteria: longest/best Kozak etc.
sequencing conditions
transcriptome quality, sequencing depth,
condition representativity/specficity
(e.g. cell type)
b) homology filter
phylogenetic taxa of origin
phylogenetic resolution
database choice: e.g. NCBI nr/RefSeq
comparability between annotated species
sequence homology
software choice for homology search
e-value/coverage cut-off
short sequences difficult
structural homology
few experimental structures available
predictions not accurate
c) non-coding homologs
detect homologs in outgroup species
choice and quality of outgroup genomes
synteny of homologs
gene age or accelerated evolution
synteny method: WGA or gene anchors
assess non-coding status
non-coding criteria: non-canonical start,
TPM threshold etc.
d) evolutionary information
method used to study selection:
dN/dS, McDonald-Kreitman...
not many (coding) homologous genes
low statistical power
e) translational evidence
translation verified by:
Ribo-seq, MS, periodicity...
correctly assign translated and transcribed
ORF, condition specificity

Figure 1: Considerations for general approaches and standards in *de novo* gene research. Related literature can be found in Table 1.

notable example being EMBOSS getorf (Rice et al., 2000). This tool conveniently provides information on the position of the ORF in the spliced transcript and its direction (forward or reverse). However, ORFs that extend to the end of a transcript without ending with a stop codon are also retrieved, which might be considered as erroneous and should be removed.

In order to extract the ORFs relevant to a given biological question, a number of steps must be followed:

1. If the RNA is stranded, detected antisense ORFs may be erroneous and should be regarded with caution.
2. Multiple transcripts may correspond to the spliced product of a single gene, and some might overlap (Lebherz et al., 2024). In such cases, removing duplicated ORFs shared among transcripts spliced from the same genomic location may be necessary.
3. The majority of transcripts contains multiple ORFs, and the choice of the ORF(s) within a transcript depends on the biological question, and various choices are valid (Xu et al., 2010).

Choice of Coding ORFs

When starting from a transcriptome with transcripts containing several ORFs, the selection of which ORFs to keep for further steps is decided by the investigator. Until recently, ORFs were typically considered potentially coding only if their size exceeded 300 nucleotides, a criterion implemented in algorithms such as those used by the Functional ANnoTation Of the Mammalian Genome (FANTOM) (Dinger et al., 2008; Leong et al., 2022). However, micropeptide and short *de novo* genes are known to have coding potential (Sandmann et al., 2023; Vakirlis et al., 2022; Patraquim et al., 2022), and *de novo* genes have been shown to be shorts (Toll-Riera et al., 2009; Palmieri et al., 2014; Guo et al., 2007). Various software tools have been developed to determine which ORF should be considered as the coding one in canonical genes, using approaches primarily based on protein homology (Varabyou et al., 2023; Vitting-Seerup et al., 2014; Kang et al., 2017). Nevertheless, even for canonical genes, the definition and number of coding ORFs are under revision, as the coding potential of genes has been shown to be significantly underestimated (Wright et al., 2022; Ardern, 2023).

In transcripts, all ORFs within a size limit can be considered. The majority of studies opt for the longest ORF (Xu et al., 2010; Dowling et al., 2020), which is also the default option for annotating protein-coding regions in most software (Rombel et al., 2002; Wang et al., 2013). Some studies only consider the first upstream ORF (uORFs) (Whiffin et al., 2020). Other studies consider the ORFs with the highest Kozak score (Kozak, 1989; Xu et al., 2010), indicating the highest likelihood of translation, or ORFs including surrounding untranslated regions (UTRs), since UTRs play crucial roles in translation initiation and transcript stability (Chatterjee and Pal, 2009; Matoulkova et al., 2012).

Importantly, the detection of ORFs with coding potential does not guarantee a translation event. Several studies have reported only a weak correlation between transcript expression levels and protein abundance (Koussounadis et al., 2015; Gry et al., 2009; Liu et al., 2016). This emphasizes that a transcribed ORF is strongly dependent on post-transcriptional and translational regulatory mechanisms for translation, which is difficult to predict without experimental evidence.

Selection of an expression threshold

Most studies include only the ORFs from transcripts that reach a minimum level of expression, which is typically determined by the transcripts per million (TPM) threshold. A threshold of 0.5 TPM has been adopted by numerous studies (Poretti et al., 2023; Vara et al., 2024; Petryszak et al., 2016) as specified by EMBL (Stoesser et al., 2002) as the minimal expression threshold. When assembling transcriptomes, low-expressed transcripts are often removed from the process as they are suspected to represent background noise (Janssen et al., 2023). However, emergence of low-expressed transcripts could be a step towards *de novo* gene emergence, and such transcripts might be important to study. The hypothesis that transcripts are produced throughout the entire genome of a species is referred to as pervasive transcription (Clark et al., 2011; Hangauer et al., 2013; Kellis et al., 2014). In cases involving splicing, it is crucial to be cautious when employing a TPM threshold. It is plausible for a gene to express multiple transcripts, where one transcript meets the specified threshold while the others do not.

Detection of genomic positions of unspliced transcripts and ORFs

In order to account for splicing events and the subsequent methodological steps, the genomic position of the selected ORFs must be detected. The software BLAT (Kent, 2002) is splicing-aware and can be used to map ORFs from a transcriptome to the corresponding genome. However, BLAT has difficulties dealing with short sequences, as *de novo* ORFs often are. Instead of aligning intact ORFs, BLAT overpredicts splicing events by splitting up ORFs to align them to multiple locations in the genome. The most precise method for retrieving the genomic location of an ORF is to extract the coordinates from the transcript it originates from. This accurate approach is only feasible if the transcriptome is assembled using reference-based algorithms. To our knowledge, such a step cannot be fulfilled by existing software and requires custom scripts.

After all these steps, all filtered ORFs and/or transcripts can be considered as candidate *de novo* genes and will be used for the next filtering steps.

2.1.3 Validation of Translation

To assess whether the selected candidate genes are coding genes, one option is to use experimental validation (Figure 1 e, Table 1). Experimental validation of a gene's coding status can be performed at the very end of the methodology, when only a subset of genes has been validated as *de novo* genes. However, when starting from a transcriptome, validating translation can be the very first step of the method. In such cases, all translated ORFs detected experimentally are mapped to the corresponding transcriptome (Wacholder et al., 2023) and subsequently sorted through several steps similar to those used in transcriptome analysis (Turcan et al., 2024).

To confirm the coding status of putative *de novo* genes, several new laboratory techniques have proven to be highly effective, particularly for small proteins. Ribosome profiling-based approaches (Ribo-Seq) (Ingolia et al., 2009; Kondo et al., 2010; Ingolia et al., 2011; Bazzini et al., 2014; Chen et al., 2020; Duffy et al., 2022) and mass spectrometry-based approaches (Ji et al., 2015; Slavoff et al., 2013; Pauli et al., 2014) assess the binding of ribosomes to transcribed ORFs or the presence of translated proteins. These two approaches can also be combined for better accuracy (Schlesinger and Elsässer, 2022; Wacholder and Carvunis, 2023; Andjus et al., 2024).

2.1.4 Genomes or Transcriptomes?

The choice between candidate *de novo* genes from annotated genomes or transcriptomes depends on the biological question being investigated. Candidate genes from an annotated genome provide a high level of confidence about the genic status of the identified *de novo* genes at the end of the pipeline. Evolutionary fixation in a species is more likely for these genes, as their genic structures are apparently stable enough to be recognised by annotation methods. Nevertheless, *de novo* genes that are lacking gene homology or genic structures, such as introns or specific transcription motifs, may not be detected by annotation tools.

Selecting candidate genes from a transcriptome generally results in the identification of a considerably higher number of *de novo* genes compared to candidate genes from an annotated genome. For example, in Roginski et al. (2024), the authors detected 89 *de novo* genes in humans when starting from a genome, while Dowling et al. (2020) identified 2,749 human-specific *de novo* expressed ORFs when starting from a transcriptome. Similarly, Roginski et al. (2024) detected 92 *de novo* genes in *Drosophila melanogaster* by analyzing an annotated genome, while Zheng and Zhao (2022) identified 993 *de novo* genes in the same species using Ribo-seq data mapped to a transcriptome. However, depending on the specific transcriptome and the applied criteria, it is possible that the majority of the detected translated ORFs may not be fixed in the species (Roginski et al., 2024).

The genic status of *de novo* candidates can be confirmed through the validation of translation as described above and subsequently only considering the translated ORFs. When starting from a transcriptome, one important issue can come from the fact that transcript expression is complicated to characterise, as expression can depend on conditions, tissues, sex, life stage, individuals or populations, among others (Nieuwenhuis et al., 2021; Xu et al., 2023; Schneider et al., 2024; Oliva et al., 2020). Consequently, particular *de novo* genes can be specific to certain conditions or tissues (Figure 1 a, Table 1). The detection of such genes can be more challenging, particularly when their expression levels are low.

