De Novo Emerged Gene *Search* in Eukaryotes with DENSE

Paul Roginski [1], Anna Grandchamp [2], Chloé Quignot [1], Anne Lopes [1,*]

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

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

*Corresponding author: E-mail: anne.lopes@i2bc.paris-saclay.fr.

Accepted: July 07, 2024

Abstract

The discovery of de novo emerged genes, originating from previously noncoding DNA regions, challenges traditional views of species evolution. Indeed, the hypothesis of neutrally evolving sequences giving rise to functional proteins is highly unlikely. This conundrum has sparked numerous studies to quantify and characterize these genes, aiming to understand their functional roles and contributions to genome evolution. Yet, no fully automated pipeline for their identification is available. Therefore, we introduce DENSE (DE Novo emerged gene SEarch), an automated Nextflow pipeline based on two distinct steps: detection of taxonomically restricted genes (TRGs) through phylostratigraphy, and filtering of TRGs for de novo emerged genes via genome comparisons and synteny search. DENSE is available as a user-friendly command-line tool, while the second step is accessible through a web server upon providing a list of TRGs. Highly flexible, DENSE provides various strategy and parameter combinations, enabling users to adapt to specific configurations or define their own strategy through a rational framework, facilitating protocol communication, and study interoperability. We apply DENSE to seven model organisms, exploring the impact of its strategies and parameters on de novo gene predictions. This thorough analysis across species with different evolutionary rates reveals useful metrics for users to define input datasets, identify favorable/unfavorable conditions for de novo gene detection, and control potential biases in genome annotations. Additionally, predictions made for the seven model organisms are compiled into a requestable database, which we hope will serve as a reference for de novo emerged gene lists generated with specific criteria combinations.

Key words: de novo emerged gene prediction, phylostratigraphy, noncoding ORFs, comparative genomics, genome evolution.

Significance

The identification and classification of de novo genes, which originate from noncoding regions of DNA, remain an ongoing challenge in genomic research. While various approaches have been employed for their identification, the lack of a standardized protocol has resulted in varying lists of de novo genes across studies. This study introduces a novel tool: DENSE (DE Novo emerged gene SEarch), that formalizes the common practices used in the field into a comprehensive and automated pipeline. DENSE streamlines the identification of taxonomically restricted genes, homology searches, and synteny analysis. This standardized methodology aims to enhance the accuracy and reliability of de novo gene identification, fostering a deeper understanding of the evolutionary mechanisms that drive gene birth and shape the genetic diversity of organisms.

Introduction

Comparative genomics has unveiled the existence of what we call de novo emerged genes—genes that arose from a DNA region that was ancestrally noncoding. Initially considered unlikely (Jacob 1977), the accumulation of sequencing data revealed that they were, in fact, widespread, being detected in various eukaryotic species and numerous, with several dozens of examples reported for different

organisms (Levine et al. 2006; Cai et al. 2008; Schlötterer 2015; Van Oss and Carvunis 2019). The discovery of these genes has highlighted that the passage from the noncoding world to the world of functional products and regulated processes was much more frequent than previously thought, challenging previous assumptions and raising an intriguing question: How can a noncoding neutrally evolving DNA sequence give rise to a functional product able to take part in the well-established biological networks and contribute to the organism's fitness? Indeed, noncoding regions are associated with different nucleotide compositions and combinations from coding regions and are consequently expected to result in nonfunctional combinations of amino acids if translated. The newly attributed role of noncoding regions and the unexpected permeability observed between these two worlds have thus captured the attention of researchers in recent years. Notably, numerous studies have been undertaken on different species, including several model organisms, to characterize de novo emerged genes (Bungard et al. 2017; Schmitz et al. 2018; Vakirlis et al. 2018; Lange et al. 2021; Papadopoulos et al. 2021; Peng and Zhao 2024). This effort not only aims to elucidate their properties and functions but also to enhance our understanding of the transition between the noncoding and the coding worlds, as well as their evolutionary relationship. The strategy employed to detect de novo emerged genes is likely to influence the resulting interpretations; therefore, it is important to design rational and reproducible methods.

So far, most model organisms have been associated with multiple lists of de novo emerged genes, as illustrated in Table 1 of the review by Van Oss and Carvunis (2019). These lists result from diverse protocols that generally rely on transcriptomics and/or comparative genomics. Nonetheless, despite this diversity, all these methods, for classifying genes as de novo emerged, impose that genes be novel or identified only in closely related species. Transcriptomics-based protocols either search for novel transcripts or require the transcription of de novo gene candidates, usually detected with an initial step of comparative genomics (Zhang et al. 2019; Blevins et al. 2021; Grandchamp et al. 2023). Moreover, ribosome profiling and proteomics can provide additional evidence for the ability of these novel transcripts to be translated into proteins (Carvunis et al. 2012; Zhang et al. 2019; Blevins et al. 2021; Prensner et al. 2023). However, all these newly expressed candidates, especially those arising from pervasive expression, are not necessarily expected to be functional. A significant fraction of them is likely to be short lived in evolutionary history and can instead be considered gene precursors (Grandchamp et al. 2023; Wacholder et al. 2023). As discussed in previous reports, transcription may predate gene emergence and, therefore, cannot be considered a sufficient criterion for a novel transcript to be classified as a gene (Cai et al. 2008; Carvunis et al. 2012; Reinhardt et al. 2013; Schlötterer 2015; Chen et al. 2020; Papadopoulos et al. 2023). Furthermore, young de novo emerged genes are typically associated with stress response or adaptation and are expected to be expressed under specific conditions (Colbourne et al. 2011; Donoghue et al. 2011; Carvunis et al. 2012; Schlötterer 2015; Van Oss and Carvunis 2019). Thus, demonstrating the expression of such genes involves finding the conditions under which they are expressed, which is not trivial. Consequently, requiring de novo gene candidates to be transcribed may be accompanied by high ratios of false negatives.

On the other hand, comparative genomics appears as an appealing solution since it can be applied once the complete genome of the species of interest (i.e. the focal species) and those of close neighboring species are available. Although these methods may involve more or less stringent criteria, all of them focus on genes that have emerged recently, and first rely on the detection of taxonomically restricted genes (TRGs; Vakirlis et al. 2018; Vakirlis and McLysaght 2019; Weisman 2022; Peng and Zhao 2024). The latter correspond to genes found in a single species (i.e. orphan genes) or closely related species, a prerequisite for finding young de novo emerged genes. The detection of TRGs generally relies on phylostratigraphy, which estimates the age of a gene of interest from the detection of its homologs across a phylogeny (Domazet-Loso et al. 2007; Arendsee, Li, Singh, Seetharam, et al. 2019; Barrera-Redondo et al. 2023). In practice, the idea is to screen each gene of the focal genome against a large sequence database (typically, nr or uniprot) with BLAST or an equivalent tool (Altschul et al. 1990; Buchfink et al. 2021). The genes are then assigned an evolutionary age corresponding to the common ancestor of all the lineages in which they have been detected, i.e. the most distant node in the phylogeny where BLAST has detected a homolog of the considered gene. As such, it becomes evident that the criteria used for the homology search are critical and deserve to be defined with caution (Moyers and Zhang 2015; Domazet-Lošo et al. 2017). Indeed, homology detection failure would be accompanied by an underestimation of gene age, thereby wrongly annotating old genes as TRGs (McLysaght and Hurst 2016; Vakirlis et al. 2020; Weisman et al. 2020). However, thorough analyses that aimed to assess the impact of BLAST criteria on gene age estimation have demonstrated that the trends deduced from phylostratigraphy-based approaches remain robust in the face of BLAST's lack of sensitivity, with some studies reporting optimal *E*-value thresholds in the context of TRG detection (Domazet-Loso et al. 2007; Domazet-Lošo et al. 2017; Vakirlis et al. 2020).