2.2 Taxonomic Group of Emergence

A *de novo* gene or expressed ORF may be specific to an individual, a population, a species, or a broader taxonomic group. When starting from a transcriptome, it may also be expressed only under specific conditions, such as in a specific tissue, age or sex. The taxonomic level of emergence can but does not have to be specified in advance, ensuring that only *de novo* genes meeting a particular condition are retained. If a gene is not species, population or condition specific, it is called a taxonomically restricted gene, and belongs to a taxonomic group of closely related species. The distinction between *de novo* genes and other genes becomes more challenging when they are shared by several rather than one single species, particularly if they have an evolutionary origin predating a loss of synteny within the taxa to which they belong, and if they exhibit a high mutation rate, although this is likely not frequent (Domazet-Lošo et al., 2017). The more distantly related the species in the taxonomic group are, the more information is lost about *de novo* gene emergence or their mechanism of emergence in general. *De novo* gene birth is easier to identify in taxonomic groups including species that diverged recently, provided that the considered evolutionary time is sufficient to characterize the genicity of the sequences. A large number of studies focuses on species-specific *de novo* genes (Broeils et al., 2023; Zhao et al., 2014; Schmitz et al., 2018; Grandchamp et al., 2023b,a; Lebherz et al., 2024; Vara et al., 2024; Zhang et al., 2019). Alternatively, there is the possibility of detecting the earliest stage of a gene emergence by studying the emergence of a *de novo* transcribed ORF in individuals or populations. In such a case, the search for homology is conducted against outgroup species, but also against outgroup populations/individuals from the same species, if such data is available (Grandchamp et al., 2023b).

2.3 Homology Filter

The main criterion for identifying a recent *de novo* gene is the lack of homology to any other coding genes outside and inside of the expected phylogenetic group/species/population of emergence. The homology search has to be performed for the full dataset of candidate genes from the previous steps. All of them that show significant homology can then be discarded from the list of potential *de novo* genes.

Each *de novo* gene is required to show no similarity to any gene outside or within the species or taxonomic group of interest, which would suggest that the candidate gene emerged via a recycling mechanism, such as duplication. The inclusion of a greater number of outgroup species in the analysis leads to more robust results.

Protein sequences as the default option

The most widely employed method for identifying homologs is to use protein sequence similarity for the purpose of database searches. Such searches may encompass proteins from a broad range of species. Distant outgroup species should be also included to rule out horizontal gene transfer and distant homologies. Large databases containing sequence data from all domains of life, such as the NCBI Reference Sequence Database (Pruitt et al., 2005) can be searched to include as many species and taxonomic groups as possible. Newly assembled genomes and corresponding proteomes that have not been incorporated into public databases can also be beneficial to search when studying a specific taxon (Figure 1 b).

With transcriptome-based analysis, it is often assumed that *de novo* candidates are not annotated in the reference genome. Consequently, annotation software might fail to identify homologous genes in outgroup genomes, leading to incomplete outgroup proteomes. In such cases, validation may rely on the subsequent identification of syntenic homologs that lack coding properties (ex ORFs) or show important frameshift, to confirm the absence of possible homologous encoded protein. Alternatively, Vakirlis and McLysaght (2019) propose performing similarity searches of six-frame translations of entire outgroup genomes. This method discards any putative coding homologs in outgroup genomes, including bona fide non-coding homologs that lacks stop, frameshift and transcription. While this approach is likely to be the most effective, it is more suitable for small genomes, as it can be computationally intensive for larger genomes. The homology search is typically conducted using the protein sequence of the genes to be tested. However, there has been an increasing trend in the use of protein structure, in addition to the sequence, depending on the specific biological question being investigated (Middendorf et al., 2024; Van Kempen et al., 2024; Alvarez-Carreño et al., 2021).

Using the DNA sequence to include ncRNAs

A homology search can also be performed based on the DNA sequence of candidate *de novo* genes. This can be useful when looking for homology in non-coding RNA (ncRNAs). In such instances, the direction of the alignment should be considered, as well as the coverage, given that two overlapping transcripts could have originated from distinct promotors (Grandchamp et al., 2023a). Furthermore, according to the biological question, it can be wanted that a *de novo* gene is not derived from a transposable element (TE), or from an annotated and conserved ncRNA. To address this, the untranslated ORF or transcript can be searched for homology against a database, comprising TEs and ncRNAs from query and outgroup species. An important caveat is that, if proteogenomic evidence of translation exists for a given genomic sequence (Slavoff et al., 2013; Chen et al., 2020; Duffy et al., 2022; Mudge et al., 2022) then such direct evidence overrules the similarity with a long non-coding RNA (lncRNA), and may in fact indicate that the lncRNA is in fact coding (Prensner et al., 2021). Importantly, the use of DNA sequences can be problematic for *de novo* genes that emerged through specific mechanisms such as overprinting or antisense emergence. More precisely, such candidates might exhibit significant DNA similarity with genes they overlap with, leading to their erroneous exclusion from a list of potential *de novo* genes.

Available tools for sequence similarity searches

Several tools are available to search for homologous sequences. BLAST (Altschul et al., 1990) is commonly used for homology searches and is recommended because of its speed and accuracy. When working with a large database such as the NCBI nr or RefSeq, a faster tool for local alignments than BLAST, such as Diamond (Buchfink et al., 2021), can be used. As *de novo* genes that show homology to existing proteins should be removed from the dataset of potential *de novo* genes, the choice of homology criteria is important. Different E-value thresholds can be used to assess homology (Vakirlis et al., 2020), even though an e-value of 10e-2 should be the highest tolerated. For example, one might want to be extremely restrictive while studying one single *de novo* gene involved in a specific function to ensure that it contains no other gene overlap. A more relaxed threshold can be applied if the phylogenetic group includes a lot of species and the homology search is performed against very distant species. An additional measure is the alignment coverage (Long and Langley, 1993; McLysaght and Hurst, 2016) (Figure 1b, Table 1).

Predicting protein structures for homology searches

Recent advancements in protein structure prediction, most importantly by AlphaFold2 (Jumper et al., 2021), have led to new opportunities for phylogenetic analyses based on protein structures (Moi et al., 2023). Protein structures exhibit greater conservation compared to their sequences (Illergård et al., 2009), suggesting the potential of putative *de novo* genes actually representing highly divergent orthologs (Casola, 2018). To further confirm a *de novo* origin, structural similarity searches can be conducted using tools such as Foldseek (Van Kempen et al., 2024). Foldseek enables rapid comparison of structural similarities across a broad range of databases, encompassing both experimental and computationally derived structures. However, the commonly used AlphaFold2 (Jumper et al., 2021) primarily relies on co-evolutionary data derived from multiple sequence alignments (MSAs), which are inherently sparse for *de novo* proteins, impacting the reliability of predictions (Figure 1 b, Table 1) (Jumper et al., 2021; Aubel et al., 2023; Liu et al., 2023). Given this limitation, there has been growing interest in structure predictors that utilize protein language models. These models are supposedly more suitable for predicting the structures of *de novo* proteins and other orphan proteins, where sequence homologies are limited or non-existent (Aubel et al., 2023; Liu et al., 2023; Michaud et al., 2022; Lin et al., 2023; Chowdhury et al., 2022; Middendorf and Eicholt, 2024). However, it is important to note that both AlphaFold2 and protein language model-based tools, such as ESMfold, have been shown to inaccurately predict structures of *de novo* proteins, and with discordant confidence scores (Middendorf and Eicholt, 2024; Aubel et al., 2023). The most recent implementation of AlphaFold - AlphaFold3 (Abramson et al., 2024) - has yet to be tested for its performance on orphan proteins and *de novo* emerged proteins. Recent studies have successfully utilized molecular dynamics (MD) simulations as refinement to explore the structural dynamics of*de novo* proteins (Lange et al., 2021; Peng and Zhao, 2024; Middendorf et al., 2024).

After the homology filtering step, the list of candidate genes is reduced to a list of potential *de novo* genes, containing only genes that don't have detected homologs outside the studied taxonomic group.

section	**considerations and literature**
genome annotation method	**completeness/quality** (Casola, 2018; Vakirlis and McLysaght, 2019; Weisman et al., 2022), ***ab initio* for novel genes** (Scalzitti et al., 2020; Baker et al., 2023; **?**; **?**)
map transcriptome to genome	**splicing, orientation** (Iyengar et al., 2024)
determine expression threshold	**exclude noise but not low expression, consider different thresholds** (Grandchamp et al., 2023a; Heames et al., 2020; Blevins et al., 2021; Lombardo et al., 2023)
select correct ORF	**criteria: length/Kozak/...** (Schmitz et al., 2018; Dowling et al., 2020; Heames et al., 2020; Blevins et al., 2021; Iyengar and Bornberg-Bauer, 2023; Xu et al., 2010; Whiffin et al., 2020)
sequencing conditions	**transcriptome quality, sequencing depth, condition specificity (e.g. cell type)** (Blevins et al., 2021; Toll-Riera et al., 2009; Schlötterer, 2015)
phylogenetic taxa of origin	**phylogenetic resolution** (Li et al., 2021), **database choice** (Vakirlis and McLysaght, 2019; Weisman et al., 2020b; Moyers and Zhang, 2015)
sequence homology	**software choice** (Altschul et al., 1990; Buchfink et al., 2021; Finn et al., 2011), **sequence similarity cutoff, especially short sequences difficult** (Moyers and Zhang, 2016, 2017, 2015, 2018; Weisman et al., 2020b; Vakirlis et al., 2020; Domazet-Loso and Tautz, 2003)
structural homology	**few experimental structures, predictions not accurate** (Aubel et al., 2023; Middendorf and Eicholt, 2024)
detect homologs in target species	**choice/quality of target genomes** (Moyers and Zhang, 2015; Vakirlis and McLysaght, 2019; Weisman et al., 2020b)
synteny of homologs	**gene age or accelerated evolution** (Vakirlis et al., 2020; Casola, 2018; Weisman et al., 2020b; Ranz et al., 2001; Zdobnov et al., 2002), **synteny method** (Casola, 2018; Roginski et al., 2024; Vakirlis et al., 2020) **WGA** (Peng and Zhao, 2024), **phylostratigraphy** (Moyers and Zhang, 2016, 2017, 2015, 2018; Prabh and Rödelsperger, 2019; Zdobnov et al., 2002; Ranz et al., 2001)
assess non-coding status	**criteria: non-canonical start, TPM threshold etc.** (Roginski et al., 2024; Vakirlis et al., 2024)
method used to study selection	**limited number of (coding) homologs** (Schlötterer, 2015; Rivard et al., 2021; Broeils et al., 2023; Gubala et al., 2017; Zhao et al., 2014; Chen et al., 2015)
translation verified	**correctly assign ORF, condition specificity** (Vakirlis et al., 2018; Zhang et al., 2019; Wilson and Masel, 2011; Papadopoulos et al., 2024; Ruiz-Orera et al., 2014; Ruiz-Orera and Albà, 2019; Papadopoulos et al., 2021; Patraquim et al., 2020, 2022)

Table 1: Considerations and related literature for general approaches and standards in *de novo* gene research.

2.4 Non-Coding Homologs

The detection of syntenic non-coding sequences, homologous to all potential *de novo* genes under investigation, in target species or populations that are outgroup to the ones expressing the potential *de novo* genes, is for now the last step to provide evidence for a *de novo* emergence. In this review, we define a "non-coding homolog" as a homologous sequence that supports the validation of a *de novo* gene emergence. However, determining whether a genomic sequence is truly non-coding can be challenging. As a result, several studies define non-coding homologs as sequences lacking an open reading frame (ORF) that could encode a protein homologous to the one produced by the *de novo* gene (Vakirlis and McLysaght, 2019; Wacholder et al., 2023; Sandmann et al., 2023). In such cases, an insertion in the homologous sequence would not necessarily prevent translation, but result in a different frame and with that loss of protein homology.

However, identification of syntenic regions and a coding status can be challenging, and the absence of a "syntenic non-coding homolog" does not necessarily invalidate a *de novo* origin.

The *de novo* origin of a potential *de novo* gene can be suspected under the following conditions:

- homologous sequences to the *de novo* gene can be detected in genome of several target species or populations. Such target species or populations must be outgroup to the phylogenetic group, species or population where the *de novo* genes under investigation are present.
- the identified homologous sequences are non-coding, or would encode a protein sufficiently different from the one encoded by the candidate, for example due to a frameshift early in the sequence.
- the identified homologous sequences are in a genomic location that is syntenic to the *de novo* gene

The following steps are required to detect syntenic non-coding homologs:

2.4.1 Selection of target genomes for synteny search

In order to identify syntenic non-coding homologs, a set of target genomes must be selected. This set of target genomes will be used to validate or invalidate a *de novo* emergence for all remaining genes from the previously filtered set. For instance, in the case of studying *de novo* genes first steps of emergence within a species, the target genomes should be those from individuals or populations of the same species that do not contain the *de novo* gene(s) of interest. Conversely, when searching for *de novo* genes specific to a taxonomic group that includes several species, the target genomes should be closely related to that taxonomic group, but have diverged earlier than the root of this group. The optimal number of target genomes required for the identification of non-coding homologs remains undetermined; however, it is generally accepted that the greater the number of genomes analysed, the more robust the conclusions drawn (Figure 1 c Table 1).

2.4.2 Homology search between the query *de novo* gene and the target genomes

Once the target species have been identified, genomic sequences homologous to the potential *de novo* gene can be searched for. During this step, the homology search is performed against the genome of all target species. One option is to use tBLASTn, by using the *de novo* translated ORF as a query (Vakirlis and McLysaght, 2019). However, the most precise option to detect homologous sequences independently of their frame of translation is to use BLASTn. If the ORF is small, and if the unspliced gene contains one or several introns, an option is to use the unspliced ORF as a query for a nucleotide BLAST against the target genome, and then splice the resulting alignment (Grandchamp et al., 2023b). If the target genome belongs to a species that is phylogenetically distant from the query species, alignment programs that allow more divergence such as exonerate (Slater and Birney, 2005) can also be used to search for homology.

2.4.3 Search for syntenic regions

Genomic synteny refers to the conservation of genomic fragments within two genomes or chromosomes. If one or several homologous hits have been detected for a single query *de novo* gene, some of these hits can be further validated in each target species by confirming their location in a genomic region that is syntenic to the *de novo* gene. This step can also be performed in reverse with the previous one, meaning that the search of homologous sequences could also be performed only in syntenic regions.

Methods for synteny detection

There are numerous methods available for synteny detection. Synteny can be compared between two complete genomes by fragmenting each chromosome into blocks based on sequence fragments, motifs, domains, etc., and determining similarity and location between blocks (Wang et al., 2012;

Liu et al., 2018). Synteny can also be examined at a genic level by studying the conservation of the order of syntenic genes between genomes. In such cases, genes are selected as anchors to determine synteny, and the detection of synteny is based on gene orthology. For instance, SynChro (Drillon et al., 2014) and Synima (Farrer, 2017) are software tools that detect synteny using reciprocal BLAST hits between genes from different genomes. Using genes as anchors for synteny is a rapid and effective approach when searching for syntenic hits of *de novo* genes that are intergenic (Vakirlis et al., 2020; Roginski et al., 2024). The genes neighboring the *de novo* gene are chosen as anchors and investigated for orthology in the target genome. If the non-coding homolog is flanked by genes orthologous to those surrounding the query *de novo* gene, the synteny is confirmed. The number of anchor genes can be adjusted based on the context. When working within populations or individuals of a single species or closely related species, a stringent requirement for complete synteny may be imposed. In such cases, non-coding sequences homologous to the candidate *de novo* gene are collected only if they are positioned between two genes homologous to those surrounding the query candidate. Other approaches also exist for synteny detection.Käther et al. (2023) introduced an approach called "Annotation-Free Identification of Potential Synteny Anchors" that does not rely on genes as anchors. Zhao and Schranz (2017) suggested using network approaches to infer synteny. One of the best ways to validate synteny is to use whole-genome alignments. In such cases, the genomic region of target genomes that aligns to the *de novo* candidate from the query genome corresponds to the syntenic homolog. For instance, Wacholder et al. (2023) aligned syntenic conserved blocks to precisely locate the coordinates of non-coding homologs compared to candidate *de novo* genes in yeasts. Similarly, Sandmann et al. (2023) used a whole-genome alignment of 120 mammalian species and another alignment of 27 primate species to search for non-coding sequences homologous to human-translated micropeptides. Whole-genome alignments have also been used to identify *de novo* genes in *Drosophila* (Peng and Zhao, 2024), though some appear to have been overlooked (Guay et al., 2025). Overall, whole-genome alignments are highly reliable but require several, high-quality genomes, which are often not available.

Caveats when using synteny

While validating synteny between *de novo* candidates and homologous sequences is necessary, this steps also is affected by methodological limitations. The definition and conservation of synteny depends on several criteria, such as the quality of genome annotation, alignments, and the selection of syntenic anchors, windows, and algorithms. Liu et al. (2018) demonstrated that synteny between species can be underestimated by up to 40% depending on the methodology chosen. Moreover, once a syntenic block is detected between a query and a target genome, the identification of a non-coding homolog also depends on the methodology. Therefore, the methodology used to detect and define synteny can vary from one project to another, leading to variable conclusions. Independently of the method used, the phylogenetic distance between the query genomes and selected target species influences synteny conservation: the greater the distance between genomes, the less conserved the synteny (Lemoine et al., 2007). For instance, macrosynteny tends to be preserved for approximately 10-100 million years, whereas microsynteny can remain conserved over several hundred million years. For example, many genes are syntenic within Chordates and Arthropods, each of which emerged around 560 million years ago (mya), but not between the two phyla (Vonica et al., 2020), which diverged approximately 708 mya (Kumar et al., 2022). Furthermore, synteny conservation can vary among taxa (e.g., plants, animals) even for similar phylogenetic distances (Roginski et al., 2024). Moreover the detection of syntenic non-coding sequences homologous to *de novo* genes often fails due to factors such as extensive genomic rearrangements. When validation of *de novo* emergence through the detection of a non-coding homolog cannot be achieved, drawing conclusions about *de novo* emergence becomes challenging. Some genes that emerge after a duplication event have been observed to evolve rapidly, diverging from their original sequence to an extent that no homology tool can reliably predict their origin (Casola, 2018; Naseeb et al., 2017; Gu et al., 2005; O'Toole et al., 2018; Pegueroles et al., 2013). Consequently, such genes may exhibit no homology to any other annotated gene and could be mistakenly identified as *de novo* genes, in the absence of non-coding homolog (Weisman et al., 2020a).

2.4.4 Assess the coding status of the detected homologous sequences

Once a syntenic homolog of a potential *de novo* gene has been detected, the final step is to determine its coding status. To do so, the query sequence and its homolog are often re-aligned before deeper investigation (Peng et al., 2024; Wacholder et al., 2023; Sandmann et al., 2023). If one homolog shares the same coding properties as the potential *de novo* gene, then such gene did not emerge *de novo*, or at least not prior to the divergence of the two studied species (query and target). On the other hand, if all homologous sequences are non-coding, then the *de novo* origin of the *de novo* candidate under investigation is assumed as the "most likely" in the query species.