Finally, TRGs consist of a heterogeneous population and encompass genes with different origins, including genes resulting from duplication or horizontal transfer followed by high divergence or gene families that have undergone

multiple loss events. Ideally, one would aim to reconstruct the ancestral noncoding sequence to retrace the evolutionary stages that led to de novo gene birth. Since it is not always possible (it requires the detection of an outgroup noncoding sequence with respect to the ancestor to be predicted), different approaches have been described in the literature to discriminate de novo emerged genes from the other TRGs (Vakirlis and McLysaght 2019; Van Oss and Carvunis 2019; Weisman 2022; Peng and Zhao 2024). Specifically, inferring that the locus of the gene candidate was noncoding in the ancestor through the analysis of the sister lineages offers an attractive alternative, since it provides solid support for de novo gene birth. Methods, therefore, require identifying homology traces of the gene of interest within the noncoding regions of a species where the gene is absent. Some approaches even impose additional constraints to strengthen the evidence that the corresponding locus was noncoding in the ancestor. Typically, these homology traces may be imposed to be found in a noncoding region of what we refer to as outgroup species (Zhang et al. 2019; Weisman 2022; Peng and Zhao 2024). The latter are defined as those for which the gene is absent and that branch in the tree after the last species where the gene is present. In addition, to further assert the ancestor's noncoding status, synteny-based approaches may be employed to guarantee the correct identification of the orthologous noncoding region. The latter, therefore, search for the syntenic region in the outgroup species and verify the noncoding status of the homologous region (Knowles and McLysaght 2009; Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016; Van Oss and Carvunis 2019). Various strategies have been undertaken for the identification of syntenic regions (Arendsee, Li, Singh, Bhandary, et al. 2019; Armstrong et al. 2020; Elghraoui et al. 2023). If synteny blocks can be readily detected within very closely related species with high-quality genome assemblies, this task becomes difficult as the evolutionary distance between species increases when studying genomes associated with high rates of chromosomal rearrangements, or simply when dealing with multiple contig genome assemblies (Ranz et al. 2001; Liu et al. 2018). Microsynteny, which searches for local gene order, nevertheless, offers a good compromise (Vakirlis and McLysaght 2019). Beyond the fact that we have no prior knowledge of the recombination rates in the regions that constitute hotbeds for de novo gene birth, the use of microsynteny enables one to handle genome assemblies of intermediate quality, therefore extending the applicability of such methods to wider genomic contexts.

Although a methodological consensus in comparative genomics-based approaches for de novo gene prediction seems to have emerged in recent years, to the best of our knowledge, the scientific community still lacks a fully automated pipeline. Moreover, different combinations of criteria and parameters are still reported, and no definitive protocol has yet been established. This hinders reproducibility and fair comparisons between studies, yet is essential to decipher and eventually reconcile contradictory trends. Therefore, we propose DE Novo emerged gene SEarch (DENSE), an automated pipeline that handles the entire process of de novo emerged gene detection, from identifying TRGs to filtering for those likely to have emerged de novo (https://github.com/i2bc/dense). Since protocols may continue to evolve, but also due to the heterogeneity in the quality of genome assemblies and/or annotations among species, DENSE has been designed to be highly flexible and offers various combinations of filters and parameters embedded in a unified Nextflow framework (Di Tommaso et al. 2017). In this manuscript, we introduce DENSE and investigate the impact of its different implemented strategies, as well as the influence of input data, on the prediction of de novo emerged genes. Finally, we present several metrics that can help users define their input dataset, identify favorable/unfavorable conditions for the detection of de novo emerged genes, and control for potential bias in genome annotations.

Results

Principle of DENSE

DENSE consists of two main independent steps: (i) search for TRGs among the annotated genes of a focal genome and (ii) identification, through a cascade of filters, of TRFs that have homology traces in the orthologous region of a species where the gene is absent (Fig. 1). Specifically, DENSE starts with the genomes of the focal species and those of its neighboring species, along with their corresponding phylogenetic tree. Then, based on the phylostratigraphy calculated by GenEra (Barrera-Redondo et al. 2023), it predicts the date of emergence of all annotated coding sequences (CDSs) of the focal genome with the assumption that horizontal transfers are rare in eukaryotic species. To do so, GenEra screens each CDS against the nr database and the annotated CDSs of the neighboring genomes with DIAMONDv2 (Buchfink et al. 2021; Fig. 1a). Alternatively, users have the option to screen other databases, such as UniProt, SwissProt, or a custom database. It is worth noting that computing phylostratigraphy with nr can be highly time-consuming for large genomes (e.g. ~23 and ~35 h on 40 CPUs for *Mus musculus* and *Homo sapiens*, respectively). However, for genomes of small or intermediate sizes, the computational time is much more acceptable (e.g. ~3, ~8, and ~11 h on 40 CPUs for *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Oryza sativa*, respectively). In any case, we recommend using nr for a more accurate age estimation. Each CDS is then assigned an evolutionary age, corresponding