Assessing the coding/non-coding status of detected homologs remains the most challenging step of the entire pipeline. Several properties can be assessed to compare the coding status of the sequence homologous to the potential *de novo* gene, such as the presence of start and stop codons, premature stop codons, frameshift mutations, and splice sites in the case of introns (Grandchamp et al., 2023b). However, the question remains: are these features, or their absence, sufficient to validate or invalidate a coding gene status? For example, the absence of an ATG start codon in a non-coding homolog to a *de novo* candidate does not necessarily prevent translation, as several weaker start codons have been shown to be adequate for translation (Cao and Slavoff, 2020), with some being conserved across evolution (Bazykin and Kochetov, 2011). More precisely, several small peptides have been shown to be often encoded by sORFs with non-AUG start codons (Peng et al., 2024). Wacholder et al. (2023) emphasise frameshift mutations as crucial features to consider, since the position of a frameshift in a putative non-coding homolog can significantly affect the divergence from the *de novo* candidate if both are translated. In Sandmann et al. (2023), authors translated the homologous ORF, if any, and calculated a score of protein homology.

Evaluating transcription of the non-coding homolog also improves the determination of a genic status. Transcription information is also useful for inferring the emergence of splice sites. Several studies have reported the presence of introns in *de novo* genes (Zhang et al., 2019; Wu et al., 2011; Grandchamp et al., 2022). Studying the emergence of these introns and the evolution/conservation of their splice sites would be essential, as the loss or gain of splicing could significantly alter the translated protein. To the best of our knowledge, such a study has not yet been conducted.

This last step must be conducted with caution, as it can lead to significant misinterpretations. Robust conclusions can only be acquired if several strategic target genomes are selected—the more, the better. The transition from a non-coding sequence to a protein-coding gene follows various steps (Ruiz-Orera et al., 2017; McLysaght and Guerzoni, 2015). All mutations and transitions can occur in different orders (Carvunis et al., 2012; Lebherz et al., 2024; Iyengar et al., 2024). More importantly, the process of acquiring a coding status can go back and forth during evolution, as the initial stages of *de novo* emergence are *a priori* not subject to selection pressures (Carvunis et al., 2012; Iyengar and Bornberg-Bauer, 2023). Therefore, the detection of non-coding sequences homologous to a candidate *de novo* gene, can only be valuable if such a non-coding status is confirmed in several target, as a coding homolog could hypothetically also be detected in more divergent species that were not studied (Figure 1 c Table 1).

After all these steps, among the set of potential *de novo* genes under investigation, the ones that have non-coding syntenic homologs in all target genomes can be validated as *de novo* genes.

2.4.5 Evolutionary Information

What selective pressures apply on a *de novo* gene? According to the model proposed in 2012 (Carvunis et al., 2012), the emergence of a new gene from a non-coding sequence involves two main steps: the first is the emergence of a proto-gene, which is a transcribed and translated ORF whose genomic sequence is not yet under selection, producing a small peptide that is likely gained and lost through evolution. The second stage is when a proto-gene becomes fixed in a species due to selection, achieving the status of a *de novo* gene (Van Oss and Carvunis, 2019). It is challenging to determine whether a *de novo* gene is fixed in a species, and by that gaining a *de novo* gene status, or whether it is not yet fixed, classifying the gene as a proto-gene. Measurements of selection pressures can be used (Feldmeyer et al., 2024) to distinguish between these two. Moreover, the method used to detect *de novo* genes influences of which type the majority of candidate genes are.

De novo genes extracted from an annotated genome are likely to become fixed or are fixed already, as their coding features are robust enough to be detected by standard annotation methods. Several studies have demonstrated that *de novo* genes extracted from annotated genomes are under purifying selection both within and between species (Li et al., 2010; Palmieri et al., 2014). Moreover, specific codons have been shown to be enriched in such *de novo* genes (Wallace et al., 2013; Hershberg and Petrov, 2008; Schlötterer, 2015).

Assessing *de novo* genes extracted from transcriptomes and/or proteomes is more challenging. Labeling such sequences as *de novo* genes should be supported by evidence of purifying selection, conservation within populations of a species and translational evidence. If no selection tests are performed, the term proto-gene is most commonly used. The term ORFans (Vakirlis and McLysaght, 2019) or newly expressed ORFs (Grandchamp et al., 2023b) is used for ORFs that were extracted from transcriptomes without evidence of translation. Newly translated ORFs is the commonly used term for ORFs with evidence of translation whose level of transcription is unknown. However, the validation of a *de novo* status does not have to be supported by all these conditions. For instance, in the case of genes annotated by *ab initio* methods, evidence of transcription is generally not provided, unless additional laboratory experiments are conducted. Moreover, *ab initio* and homology-based methods do not provide evidence of selection for the identified genes (Kryazhimskiy and Plotkin, 2008; Burge and Karlin, 1997). Conversely, if an unannotated ORF exhibits direct evidence of both transcription and translation, there is no conceptually valid reason to apply more restrictive criteria than for canonical genes.

Unfortunately, assessing evidence of selection in *de novo* genes remains extremely challenging (Figure

1 d Table 1). Selection pressure is often assessed using metrics such as the *dN/dS* ratio (Hurst, 2002; Yang and Bielawski, 2000; Kosakovsky Pond and Frost, 2005) or the *pN/pS* ratio (McDonald and Kreitman, 1991). However, both of these metrics are designed for coding sequences. Therefore, the presence of non-coding homologs or non-coding variants of a *de novo* emerged ORF poses problems for their calculation. While these difficulties do not prevent the study of selection among all coding samples of a *de novo* emerged ORF, a future challenge would be to incorporate non-coding sequences into a calculation of selective pressure, to gain a clearer understanding of selection dynamics in the earliest stages of emergence.

Lastly, most *de novo* ORFs are shorter than canonical ORFs and are present in a limited number of species or populations, which limits the statistical power to confidently detect selection (Wacholder et al., 2023). Several studies have addressed the challenge of assessing selection on *de novo* emerged ORFs. For example, Ward and Kellis (2012) attempted to understand whether the large portion of the human genome that is biochemically active shows evidence of purifying selection. By using genome alignments and studying sequence conservation, they found that 4% of the human genome is subject to lineage-specific constraint, in addition to the 5% already known. In 2003, Kellis et al. (2003) developed a reading frame conservation (RFC) test to classify all ORFs of *S. cerevisiae* as either biologically meaningful or meaningless. This RFC test was later adapted by Wacholder et al. (2023) to distinguish ORFs evolving under selection from other ORFs in the yeast genome particularly those showing weak signals in more classical selection tests. While they found no evidence of purifying selection acting on most of these *de novo* emerged ORFs, a few samples showed selection.

2.5 Available Software

The identification of *de novo* genes is contingent on numerous methodological decisions, with custom scripts or programs frequently required for multiple steps in the process. Fortunately, recent advancements have led to the publication of various tools and software that automate *de novo* gene detection, either completely or partially. Singh and Wurtele (2021) developed *orfipy*, which facilitates the detection of ORFs in new transcriptomes that can be used subsequently to search for *de novo* genes in transcriptomic data. The R package *phylostratr* (Arendsee et al., 2019b) allows to infer a phylostratum for all input query genes, thereby enabling the identification of homology to a candidate gene. GenEra (Barrera-Redondo et al., 2023) allows to detect taxonomically restricted genes. The softwares *fagin* partially automate (Arendsee et al., 2019a) and DENSE (Roginski et al., 2024) automate the detection of *de novo* genes in an annotated genome. An automated tool for detection of *de novo* genes based on transcriptomic data is unfortunately not yet available.

2.6 Challenges & Conclusions

In conclusion, despite significant advances in understanding *de novo* gene emergence, two major challenges remain. Firstly, current methods for detecting *de novo* genes are largely limited to evolutionary young genes, making it difficult to discern the origins of ancient genes within large and complex gene families. This limitation stems from the fact that existing approaches can only trace the recent origin of a gene, which becomes increasingly challenging as the gene ages and undergoes multiple rounds of duplication and divergence of sequence and function. As a result, our current understanding of *de novo* gene emergence is biased towards recently evolved genes, leaving a significant gap in our knowledge of how older *de novo* genes originated. Novel approaches for remote homology detection and improved structure predictions could help us address this bias in the future.

Secondly, the lack of standardisation in methodology and terminology hinders comparability between studies, with different approaches and thresholds yielding disparate results even when analysing the same species. We address this problem directly in our accompanying paper by providing a standardised annotation format based on the identified classifications described in this review. Such a standardised annotation format represents a crucial step towards achieving a common framework, enabling researchers to compare and build upon each other's work more effectively.

By establishing a common framework for describing, analysing and comparing *de novo* gene studies, we can enhance reproducibility, comparability, and ultimately, drive progress in this rapidly evolving field. Albeit the remaining challenges in this young field, our work paves the way for future studies to refine methods and integrate *de novo* gene searches into standard gene annotation pipelines, unlocking new biological insights into the origins of genes.

Competing interests

No competing interest is declared.

Author contributions statement

AG and ED were responsible for the conceptualisation of the review and handled project administration. AG wrote the first draft of the review. MA, LE and ED restructured the text and implemented

sections. MA edited the figures. PR, VL and AK provided feedback and modifications on the text. The final manuscript was edited and reviewed by all authors. Erich Bornberg-Bauer provided general administrative support and acquisition of the financial support for the project leading to this publication.

Acknowledgments

We thank Erich Bornberg-Bauer (EBB), Nikolaos Vakirlis, Li Zhao, Anne Lopes, Bharat Ravi Iyengar and Andreas Lange for their useful feedback during the planning phase of the project. ED was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 503348080. This grant with the additional grant number BO2544-22-1 was awarded to EBB. MA received funding from the Volkswagen foundation with grant code 98183 awarded to EBB. AG's work was supported by the Deutsche Forschungsgemeinschaft priority program "Genomic Basis of Evolutionary Innovations" (SPP 2349) BO 2544/20-1 to EBB, and by the (HFSP) Postdoctoral Fellowship (Grant No. 2023-981550) awarded to EBB, Anne-Ruxandra Carvunis and Christine Brun. LAE has been supported by EMBO Scientific Exchange Grant 10944. VL was supported by NIH grant R01NS095654 (to Nenad Sestan).

References

J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

C. Alvarez-Carreño, P. I. Penev, A. S. Petrov, and L. D. Williams. Fold Evolution before LUCA: Common Ancestry of SH3 Domains and OB Domains. *Molecular Biology and Evolution*, 38(11):5134–5143, Nov. 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab240. URL https://doi.org/10.1093/molbev/msab240.

S. Andjus, U. Szachnowski, N. Vogt, S. Gioftsidi, I. Hatin, D. Cornu, C. Papadopoulos, A. Lopes, O. Namy, M. Wery, et al. Pervasive translation of xrn1-sensitive unstable long noncoding rnas in yeast. *RNA*, 30(6): 662–679, 2024.

Z. Ardern. Alternative reading frames are an underappreciated source of protein sequence novelty. *Journal of Molecular Evolution*, 91(5):570–580, 2023.

Z. Ardern, K. Neuhaus, and S. Scherer. Are antisense proteins in prokaryotes functional? *Frontiers in molecular biosciences*, 7:187, 2020.

Z. Arendsee, J. Li, U. Singh, P. Bhandary, A. Seetharam, and E. S. Wurtele. Fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC bioinformatics*, 20(1):1–14, 2019a.

Z. Arendsee, J. Li, U. Singh, A. Seetharam, K. Dorman, and E. S. Wurtele. phylostratr: A framework for phylostratigraphy. *Bioinformatics*, 35(19):3617–3627, 2019b.

M. Aubel, L. Eicholt, and E. Bornberg-Bauer. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research*, 12(347):347, 2023.

L. Baker, C. David, and D. J. Jacobs. Ab initio gene prediction for protein-coding regions. *Bioinformatics Advances*, 3(1):vbad105, 2023.

D. Baltimore. Viral rna-dependent dna polymerase: Rna-dependent dna polymerase in virions of rna tumour viruses. *Nature*, 226(5252):1209–1211, 1970.

J. Barrera-Redondo, J. S. Lotharukpong, H.-G. Drost, and S. M. Coelho. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using genera. *Genome Biology*, 24 (1):54, 2023.

G. A. Bazykin and A. V. Kochetov. Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic acids research*, 39(2):567–577, 2011.

A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, et al. Identification of small orf s in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, 33(9):981–993, 2014.

D. J. Begun, H. A. Lindfors, M. E. Thompson, and A. K. Holloway. Recently evolved genes identified from drosophila yakuba and d. erecta accessory gland expressed sequence tags. *Genetics*, 172(3):1675–1681, 2006.

D. J. Begun, H. A. Lindfors, A. D. Kern, and C. D. Jones. Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics*, 176(2):1131–1137, 2007.

W. R. Blevins, J. Ruiz-Orera, X. Messeguer, B. Blasco-Moreno, J. L. Villanueva-Cañas, L. Espinar, J. Díez, L. B. Carey, and M. M. Albà. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature communications*, 12(1):604, 2021.

L. A. Broeils, J. Ruiz-Orera, B. Snel, N. Hubner, and S. van Heesch. Evolution and implications of de novo genes in humans. *Nature ecology & evolution*, pages 1–12, 2023.

B. Buchfink, K. Reuter, and H.-G. Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18(4):366–368, 2021.

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic dna. *Journal of molecular biology*, 268(1):78–94, 1997.

J. Cai, R. Zhao, H. Jiang, and W. Wang. De novo origination of a new protein-coding gene in saccharomyces cerevisiae. *Genetics*, 179(1):487–496, 2008.

X. Cao and S. A. Slavoff. Non-aug start codons: expanding and regulating the small and alternative orfeome. *Experimental cell research*, 391(1):111973, 2020.

A.-R. Carvunis, T. Rolland, I. Wapinski, M. A. Calderwood, M. A. Yildirim, N. Simonis, B. Charloteaux, C. A. Hidalgo, J. Barbette, B. Santhanam, et al. Proto-genes and de novo gene birth. *Nature*, 487(7407):370–374, 2012.

C. Casola. From de novo to "de nono": the majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. *Genome Biology and Evolution*, 10(11):2906–2918, 2018.

S. Chatterjee and J. K. Pal. Role of 5-and 3-untranslated regions of mrnas in human diseases. *Biology of the Cell*, 101(5):251–262, 2009.

J. Chen, A.-D. Brunner, J. Z. Cogan, J. K. Nuñez, A. P. Fields, B. Adamson, D. N. Itzhak, J. Y. Li, M. Mann, M. D. Leonetti, et al. Pervasive functional translation of noncanonical human open reading frames. *Science*, 367(6482):1140–1146, 2020.

J.-Y. Chen, Q. S. Shen, W.-Z. Zhou, J. Peng, B. Z. He, Y. Li, C.-J. Liu, X. Luan, W. Ding, S. Li, C. Chen, B. C.-M. Tan, Y. E. Zhang, A. He, and C.-Y. Li. Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. *PLOS Genet.*, 11(7): e1005391, 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005391. Publisher: Public Library of Science.

R. Chowdhury, N. Bouatta, S. Biswas, C. Floristean, A. Kharkar, K. Roy, C. Rochereau, G. Ahdritz, J. Zhang, G. M. Church, et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

M. B. Clark, P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn, C. P. Ponting, P. F. Stadler, K. V. Morris, A. Morillon, et al. The reality of pervasive transcription. *PLoS biology*, 9(7):e1000625, 2011.

J. M. Coffin and H. Fan. The discovery of reverse transcriptase. *Annual review of virology*, 3:29–51, 2016.

L. Delaye, A. DeLuna, A. Lazcano, and A. Becerra. The origin of a novel gene through overprinting in escherichia coli. *BMC Evolutionary Biology*, 8:1–10, 2008.

M. E. Dinger, K. C. Pang, T. R. Mercer, and J. S. Mattick. Differentiating protein-coding and noncoding rna: challenges and ambiguities. *PLoS computational biology*, 4(11):e1000176, 2008.

E. Dohmen, M. Aubel, L. A. Eicholt, P. Roginski, V. Luria, A. Karger, and A. Grandchamp. Denofo: a file format and toolkit for standardised, comparable de novo gene annotation. *bioRxiv*, 2025. doi: 10.1101/2025.03.31.644673. URL https://www.biorxiv.org/content/early/2025/04/01/2025.03.31.644673.

T. Domazet-Loso and D. Tautz. An Evolutionary Analysis of Orphan Genes in Drosophila. *Genome Research*, 13(10):2213–2219, 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1311003.

T. Domazet-Lošo, A.-R. Carvunis, M. M. Albà, M. S. Šestak, R. Bakarić, R. Neme, and D. Tautz. No Evidence for Phylostratigraphic Bias Impacting Inferences on Patterns of Gene Emergence and Evolution. *Molecular Biology and Evolution*, 34(4):843–856, Apr. 2017. ISSN 0737-4038. doi: 10.1093/molbev/msw284. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5400388/.

D. Dowling, J. F. Schmitz, and E. Bornberg-Bauer. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. *Genome biology and evolution*, 12(11):2183–2195, 2020.

G. Drillon, A. Carbone, and G. Fischer. Synchro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PloS one*, 9(3):e92621, 2014.

E. E. Duffy, B. Finander, G. Choi, A. C. Carter, I. Pritisanac, A. Alam, V. Luria, A. Karger, W. Phu, M. A. Sherman, et al. Developmental dynamics of rna translation in the human brain. *Nature neuroscience*, 25 (10):1353–1365, 2022.

S. R. Eddy. A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009: Genome Informatics Series Vol. 23*, pages 205–211. World Scientific, 2009.

R. A. Farrer. Synima: a synteny imaging tool for annotated genome assemblies. *BMC bioinformatics*, 18:1–4, 2017.

B. Feldmeyer, E. Bornberg-Bauer, E. Dohmen, B. Fouks, J. Heckenhauer, A. K. Huylmans, A. R. C. Jones, E. Stolle, and M. C. Harrison. Comparative evolutionary genomics in insects. *Methods Mol. Biol.*, 2802: 473–514, 2024.

R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–W37, July 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr367.

V. J. Freeman. Studies on the virulence of bacteriophage-infected strains of corynebacterium diphtheriae. *Journal of bacteriology*, 61(6):675–688, 1951.

W. Gilbert. Why genes in pieces? *Nature*, 271(5645):501–501, 1978.

A. Grandchamp, K. Berk, E. Dohmen, and E. Bornberg-Bauer. New genomic signals underlying the emergence of human proto-genes. *Genes*, 13(2):284, 2022.