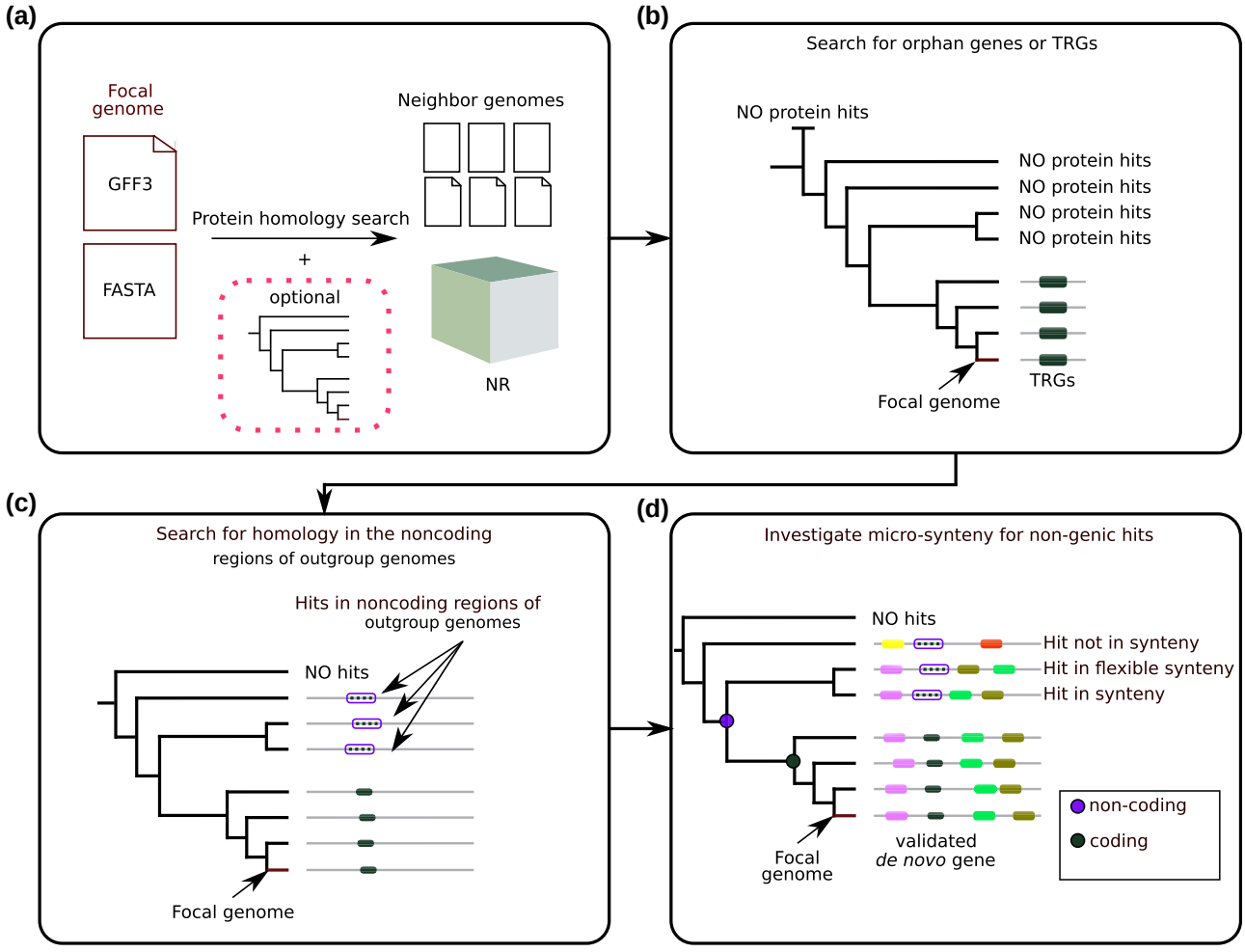

Fig. 1. DENSE workflow. a) DENSE, through GenEra (Barrera-Redondo et al. 2023), screens all the focal genome's (FGs) annotated CDSs against the nr database and the genomes of the focal's neighbors. Depending on the chosen strategy (mandatory for Strategies 1 and 3), the tree of the FG and its neighbors (local tree) must be provided. Then, DENSE outputs the estimated ages of each CDS of the FG and extracts its TRGs, including orphans, according to the phylostratum threshold indicated by the user (genus level by default). b) Example of the conservation of a focal TRG across the local tree. If no local tree has been provided, only the presence/absence across the neighbors is considered. c) Here, we assume that Strategy 1 has been called. DENSE screens the noncoding regions of the outgroup species (only applicable for the focal's neighbors) with tblastn. In this example, three genomes are associated with a tblastn hit in a noncoding region (dashed lines surrounded by purple borders). d) We assume that the user requires the noncoding hits to be found in synteny with the de novo emerged gene candidate. Therefore, DENSE verifies that at least one of the noncoding hits of the outgroup species is effectively detected in a region that is syntenic to that of the candidate. If so, the corresponding locus of their associated Most Recent Common Ancestor (MRCA) is considered as noncoding while that of the MRCA of the lineages where the candidate has been detected is considered as coding.

to the most distant node in the NCBI phylogeny where DIAMONDv2 has detected the CDS. In cases where a gene has multiple isoforms, the age of the oldest one is assigned to all isoforms. DENSE provides the user with all CDSs associated with their predicted evolutionary age and the list of the TRGs of the focal species (Fig. 1b). One should note that, by default, the phylostratum level used to define TRGs is set to the genus level, but depending on the studied species and the aims of the user, DENSE offers the possibility to modify it to younger or older phylostrata. To limit the number of false positives in TRG detection, DIAMONDv2 is called the sensitive mode with a threshold *E*-value of 10^{-5} that has been shown to be optimal for the identification of orphan genes in *S. cerevisiae*, *D. melanogaster*, and *H. sapiens* (Vakirlis et al. 2020).

During the second step, DENSE focuses on the focal genome and the neighboring genomes provided by the user to apply a combination of filters in order to distinguish de novo emerged genes from the TRGs identified by GenEra in the initial step. Users can choose from several strategies, each associated with specific filter and parameter combinations (Fig. 2a). Here, we present one of the most stringent strategies, Strategy 1, which mandates at least one outgroup noncoding hit combined with a search of synteny.