A. Grandchamp, P. Czuppon, and E. Bornberg-Bauer. Quantification and modeling of turnover dynamics of de novo transcripts in drosophila melanogaster. *Nucleic Acids Research*, page gkad1079, 2023a.

A. Grandchamp, L. Kühl, M. Lebherz, K. Brüggemann, J. Parsch, and E. Bornberg-Bauer. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in drosophila melanogaster. *Genome Research*, 33(6):872–890, 2023b.

F. Griffith. The significance of pneumococcal types. *Epidemiology & Infection*, 27(2):113–159, 1928.

M. Gry, R. Rimini, S. Strömberg, A. Asplund, F. Pontén, M. Uhlén, and P. Nilsson. Correlations between rna and protein expression profiles in 23 human cell lines. *BMC genomics*, 10:1–14, 2009.

X. Gu, Z. Zhang, and W. Huang. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences*, 102(3):707–712, 2005.

S. Y. Guay, P. H. Patel, J. M. Thomalla, K. L. McDermott, J. M. O'Toole, S. E. Arnold, S. J. Obrycki, M. F. Wolfner, and G. D. Findlay. An orphan gene is essential for efficient sperm entry into eggs in drosophila melanogaster. *Genetics*, page iyaf008, 2025.

A. M. Gubala, J. F. Schmitz, M. J. Kearns, T. T. Vinh, E. Bornberg-Bauer, M. F. Wolfner, and G. D. Findlay. The goddard and saturn genes are essential for drosophila male fertility and may have arisen de novo. *Molecular biology and evolution*, 34(5):1066–1082, 2017.

W.-J. Guo, P. Li, J. Ling, and S.-P. Ye. Significant comparative characteristics between orphan and nonorphan genes in the rice (oryza sativa l.) genome. *International Journal of Genomics*, 2007(1):021676, 2007.

M. J. Hangauer, I. W. Vaughn, and M. T. McManus. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding rnas. *PLoS genetics*, 9(6):e1003569, 2013.

B. Heames, J. Schmitz, and E. Bornberg-Bauer. A continuum of evolving de novo genes drives protein-coding novelty in drosophila. *Journal of molecular evolution*, 88(4):382–398, 2020.

T. J. Heinen, F. Staubach, D. Häming, and D. Tautz. Emergence of a new gene from an intergenic region. *Current biology*, 19(18):1527–1531, 2009.

R. Hershberg and D. A. Petrov. Selection on codon bias. *Annual review of genetics*, 42(1):287–299, 2008.

L. D. Hurst. The ka/ks ratio: diagnosing the form of sequence evolution. *TRENDS in Genetics*, 18(9):486–487, 2002.

K. Illergård, D. H. Ardell, and A. Elofsson. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009.

N. T. Ingolia, L. F. Lareau, and J. S. Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, 2011.

H. Innan and F. Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, 2010.

B. R. Iyengar and E. Bornberg-Bauer. Neutral models of de novo gene emergence suggest that gene evolution has a preferred trajectory. *Molecular Biology and Evolution*, 40(4):msad079, 2023.

B. R. Iyengar, A. Grandchamp, and E. Bornberg-Bauer. How antisense transcripts can evolve to encode novel proteins. *Nature Communications*, 15(1):6187, 2024.

P. Janssen, Z. Kliesmete, B. Vieth, X. Adiconis, S. Simmons, J. Marshall, C. McCabe, H. Heyn, J. Z. Levin, W. Enard, et al. The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biology*, 24(1):140, 2023.

Z. Ji, R. Song, A. Regev, and K. Struhl. Many lncrnas, 5'utrs, and pseudogenes are translated and some are likely to express functional proteins. *elife*, 4:e08890, 2015.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, 2021.

Y.-J. Kang, D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, L. Wei, and G. Gao. Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, 45(W1):W12–W16, 2017.

K. Käther, S. Lemke, and P. F. Stadler. Annotation-free identification of potential synteny anchors. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 217–230. Springer, 2023.

D. M. Keeling, P. Garza, C. M. Nartey, and A.-R. Carvunis. The meanings of'function'in biology and the problematic case of de novo gene emergence. *Elife*, 8:e47014, 2019.

P. K. Keese and A. Gibbs. Origins of genes:" big bang" or continuous creation? *Proceedings of the National Academy of Sciences*, 89(20):9489–9493, 1992.

M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.

M. Kellis, B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, et al. Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138, 2014.

W. J. Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

T. Kondo, S. Plaza, J. Zanet, E. Benrabah, P. Valenti, Y. Hashimoto, S. Kobayashi, F. Payre, and Y. Kageyama. Small peptides switch the transcriptional activity of shavenbaby during drosophila embryogenesis. *Science*, 329(5989):336–339, 2010.

S. L. Kosakovsky Pond and S. D. Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5):1208–1222, 2005.

A. Koussounadis, S. P. Langdon, I. H. Um, D. J. Harrison, and V. A. Smith. Relationship between differentially expressed mrna and mrna-protein correlations in a xenograft model system. *Scientific reports*, 5(1):10775, 2015.

S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea. Transcriptome assembly from long-read rna-seq alignments with stringtie2. *Genome biology*, 20(1):1–13, 2019.

M. Kozak. The scanning model for translation: an update. *The Journal of cell biology*, 108(2):229–241, 1989.

S. Kryazhimskiy and J. B. Plotkin. The population genetics of dn/ds. *PLoS genetics*, 4(12):e1000304, 2008.

S. Kumar, M. Suleski, J. M. Craig, A. E. Kasprowicz, M. Sanderford, M. Li, G. Stecher, and S. B. Hedges. Timetree 5: an expanded resource for species divergence times. *Molecular biology and evolution*, 39(8): msac174, 2022.

A. Lange, P. H. Patel, B. Heames, A. M. Damry, T. Saenger, C. J. Jackson, G. D. Findlay, and E. Bornberg-Bauer. Structural and functional characterization of a putative de novo gene in drosophila. *Nature communications*, 12(1):1667, 2021.

M. K. Lebherz, B. Ravi Iyengar, and E. Bornberg-Bauer. Modeling length changes in de novo orfs during neutral evolution. *Genome Biology and Evolution*, page evae129, 2024.

F. Lemoine, O. Lespinet, and B. Labedan. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC evolutionary biology*, 7:1–18, 2007.

A. Z.-X. Leong, P. Y. Lee, M. A. Mohtar, S. E. Syafruddin, Y.-F. Pung, and T. Y. Low. Short open reading frames (sorfs) and microproteins: an update on their identification and validation measures. *Journal of biomedical science*, 29(1):19, 2022.

C.-Y. Li, Y. Zhang, Z. Wang, Y. Zhang, C. Cao, P.-W. Zhang, S.-J. Lu, X.-M. Li, Q. Yu, X. Zheng, et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS computational biology*, 6(3):e1000734, 2010.

J. Li and C. Liu. Coding or noncoding, the converging concepts of rnas. *Frontiers in genetics*, 10:496, 2019.

J. Li, U. Singh, P. Bhandary, J. Campbell, Z. Arendsee, A. S. Seetharam, and E. S. Wurtele. Foster thy young: enhanced prediction of orphan genes in assembled genomes. *Nucleic Acids Research*, 50(7):e37, Dec. 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1238.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

D. Liu, M. Hunt, and I. J. Tsai. Inferring synteny between genome assemblies: a systematic evaluation. *BMC bioinformatics*, 19(1):1–13, 2018.

J. Liu, R. Yuan, W. Shao, J. Wang, I. Silman, and J. L. Sussman. Do "newly born" orphan proteins resemble "never born" proteins? a study using three deep learning algorithms. *Proteins: Structure, Function, and Bioinformatics*, 2023.

Y. Liu, A. Beyer, and R. Aebersold. On the dependency of cellular protein levels on mrna abundance. *Cell*, 165 (3):535–550, 2016.

K. D. Lombardo, H. K. Sheehy, J. M. Cridland, and D. J. Begun. Identifying candidate de novo genes expressed in the somatic female reproductive tract of drosophila melanogaster. *G3: Genes, Genomes, Genetics*, 13(8): jkad122, 2023.

M. Long and C. H. Langley. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science*, 260(5104):91–95, 1993. Publisher: American Association for the Advancement of Science.

E. Matoulkova, E. Michalova, B. Vojtesek, and R. Hrstka. The role of the 3'untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA biology*, 9(5):563–576, 2012.

J. H. McDonald and M. Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351 (6328):652–654, 1991.

A. McLysaght and D. Guerzoni. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140332, 2015.

A. McLysaght and L. D. Hurst. Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, 17(9):567–578, 2016.

J. M. Michaud, A. Madani, and J. S. Fraser. A language model beats alphafold2 on orphans. *Nature Biotechnology*, 40(11):1576–1577, 2022.

L. Middendorf and L. A. Eicholt. Random, de novo, and conserved proteins: how structure and disorder predictors perform differently. *Proteins: Structure, Function, and Bioinformatics*, 92(6):757–767, 2024.

L. Middendorf, B. Ravi Iyengar, and L. A. Eicholt. Sequence, structure, and functional space of drosophila de novo proteins. *Genome Biology and Evolution*, 16(8):evae176, 2024.

F. Mitelman, B. Johansson, and F. Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7(4):233–245, 2007.

D. Moi, C. Bernard, M. Steinegger, Y. Nevers, M. Langleib, and C. Dessimoz. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *bioRxiv*, pages 2023–09, 2023.

B. A. Moyers and J. Zhang. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution. 32(1): 258–267, 2015. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msu286.