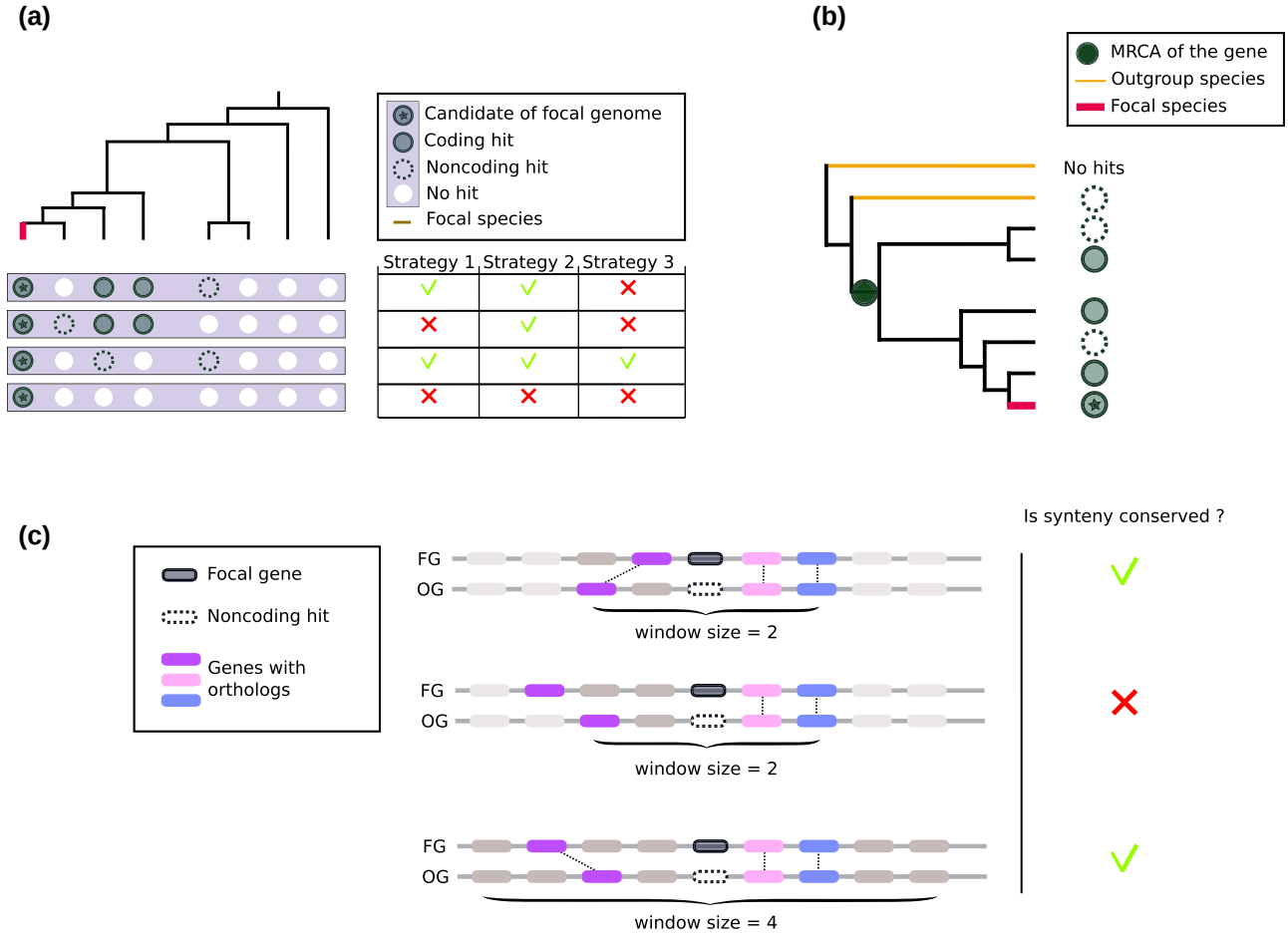

Fig. 2. DENSE strategies and definitions. a) Strategies proposed by DENSE. Strategy 1 requires a noncoding hit in at least one outgroup species. The number of outgroup species in which noncoding hits have been observed can be modified by the user. Strategy 2 only asks the candidate to have a noncoding hit, irrespective of the location of the corresponding species in the tree. Strategy 3 is the most stringent one since it requires candidates retained by Strategy 1 to be orphans with a noncoding hit. Here, the absence/presence of coding and noncoding hits of four fictive genes is represented. The table indicates for each of them whether it would be predicted as de novo emerged according to the employed strategy. Note that all strategies can be combined to the additional requirement of synteny conservation. b) Definition of what is referred to as an outgroup species in DENSE for a given gene. Same tree, as (a), with the focal species represented in red (large branch at the bottom of the tree). The absence/presence of coding and noncoding hits of the gene is represented on the right of the tree following the same scheme as in (a). The dark green node corresponds to the MRCA of the lineages where the gene is present, and all species that branch after this node are considered outgroup species according to this gene. c) Examples of synteny conservation checking for a de novo gene candidate (in black) of a focal genome (FG) and its homologous noncoding hit (in white, dashed borders) in an outgroup genome (OG). DENSE defines two windows of the same size located upstream and downstream of the focal gene (focal window) and its noncoding counterpart (target window). In all examples, at least one gene from each focal window must be detected in the target windows, with one gene located in the upstream and the other in the downstream target window (i.e. number of anchor pairs set to one). In the first example, one of the two genes of the upstream focal window is retrieved in the upstream target window (purple orthologous pair forming the upstream anchor), while the two genes of the downstream focal window are retrieved in the downstream target window (orthologous pairs in pink and blue forming the downstream anchors). The synteny is then considered conserved since each window is associated with at least one anchor gene. In the second example, no anchor is found within the upstream window (i.e. the purple ortholog is located outside the target window), thus the conservation of synteny is not validated. In the last example, the window size is extended to four genes, enabling the purple ortholog in the outgroup species to be retrieved. The two focal windows are associated with an anchor, thus the conservation of synteny is validated.

This involves employing tblastn for each TRG to search for sequence homology within the noncoding regions of the outgroup species. These outgroup species represent those branching after the last common ancestor node in the phylogenetic tree where the TRG's homologs were detected (Figs. 1c and d, 2b; Gertz et al. 2006). Then, using microsynteny, DENSE checks whether the noncoding hit(s) detected in the outgroup species are located in synteny with the focal de novo gene candidate (Figs. 1d and 2c). Therefore, DENSE defines two gene windows of a specific size (set to four genes by default), flanking the focal gene (focal windows) and its homologous noncoding hit

(target windows). It then requires at least one upstream gene and one downstream gene from the focal windows (number of anchor pairs set to one) to be present in the target windows. To do so, DENSE employs the Reciprocal Best Hit method (E-value = 10^{-3}, and query coverage = 70%) to search for the orthologs of the focal windows within the outgroup neighbors. It then verifies the presence of at least two of these orthologs in the target windows, with one in the upstream window and the other in the downstream one (Fig. 2c). One should notice that the number of anchor gene pairs and the size of the windows are parameterizable and that synteny search can be combined with any of the three strategies available in DENSE. Finally, all the TRGs that fulfill these criteria (i.e. one hit in a syntenic noncoding region of at least one outgroup species) are considered genes that have recently emerged de novo (i.e. de novo gene candidates). Nevertheless, DENSE is very flexible, and users can provide their own list of TRGs and enter the pipeline directly at the filtering step (Fig. 1c) or define their own criteria. Notably, users can deactivate the verification of microsynteny and/or focus on candidates with hits in the noncoding regions of the neighboring species regardless of the fact that the species are outgroups. If removing filters may be expected to be accompanied by increases in false positive rates, this can nevertheless be useful when the phylogeny of the considered species is unknown or incomplete or when the outgroup species are too far to detect homology traces in their noncoding regions.

Application of DENSE to Seven Model Organisms

To illustrate the usage of DENSE, we applied it to seven model organisms that are well annotated and for which at least four neighboring species were sequenced and annotated (see list in Table 1 and complete list of neighbors in supplementary table S1, Supplementary Material online). We used the default parameters, which require genes to have a tblastn hit in a syntenic noncoding region of an outgroup species in order to be predicted as de novo emerged (Strategy 1 with synteny, using a window of four genes with one anchor pair). With the exception of *O. sativa*, we detected several dozen to a few hundred de novo gene candidates in all model organisms (Table 1 for the number of candidates detected at each step). The majority of detected de novo gene candidates, irrespective of the focal species, are orphans (68%; Fig. 3) or identified in the closest neighbors, illustrating the difficulty of detecting events of de novo gene emergence older than several million years ago. Interestingly, DENSE predicted ~2,500 de novo emerged genes for *O. sativa*. Most of them (77%) are also very young, exclusively detected within *O. sativa*, despite the close proximity of its nearest neighbors, which are <1.5 million years distant (Fig. 3). In fact, this result must be interpreted with caution as it does not necessarily imply a higher propensity for de novo gene birth in *O. sativa* compared with the other species, and may result from methodological reasons. To classify a gene as de novo emerged, DENSE needs to detect, for each de novo gene candidate, a noncoding hit in the syntenic noncoding region of an outgroup species (Fig. 1d). This becomes more challenging, as the distance between the focal and the outgroup species increases. Indeed, noncoding regions evolve fast, and the detection of homology relationships with such fast-evolving sequences is expected to decrease rapidly with the evolutionary distance separating the focal and the screened neighbors. Consequently, our ability to confidently support a TRG as a de novo emerged gene is bounded by our capacity to detect its orthologous noncoding hit(s) in the outgroup species, which, in turn, is directly limited by the distance of the latter. Precisely, *O. sativa* has 7 neighbors with <1.5 million years of divergence, which probably facilitates the detection of outgroup noncoding hits. In contrast, the closest neighbors of the other focal species (except *Arabidopsis thaliana*) are >4 million years distant, according to TimeTree (Kumar et al. 2022). However, it is noteworthy that geological time is not well suited for fair comparisons between such diverse species. Each species is characterized by different generation times and may evolve at distinct evolutionary rates.