B. A. Moyers and J. Zhang. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Molecular Biology and Evolution*, 33(5):1245–1256, 2016. ISSN 1537-1719. doi: 10.1093/molbev/msw008.

B. A. Moyers and J. Zhang. Further Simulations and Analyses Demonstrate Open Problems of Phylostratigraphy. 9(6):1519–1527, 2017. ISSN 1759-6653. doi: 10.1093/gbe/evx109.

B. A. Moyers and J. Zhang. Toward Reducing Phylostratigraphic Errors and Biases. *Genome Biol Evol*, 10(8): 2037–2048, Aug. 2018. doi: 10.1093/gbe/evy161. Publisher: Oxford Academic.

J. M. Mudge, J. Ruiz-Orera, J. R. Prensner, M. A. Brunet, F. Calvet, I. Jungreis, J. M. Gonzalez, M. Magrane, T. F. Martinez, J. F. Schulz, et al. Standardized annotation of translated open reading frames. *Nature biotechnology*, 40(7):994–999, 2022.

S. Naseeb, R. M. Ames, D. Delneri, and S. C. Lovell. Rapid functional and evolutionary changes follow gene duplication in yeast. *Proceedings of the Royal Society B: Biological Sciences*, 284(1861):20171393, 2017.

R. Neme and D. Tautz. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC genomics*, 14:1–13, 2013.

T. O. Nieuwenhuis, A. Z. Rosenberg, M. N. McCall, and M. K. Halushka. Tissue, age, sex, and disease patterns of matrisome expression in gtex transcriptome data. *Scientific reports*, 11(1):21549, 2021.

P. Nowell and D. Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Landmarks in medical genetics: classic papers with commentaries*, 132(51):103, 2004.

S. Ohno. *Evolution by Gene Duplication*. 1970. doi: 10.1007/978-3-642-86659-3.

M. Oliva, M. Muñoz-Aguirre, S. Kim-Hellmuth, V. Wucher, A. D. Gewirtz, D. J. Cotter, P. Parsana, S. Kasela, B. Balliu, A. Viñuela, et al. The impact of sex on gene expression across human tissues. *Science*, 369(6509): eaba3066, 2020.

V. Orgogozo, A. E. Peluffo, and B. Morizot. The "mendelian gene" and the "molecular gene": two relevant concepts of genetic units. *Current topics in developmental biology*, 119:1–26, 2016.

Á. N. O'Toole, L. D. Hurst, and A. McLysaght. Faster evolving primate genes are more likely to duplicate. *Molecular biology and evolution*, 35(1):107–118, 2018.

N. Palmieri, C. Kosiol, and C. Schlötterer. The life cycle of drosophila orphan genes. *elife*, 3:e01311, 2014.

C. Papadopoulos, I. Callebaut, J.-C. Gelly, I. Hatin, O. Namy, M. Renard, O. Lespinet, and A. Lopes. Intergenic orfs as elementary structural modules of de novo gene birth and protein evolution. *Genome Research*, 31 (12):2303–2315, 2021.

C. Papadopoulos, H. Arbes, D. Cornu, N. Chevrollier, S. Blanchet, P. Roginski, C. Rabier, S. Atia, O. Lespinet, O. Namy, and A. Lopes. The ribosome profiling landscape of yeast reveals a high diversity in pervasive translation. *Genome Biology*, 25(1):268, Oct. 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03403-7. URL `https://doi.org/10.1186/s13059-024-03403-7`.

P. Patraquim, M. A. S. Mumtaz, J. I. Pueyo, J. L. Aspden, and J.-P. Couso. Developmental regulation of canonical and small ORF translation from mRNAs. *Genome Biology*, 21(1):128, May 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02011-5. URL `https://doi.org/10.1186/s13059-020-02011-5`.

P. Patraquim, E. G. Magny, J. I. Pueyo, A. I. Platero, and J. P. Couso. Translation and natural selection of micropeptides from long non-canonical RNAs. *Nature Communications*, 13(1):6515, Oct. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34094-y.

A. Pauli, M. L. Norris, E. Valen, G.-L. Chew, J. A. Gagnon, S. Zimmerman, A. Mitchell, J. Ma, J. Dubrulle, D. Reyon, et al. Toddler: an embryonic signal that promotes cell movement via apelin receptors. *Science*, 343(6172):1248636, 2014.

A. Pavesi. Origin and evolution of overlapping genes in the family microviridae. *Journal of General Virology*, 87(4):1013–1017, 2006.

C. Pegueroles, S. Laurie, and M. M. Albà. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular biology and evolution*, 30(8):1830–1842, 2013.

J. Peng and L. Zhao. The origin and structural evolution of de novo genes in drosophila. *Nature Communications*, 15(1):810, 2024.

M. Peng, T. Wang, Y. Li, Z. Zhang, and C. Wan. Mapping start codons of small open reading frames by n-terminomics approach. *Molecular & Cellular Proteomics*, 23(11):100860, 2024.

R. Petryszak, M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett, A. Füllgrabe, A. M.-P. Fuentes, S. Jupp, S. Koskinen, et al. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic acids research*, 44(D1):D746–D752, 2016.

M. Poretti, C. R. Praz, A. G. Sotiropoulos, and T. Wicker. A survey of lineage-specific genes in triticeae reveals de novo gene evolution from genomic raw material. *Plant Direct*, 7(3):e484, 2023.

N. Prabh and C. Rödelsperger. De Novo, Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in Pristionchus Nematodes. *G3: Genes|Genomes|Genetics*, 9(7):2277–2286, May 2019. ISSN 2160-1836. doi: 10.1534/g3.119.400326. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6643871/.

J. R. Prensner, O. M. Enache, V. Luria, K. Krug, K. R. Clauser, J. M. Dempster, A. Karger, L. Wang, K. Stumbraite, V. M. Wang, et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nature biotechnology*, 39(6):697–704, 2021.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(suppl_1):D501–D504, 2005.

A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

V. Raghavan, L. Kraft, F. Mesny, and L. Rigerte. A simple guide to de novo transcriptome assembly and annotation. *Briefings in bioinformatics*, 23(2):bbab563, 2022.

C. Rancurel, M. Khosravi, A. K. Dunker, P. R. Romero, and D. Karlin. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *Journal of virology*, 83(20): 10719–10736, 2009.

J. M. Ranz, F. Casals, and A. Ruiz. How Malleable is the Eukaryotic Genome? Extreme Rate of Chromosomal Rearrangement in the Genus Drosophila. *Genome Research*, 11(2):230–239, Feb. 2001. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.162901.

P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends in genetics*, 16(6):276–277, 2000.

E. L. Rivard, A. G. Ludwig, P. H. Patel, A. Grandchamp, S. E. Arnold, A. Berger, E. M. Scott, B. J. Kelly, G. C. Mascha, E. Bornberg-Bauer, et al. A putative de novo evolved gene required for spermatid chromatin condensation in drosophila melanogaster. *PLoS genetics*, 17(9):e1009787, 2021.

P. Roginski, A. Grandchamp, C. Quignot, and A. Lopes. De novo emerged gene search in eukaryotes with dense. *Genome Biology and Evolution*, 16(8):evae159, 2024.

I. B. Rogozin, A. N. Spiridonov, A. V. Sorokin, Y. I. Wolf, I. K. Jordan, R. L. Tatusov, and E. V. Koonin. Purifying and directional selection in overlapping prokaryotic genes. *Trends in Genetics*, 18(5):228–232, 2002.

I. T. Rombel, K. F. Sykes, S. Rayner, and S. A. Johnston. Orf-finder: a vector for high-throughput gene identification. *Gene*, 282(1-2):33–41, 2002.

J. Ruiz-Orera and M. M. Albà. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation, 2019. ISSN: 13624555 Publication Title: Trends in Genetics.

J. Ruiz-Orera, X. Messeguer, J. A. Subirana, and M. M. Alba. Long non-coding RNAs as a source of new peptides. *eLife*, 3:1–24, 2014. ISSN 2050084X. doi: 10.7554/eLife.03523.

J. Ruiz-Orera, J. L. Villanueva-Cañas, W. Blevins, and M. Albà. De novo gene evolution: How do we transition from non-coding to coding? *PeerJ Preprints*, 2017.

C.-L. Sandmann, J. F. Schulz, J. Ruiz-Orera, M. Kirchner, M. Ziehm, E. Adami, M. Marczenke, A. Christ, N. Liebe, J. Greiner, et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Molecular cell*, 83(6):994–1011, 2023.

N. Scalzitti, A. Jeannin-Girardon, P. Collet, O. Poch, and J. D. Thompson. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC genomics*, 21:1–20, 2020.

D. Schlesinger and S. J. Elsässer. Revisiting sorfs: overcoming challenges to identify and characterize functional microproteins. *The FEBS journal*, 289(1):53–74, 2022.

C. Schlötterer. Genes from scratch–the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4):215–219, 2015.

J. Schmitz and J. Brosius. Exonization of transposed elements: a challenge and opportunity for evolution. *Biochimie*, 93(11):1928–1934, 2011.

J. F. Schmitz, K. K. Ullrich, and E. Bornberg-Bauer. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature ecology & evolution*, 2(10):1626–1632, 2018.

A. L. Schneider, R. Martins-Silva, A. Kaizeler, N. Saraiva-Agostinho, and N. L. Barbosa-Morais. voyager, a free web interface for the analysis of age-related gene expression alterations in human tissues. *Elife*, 12:RP88623, 2024.