Accordingly, we sought to compare, for each focal species, our ability to detect the homology traces of its noncoding ORFs across its neighbors by establishing what we call a noncoding segments' detectability profile. We, therefore, randomly selected subsets of 1,000 noncoding

Table 1 DENSE predictions for the seven studied species: for each species, the number of genes, the phylostratum threshold used for TRG detection, the number of predicted TRGs, and the number of de novo emerged genes predicted with Strategy 2, Strategy 1, or Strategy combined with synteny criterion are indicated

Organism	# Genes	TRG node	# TRGs	# De novo strat 2	# De novo strat 1	# De novo strat 1 + synteny
S. cerevisiae	5,997	Saccharomyces	406	296	269	230
H. sapiens	23,140	Hominidae	287	184	176	89
M. musculus	22,122	Murinae	285	145	128	84
D. melanogaster	13,900	Drosophila	966	516	117	92
O. sativa	34,177	Oryza	5,104	4,298	3,640	2,455
A. thaliana	27,499	Arabidopsis	953	601	511	289
C. elegans	19,984	Caenorhabditis	5,811	927	135	54

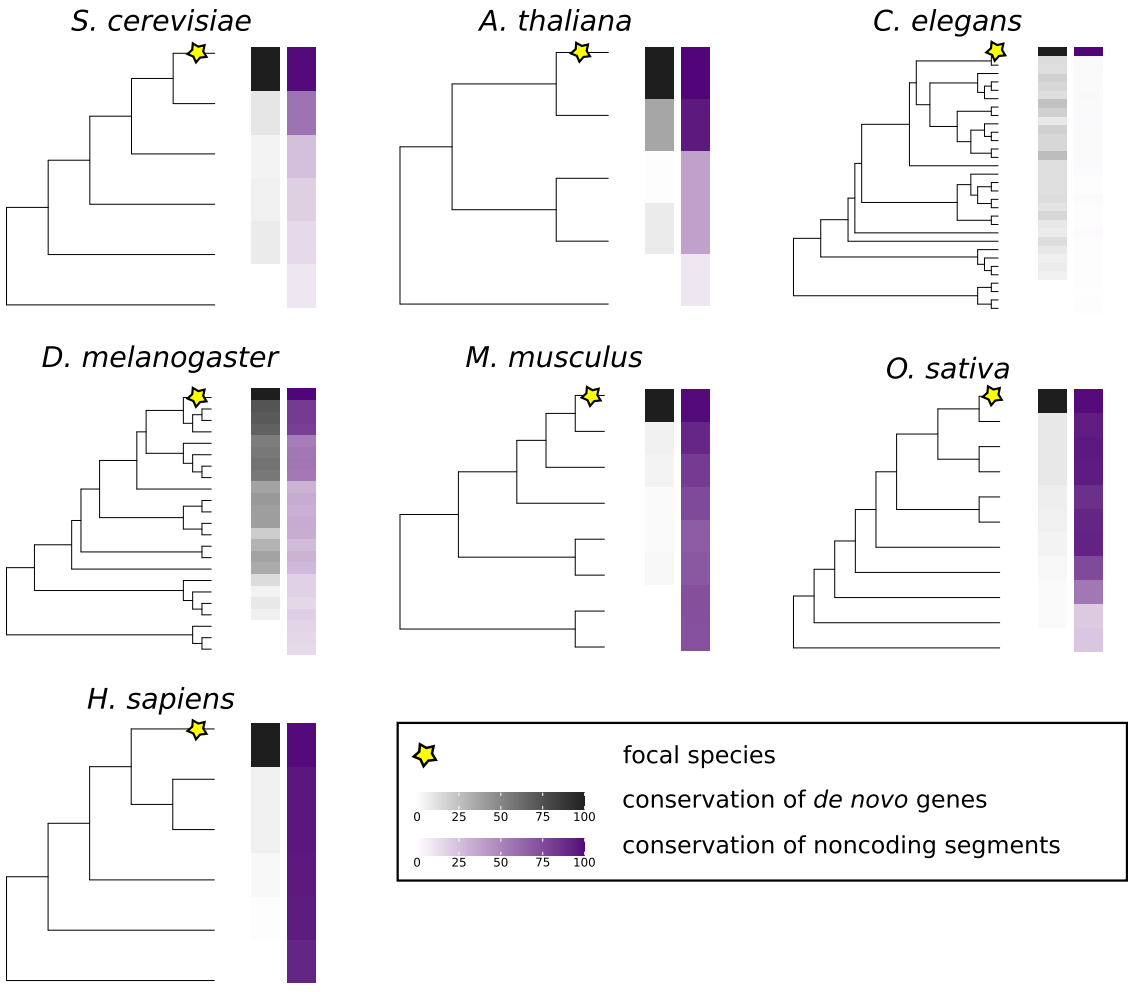

Fig. 3. Conservation profiles of predicted de novo emerged genes and noncoding segments across the focal species's neighbors. The trees are represented for the seven studied species and their neighbors (see supplementary fig. S1, Supplementary Material online for the trees with the names of species). Each tree is associated with two heat maps. The left one (gradient of grays) represents the conservation levels of all the focal de novo emerged genes predicted by DENSE. The color of a given species in the tree represents the fraction of the focal de novo emerged genes that were detected until this species, with dark colors corresponding to species for which an important fraction of the focal de novo genes has been detected in it. The right heat map (gradient of purple) represents the conservation level of a subset of 1,000 noncoding segments selected randomly as calculated for the left heat map.

segments from each focal genome and screened them with tblastn against the noncoding regions of the neighbors. We then calculated the fraction of these segments with a significant tblastn hit within the neighbors (Fig. 3). This served as a proxy to estimate the speed at which the homology signal within neutrally evolving regions disappears among related species, providing us with a detectability profile that reflects the progression of noncoding regions' detectability across the tree. The latter that is calculated between noncoding regions gives a lower bounded estimation of the homology signal that can be expected between de novo emerged genes (i.e. under selection) and their orthologous noncoding regions (i.e. neutral evolution). For *S. cerevisiae*, *Caenorhabditis elegans*, *D. melanogaster*, and *A. thaliana*, the conservation of the predicted de novo emerged genes across the neighbors overall correlates with the detectability profiles of noncoding segments, suggesting that their detection is bounded by the detectability of the noncoding hits within the outgroup species (see supplementary table S2, Supplementary Material online for the corresponding Spearman's correlation coefficients). This opens the question of whether additional young de novo emerged genes are missed in these four species due to the high phylogenetic distance separating the focal and its closest neighbors. In contrast, in the case of *O. sativa*, as many as 76% of the randomly selected noncoding segments still exhibit significant hits to *Ornithoptera meridionalis*, which may explain the higher count of de novo gene candidates predicted for this species (Fig. 3, supplementary fig. S1, Supplementary Material online). To assess the influence of the neighboring species proximity on the detection of de novo emerged genes, we iteratively designated each *O. sativa*'s

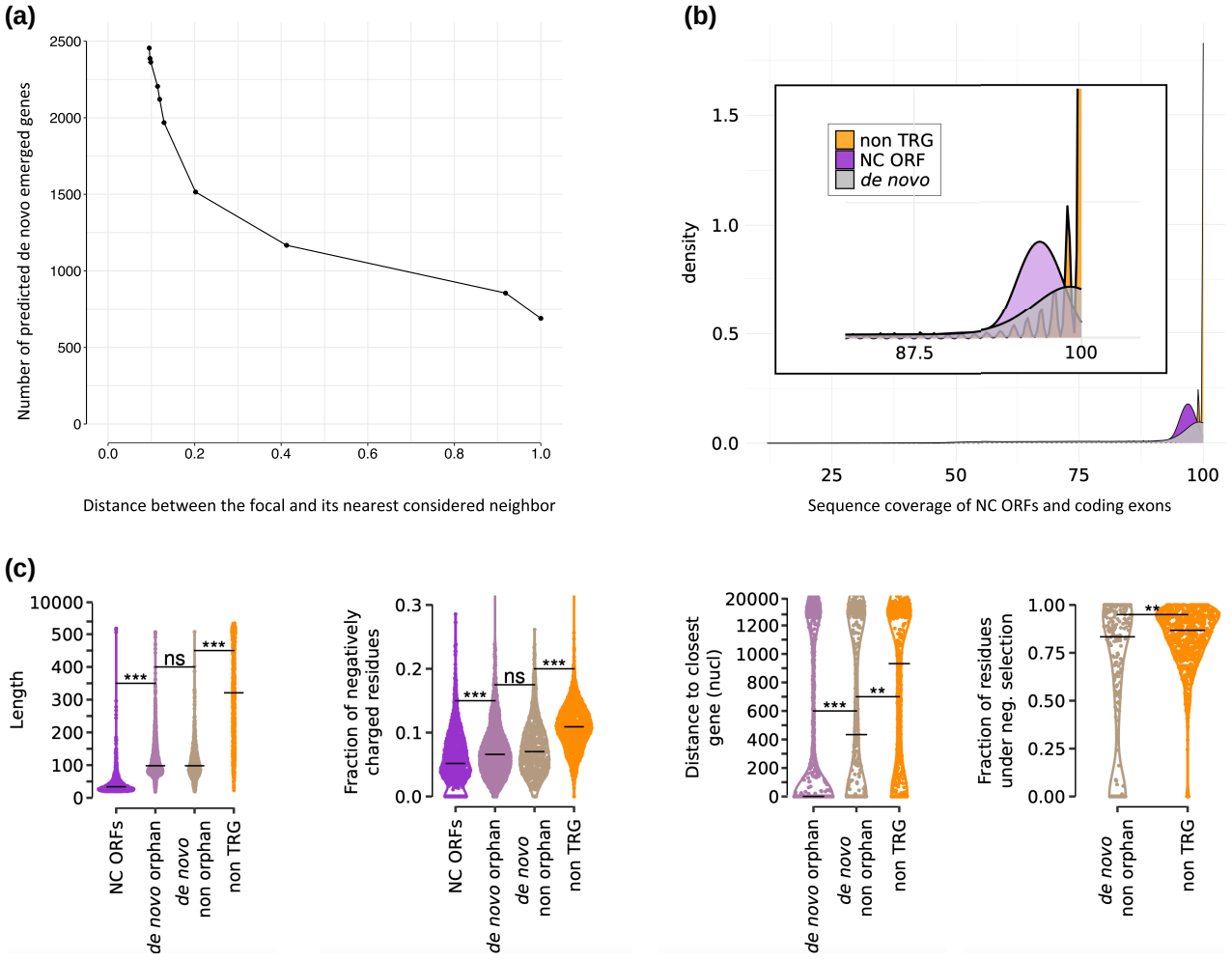

Fig. 4. Analysis of the de novo emerged gene candidates of *O. sativa*. a) Number of de novo emerged genes that are predicted by DENSE (Strategy 1 with synteny criterion) with respect to the phylogenetic distance between the focal species and the nearest considered neighbor. The phylogenetic distance was calculated by OrthoFinder (Emms and Kelly 2019). b) Distribution of the intactness of 1,000 non-TRGs (orange), 1,000 noncoding ORFs (purple), and the orphan de novo emerged genes (gray) of *O. sativa* across its neighboring species. The intactness represents the percentage of a noncoding ORF or coding sequence exon that is aligned with a homologous sequence detected in the neighbors (i.e. alignment coverage). c) Several properties calculated for 1,000 noncoding ORFs and 1,000 non-TRGs (purple and orange, respectively), and all the de novo genes of *O. sativa* separated in orphans and nonorphans. From left to right: ORF length (in amino acids), fraction of negatively charged residues, distance to the closest neighboring gene (in nucleotides), and fraction of residues undergoing negative selection. *P*-values were computed with the Mann–Whitney *U* test (one-sided). Asterisks denote level of significance: *, **, *** for $P < 1 \times 10^{-1}$, 1×10^{-2}, 1×10^{-3}, respectively.

neighboring species as its new closest neighbor by systematically excluding neighbors that were phylogenetically closer than the newly assigned nearest neighbor (Fig. 4a). By doing so, we effectively demonstrated that the prediction of de novo emerged genes in *O. sativa* is significantly impacted by the choice of the closest neighbor. The number of predicted de novo emerged genes diminishes rapidly with the distance to the closest considered neighbor. Notably, a sharp decrease in the number of predicted de novo emerged genes is observed when *Oryza punctata* is considered the nearest neighbor, aligning precisely with the lineage associated with a noticeable decline in the detectability of noncoding ORFs. This trend holds throughout the tree, as shown by the strong correlation between the number of predicted de novo emerged genes and the distance with the closest neighbor (Spearman's correlation coefficient: Rho $= -1$, $P = 1.7 \times 10^{-61}$). This result further strengthens the importance of the proximity of the neighbors considered in the prediction of de novo emerged genes, again supporting the hypothesis that other young de novo gene candidates may be missed for species whose neighbors are too distant. This also supports the utility of noncoding segments' detectability profiles as a valuable proxy for estimating our ability to identify de novo emerged genes.

Surprisingly, in the cases of *M. musculus* and *H. sapiens*, the number of predicted de novo emerged genes is very