C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020.

U. Singh and E. S. Wurtele. orfipy: a fast and flexible tool for extracting orfs. *Bioinformatics*, 37(18):3019–3020, 2021.

G. S. C. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6:1–11, 2005.

S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. A. Budnik, J. L. Rinn, and A. Saghatelian. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology*, 9(1):59–64, 2013.

J. Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

R. Sorek. The birth of new exons: mechanisms and evolutionary consequences. *Rna*, 13(10):1603–1608, 2007.

G. Stoesser, W. Baker, A. van den Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, et al. The embl nucleotide sequence database. *Nucleic acids research*, 30(1):21–26, 2002.

D. Tautz and T. Domazet-Lošo. The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10): 692–702, 2011.

H. M. Temin, S. Mizutami, et al. Rna-dependent dna polymerase in virions of rous sarcoma virus. *Nature*, 226: 1211–1213, 1970.

K. E. Thomas, P. A. Gagniuc, and E. Gagniuc. Moonlighting genes harbor antisense orfs that encode potential membrane proteins. *Scientific Reports*, 13(1):12591, 2023.

M. Toll-Riera, N. Bosch, N. Bellora, R. Castelo, L. Armengol, X. Estivill, and M. Mar Alba. Origin of primate orphan genes: a comparative genomics approach. *Molecular biology and evolution*, 26(3):603–612, 2009.

A. Turcan, J. Lee, A. Wacholder, and A.-R. Carvunis. Integrative detection of genome-wide translation using iribo. *STAR Protocols*, 5(1):102826, 2024. ISSN 2666-1667. doi: https://doi.org/10.1016/j.xpro.2023.102826.

N. Vakirlis and A. McLysaght. Computational prediction of de novo emerged protein-coding genes. *Computational methods in protein evolution*, pages 63–81, 2019.

N. Vakirlis, A. S. Hebert, D. A. Opulente, G. Achaz, C. T. Hittinger, G. Fischer, J. J. Coon, and I. Lafontaine. A molecular portrait of de novo genes in yeasts. *Molecular Biology and Evolution*, 35(3):631–645, 2018.

N. Vakirlis, A.-R. Carvunis, and A. McLysaght. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife*, 9:e53500, 2020.

N. Vakirlis, Z. Vance, K. M. Duggan, and A. McLysaght. De novo birth of functional microproteins in the human lineage. *Cell reports*, 41(12), 2022.

N. Vakirlis, O. Acar, V. Cherupally, and A.-R. Carvunis. Ancestral Sequence Reconstruction as a Tool to Detect and Study De Novo Gene Emergence. *Genome Biology and Evolution*, 16(8):evae151, Aug. 2024. ISSN 1759-6653. doi: 10.1093/gbe/evae151. URL https://doi.org/10.1093/gbe/evae151.

M. Van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. Gilchrist, J. Söding, and M. Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

S. B. Van Oss and A.-R. Carvunis. De novo gene birth. *PLoS genetics*, 15(5):e1008160, 2019.

C. Vara, J. C. Montañés, and M. M. Albà. High polymorphism levels of de novo orfs in a yoruba human population. *Genome Biology and Evolution*, 16(7):evae126, 2024.

A. Varabyou, B. Erdogdu, S. L. Salzberg, and M. Pertea. Investigating open reading frames in known and novel transcripts using orfanage. *Nature computational science*, 3(8):700–708, 2023.

K. Vitting-Seerup, B. T. Porse, A. Sandelin, and J. Waage. splicer: an r package for classification of alternative splicing and prediction of coding potential from rna-seq data. *BMC bioinformatics*, 15:1–7, 2014.

A. Vonica, N. Bhat, K. Phan, J. Guo, L. Iancu, J. A. Weber, A. Karger, J. W. Cain, E. C. Wang, G. M. DeStefano, et al. Apcdd1 is a dual bmp/wnt inhibitor in the developing nervous system and skin. *Developmental biology*, 464(1):71–87, 2020.

A. Wacholder and A.-R. Carvunis. Biological factors and statistical limitations prevent detection of most noncanonical proteins by mass spectrometry. *PLoS Biology*, 21(12):e3002409, 2023.

A. Wacholder, S. B. Parikh, N. C. Coelho, O. Acar, C. Houghton, L. Chou, and A.-R. Carvunis. A vast evolutionarily transient translatome contributes to phenotype and fitness. *Cell Systems*, 14(5):363–381, 2023.

E. W. Wallace, E. M. Airoldi, and D. A. Drummond. Estimating selection on synonymous codon usage from noisy experimental data. *Molecular biology and evolution*, 30(6):1438–1453, 2013.

L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li. Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6):e74–e74, 2013.

Y. Wang, H. Tang, J. D. DeBarry, X. Tan, J. Li, X. Wang, T.-h. Lee, H. Jin, B. Marler, H. Guo, et al. Mcscanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*, 40 (7):e49–e49, 2012.

Z. Wang, Y. Chen, and Y. Li. A brief review of computational gene prediction methods. *Genomics, proteomics & bioinformatics*, 2(4):216–221, 2004.

L. D. Ward and M. Kellis. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337(6102):1675–1678, 2012.

C. M. Weisman. The Origins and Functions of De Novo Genes: Against All Odds? *Journal of Molecular Evolution*, 90(3):244–257, Aug. 2022. ISSN 1432-1432. doi: 10.1007/s00239-022-10055-3.

C. M. Weisman, A. W. Murray, and S. R. Eddy. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS biology*, 18(11):e3000862, 2020a.

C. M. Weisman, A. W. Murray, and S. R. Eddy. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biology*, 18(11):e3000862, Nov. 2020b. ISSN 1545-7885. doi: 10.1371/journal.pbio.3000862. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000862. Publisher: Public Library of Science.

C. M. Weisman, A. W. Murray, and S. R. Eddy. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology*, 32(12):2632–2639.e2, 2022. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2022.04.085. URL https://www.sciencedirect.com/science/article/pii/S0960982222007217.

N. Whiffin, K. J. Karczewski, X. Zhang, S. Chothani, M. J. Smith, D. G. Evans, A. M. Roberts, N. M. Quaife, S. Schafer, O. Rackham, et al. Characterising the loss-of-function impact of 5'untranslated region variants in 15,708 individuals. *Nature communications*, 11(1):2523, 2020.

B. A. Wilson and J. Masel. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biology and Evolution*, 3(1):1245–1252, 2011. ISSN 17596653. doi: 10.1093/gbe/evr099.

B. W. Wright, Z. Yi, J. S. Weissman, and J. Chen. The dark proteome: translation from noncanonical open reading frames. *Trends in Cell Biology*, 32(3):243–258, 2022.

D.-D. Wu, D. M. Irwin, and Y.-P. Zhang. De novo origin of human protein-coding genes. *PLoS genetics*, 7 (11):e1002379, 2011.

S. Xia, J. Chen, D. Arsala, J. Emerson, and M. Long. Functional innovation through new genes as a general evolutionary process. *Nature genetics*, pages 1–15, 2025.

A. Xu, B. B. Teefy, R. J. Lu, S. Nozownik, A. M. Tyers, D. R. Valenzano, and B. A. Benayoun. Transcriptomes of aging brain, heart, muscle, and spleen from female and male african turquoise killifish. *Scientific Data*, 10 (1):695, 2023.

H. Xu, P. Wang, Y. Fu, Y. Zheng, Q. Tang, L. Si, J. You, Z. Zhang, Y. Zhu, L. Zhou, et al. Length of the orf, position of the first aug and the kozak motif are important factors in potential dual-coding transcripts. *Cell research*, 20(4):445–457, 2010.

Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution*, 15(12):496–503, 2000.

E. M. Zdobnov, C. von Mering, I. Letunic, D. Torrents, M. Suyama, R. R. Copley, G. K. Christophides, D. Thomasova, R. A. Holt, G. M. Subramanian, H.-M. Mueller, G. Dimopoulos, J. H. Law, M. A. Wells, E. Birney, R. Charlab, A. L. Halpern, E. Kokoza, C. L. Kraft, Z. Lai, S. Lewis, C. Louis, C. Barillas-Mury, D. Nusskern, G. M. Rubin, S. L. Salzberg, G. G. Sutton, P. Topalis, R. Wides, P. Wincker, M. Yandell, F. H. Collins, J. Ribeiro, W. M. Gelbart, F. C. Kafatos, and P. Bork. Comparative Genome and Proteome Analysis of Anopheles gambiae and Drosophila melanogaster. 298(5591):149–159, 2002. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1077061.

L. Zhang, Y. Ren, T. Yang, G. Li, J. Chen, A. R. Gschwend, Y. Yu, G. Hou, J. Zi, R. Zhou, et al. Rapid evolution of protein diversity by de novo origination in oryza. *Nature ecology & evolution*, 3(4):679–690, 2019.

L. Zhao, P. Saelao, C. D. Jones, and D. J. Begun. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science (New York, N.Y.)*, 343(6172):769–772, Feb. 2014. ISSN 0036-8075. doi: 10.1126/science.1248286. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391638/`.

T. Zhao and M. E. Schranz. Network approaches for plant phylogenomic synteny analysis. *Current opinion in plant biology*, 36:129–134, 2017.

E. B. Zheng and L. Zhao. Protein evidence of unannotated orfs in drosophila reveals diversity in the evolution and properties of young proteins. *Elife*, 11:e78772, 2022.