low, despite the high noncoding segments' detectability across their neighbors. In fact, the number of predicted de novo emerged genes is also limited by the count of detected TRGs, which is quite low for these two species compared with the others (see Table 1). However, whether this low number results from biological factors or from the phylostratum threshold chosen for these species (e.g. too young) remains unknown. Furthermore, in *O. sativa* and the two mammals, most of their de novo gene candidates (*O. sativa*: 77%, *M. musculus* 87%, *H. sapiens* 87%) are orphans, though, according to their noncoding segments' detectability profiles, one might anticipate higher levels of de novo genes being shared by their closest neighbors. Yet, the fractions of de novo emerged genes conserved beyond the focal species remain very low (23%, 13%, and 13% for *O. sativa*, *M. musculus*, and *H. sapiens*, respectively). We can hypothesize that these genes emerged earlier and were subsequently lost in the sister lineages, recalling the high turnover of novel ORFs reported by Grandchamp et al. (2023); alternatively, they may reflect very young genes that have emerged recently in the lineage of the focal species, or they may simply result from annotation bias. Model organisms may have been overannotated with respect to their neighbors, as is the case, for example, with the manual annotation of yeast available in the Saccharomyces Genome Database (Cherry et al. 2012). If so, the young genes detected in the focal species may also be present in the neighbors while not annotated as coding. However, one should note that for these three species, the genomes of the focal and the neighbors were annotated using the same pipeline (Ensembl 53 for *O. sativa* and NCBI for *M. musculus* and *H. sapiens*), indicating that the annotation bias, if any, is likely to be the same for all the compared genomes. In the following section, we chose to focus on the case of *O. sativa*, which harbors a significant number of de novo emerged genes for meaningful statistical analyses. To explore the possibility that the high number of de novo gene candidates specific to *O. sativa* could be attributed to potential overannotation, we investigated whether their corresponding exons were intact in the neighboring species by calculating the proportion of each exon that was aligned with its homologous hits (i.e. exon coverage). A high level of exon intactness would suggest the presence of older genes that might have been overlooked when annotating the neighboring genomes. As a control, we repeated the experiment for a subset of 1,000 non-TRGs (i.e. old CDSs) and 1,000 noncoding ORFs randomly extracted from *O. sativa*. These subsets enabled us to estimate the conservation levels expected for established CDSs and noncoding ORFs, respectively, over this timescale. Firstly, it is worth noting that the exons of non-TRGs, on average, are retrieved in 85% of the screened *Oryza* species, while the noncoding ORFs and the exons of the orphan de novo emerged genes are less conserved, being detected in 70% of the genomes. Figure 4b represents the distribution of the alignment coverage of *O. sativa*'s noncoding ORFs and that of the exons of its non-TRGs and de novo genes when aligned with their homologs detected in the sister lineages (i.e. intactness). We show that the distribution of the intactness of de novo orphan exons lies between those of the noncoding ORFs and the non-TRG exons. On average, the coverage between the aligned homologous exons is significantly lower for de novo emerged genes than for older CDSs (Mann–Whitney *U* test [one-sided], $P = 1.6 \times 10^{-29}$), supporting the hypothesis that these young genes do not overall correspond to older genes missed by the annotation pipeline of *O. sativa*'s neighbors. As an additional control, we used DENSE to search for the de novo genes of the three closest neighbors of *O. sativa*, which share the same topological position within the tree. By doing so, we detected comparable, albeit lower, numbers of de novo emerged genes (1,451, 1,762, and 2,021 for *Oryza indica*, *Oryza nivara*, and *Oryza rufipogon*, respectively), indicating that *Oryza* species, overall, exhibit higher numbers of de novo emerged genes. However, it remains difficult to disentangle the effective contribution of the close proximity of *Oryza* species from that of potential biological factors that could lead to high levels of de novo emerged genes. As observed for *O. sativa*, the majority of the de novo emerged genes of *O. indica* are orphans (59%; supplementary fig. S2, Supplementary Material online). *Oryza nivara* and *O. rufipogon,* which are very close (about 700K years, according to TimeTree [Kumar et al. 2022]), share on average 44% of their de novo emerged genes and are respectively associated with 38% and 36% of orphan de novo genes. These results highlight that despite the close proximity of these species, the latter are nevertheless associated with high levels of de novo emerged genes specific to closely related lineages. Although the evolutionary fate of these genes is unpredictable, their remarkably low conservation levels suggest that a significant fraction of them may be short lived in evolutionary history.

This prompted us to ask whether these ORFs are indeed coding and do not indicate segregating protogenes or noncoding ORFs erroneously annotated as coding. Indeed, it is worth noting that DENSE relies on genome annotation, assuming all genomic elements annotated as CDS as coding. However, depending on the annotation criteria, annotated CDSs may include young de novo emerged genes that might be short lived in evolutionary history but also nonfunctional protogenes that might never reach fixation. Discriminating between these two categories is complex, as it involves discretizing a population that is, in fact, continuous. Notably, it involves a twilight zone where ORFs are neither coding nor silent, providing an entry point into the coding world. One may simply divide the noncoding and coding worlds from a functional perspective

as "to be functional or not functional." However, this appealing dichotomy is not that simple, as it implies a clear definition of the concept of function and the ability to reliably assess whether an ORF is functional (Doolittle et al. 2014; Doolittle 2018; Keeling et al. 2019). Nevertheless, the scientific community has identified a set of features that are typical hallmarks of functional sequences and that turned out to be efficient in characterizing young ORFs (Carvunis et al. 2012; McLysaght and Hurst 2016; Couso and Patraquim 2017; Papadopoulos et al. 2021; Peng and Zhao 2024). Figure 4c represents for *O. sativa*, several of these properties calculated for a subset of 1,000 noncoding ORFs, 1,000 non-TRGs, and all predicted de novo emerged genes classified as orphans or nonorphans. The ORF length has long been recognized as a key feature for detecting coding ORFs, with annotated CDSs being longer than noncoding ORFs regardless of the considered species. In *O. sativa*, genes predicted as de novo emerged, irrespective of being orphan or not, generally exhibit intermediate sizes between noncoding ORFs and old CDSs. It is important to note, however, that the ORF size is a feature that is explicitly taken into account in classical annotation pipelines, potentially introducing bias in the size of these young ORFs, since smaller ones may have been automatically excluded during the annotation process. Previously, we and others have demonstrated that CDSs in yeast and fly were enriched in negatively charged residues compared with noncoding ORFs, likely contributing to translation efficiency and/or preventing promiscuous interactions with the highly abundant and negatively charged ribosomes (Couso and Patraquim 2017; Papadopoulos et al. 2021). This feature, therefore, offers a good proxy to interrogate the codability of subsets of candidates. Precisely, Fig. 4c shows that the de novo gene candidates, including the orphan ones, display distributions of negatively charged residue fractions that lie between those of noncoding ORFs and old CDSs. This observation underscores that these candidates do not resemble noncoding ORFs and provides support for their classification as young, recently de novo emerged genes. While not applicable to the orphan de novo gene candidates, the calculation of the d*n*/d*s* ratio for the nonorphan ones further substantiates this hypothesis with older de novo gene candidates exhibiting >62% of their residues under negative selection, a fraction comparable with that observed for older CDSs (see the Materials and methods section for more details). Orphan candidates display significantly closer proximity to their surrounding genes than older de novo emerged genes and canonical CDSs. The genomic environment of nearby genes may contribute to their expression through transcription read-through, bidirectional promoter activity, or pervasive expression of regions of open chromatin (Gotea et al. 2013; Wu and Sharp 2013; McLysaght and Hurst 2016). Finally, while we cannot exclude that the orphan de novo emerged genes may include nonfunctional segregating protogenes, their similarity to their nonorphan counterparts suggests a homogeneous population of young genes whose future trajectory remains uncertain.

All these findings highlight the importance of having close neighbors in the detection of de novo emerged genes. Additionally, our analysis suggests that using very close species (i.e. associated with high noncoding segments' detectability) may unveil populations of very young ORFs specific to the lineage of the focal species. Although the young ORFs specific to *O. sativa* overall resemble the older de novo gene candidates, the fact that most of them are not retrieved in the neighboring species (see supplementary fig. S2, Supplementary Material online) suggests a rapid turnover of de novo emerged ORFs, at least in *Oryza* species, whose majority may be short lived in evolutionary history, though the fate of each individual ORFs remains currently unpredictable.

Application of DENSE to Short Timescales

In this section, we illustrate an application of DENSE to short timescales, which can be very useful when one aims to investigate the distribution and/or conservation of de novo emerged ORFs in a population. DENSE can handle genomes of different strains or lines, thereby enabling the characterization of the emergence of novel ORFs during very short timescales. Here, we sought to assess the conservation of the de novo gene candidates identified in the reference line of *D. melanogaster*, in six other lines sequenced and annotated in Grandchamp et al. (2023). Therefore, we started from the TRGs of the reference line detected previously with the default parameters of DENSE (see Table 1), and directly entered the DENSE pipeline at the filtering step (Fig. 1c). In this configuration, the focal is the reference line, and the neighbors consist of seven genomes of the *Drosophila* genus and those of the other six *D. melanogaster* lines (see the list of fly lines in supplementary table S3, Supplementary Material online). As in the previous section, genes are considered as de novo emerged according to Strategy 1 combined with the synteny criterion. Figure 5a shows that all the de novo emerged genes predicted for the reference line are present in the other six fly lines. This result contrasts with the observation made by Grandchamp et al. (2023), where most newly expressed ORFs (i.e. neORFs that consist of not-yet fixed precursors of de novo genes) are generally observed in a single line, supporting a high birth-death rate (Fig. 5b). Figure 5c to e represents the size, the fraction of negatively charged residues, and the genomic distance to the closest gene of these two ORF categories, along with those of a subset of 1,000 noncoding ORFs. The neORFs and the de novo gene candidates display comparable size, being significantly longer than noncoding ORFs (Mann–Whitney *U* test

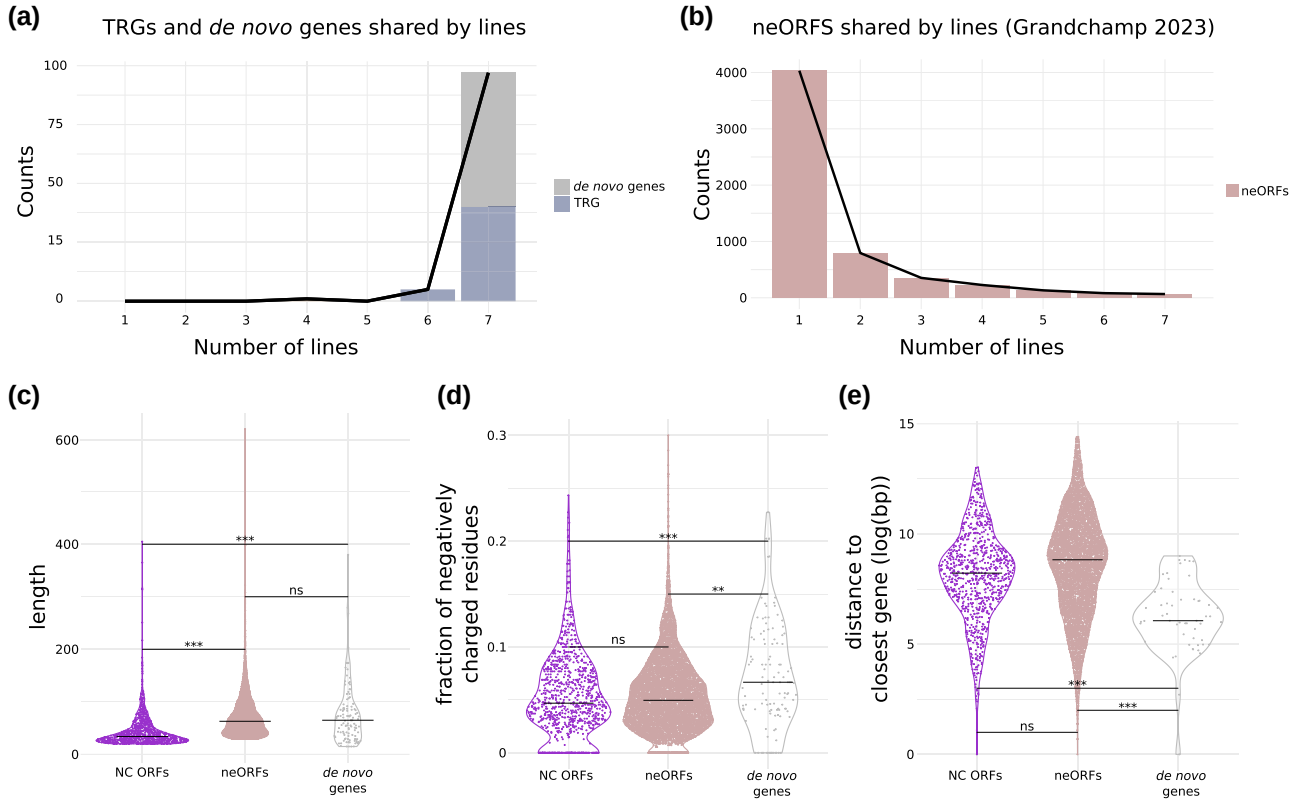

Fig. 5. Conservation and properties of de novo emerged genes and neORFs of *D. melanogaster*. a) Distribution of the conservation levels of the de novo emerged genes predicted for *D. melanogaster* with DENSE (Strategy 1 with synteny) across the seven fly lines. b) Same as (a) for the neORFs of *D. melanogaster* detected in Grandchamp et al. 2023. c) ORF length for 1,000 randomly selected noncoding ORFs of *D. melanogaster*, its neORFs, and de novo emerged genes. d) Same as (c) for the distributions of negatively charged residue fractions. e) Same as (c) for the distance to the closest gene. Asterisks denote level of significance: *, **, *** for $P < 1 \times 10^{-1}$, 1×10^{-2}, 1×10^{-3}, respectively.

[one-sided], both $P < 5.5 \times 10^{-11}$). However, the fraction of negatively charged residues of the neORFs is similar to that of noncoding ORFs and lower than that of the de novo emerged genes. This suggests two distinct populations: an older population of ORFs conserved across *D. melanogaster* lines and whose amino acid content is undergoing optimization, and a younger population of potential gene precursors characterized by high turnover. Furthermore, the fact that young de novo genes are generally located closer to old genes than neORFs suggests that the ORFs present in the gene vicinity may benefit from a favorable genomic environment, which could facilitate their "fixation," at least across *D. melanogaster* lines.

Impact of Parameters

We then investigated the impact of the different filters that can be applied to TRGs to classify them as de novo emerged. Unfortunately, no benchmark for de novo emerged genes is available, hampering the systematic study of parameter effects on the predictions. Furthermore, as de novo emerged genes get older, they become more challenging to detect since the criteria to classify them as de novo emerged (e.g. microsynteny, hit in noncoding regions of outgroup species) become more difficult to satisfy. As a result, older de novo emerged genes are likely to no longer conform to the constraints imposed by classical de novo gene detection protocols, and assessing whether individual candidates are true or false positives becomes a nontrivial task. Therefore, we deliberately removed some of the filters proposed by DENSE and analyzed the aforementioned properties of the resulting candidates, in order to see whether the latter are associated with properties similar to those of confidently identified candidates, i.e. with the complete set of filters as shown with the protocol presented in Fig. 1 (Strategy 1 with synteny criterion).

Removing the synteny filter adds, on average, 67% of de novo emerged candidates for each focal species (see Table 1). Generally, the candidates identified without synteny validation exhibit properties similar to those with a noncoding hit found in synteny, especially considering their fractions of negatively charged residues or positions undergoing negative selection (Fig. 6). In several species, they are, however, associated with longer size, higher distance to the

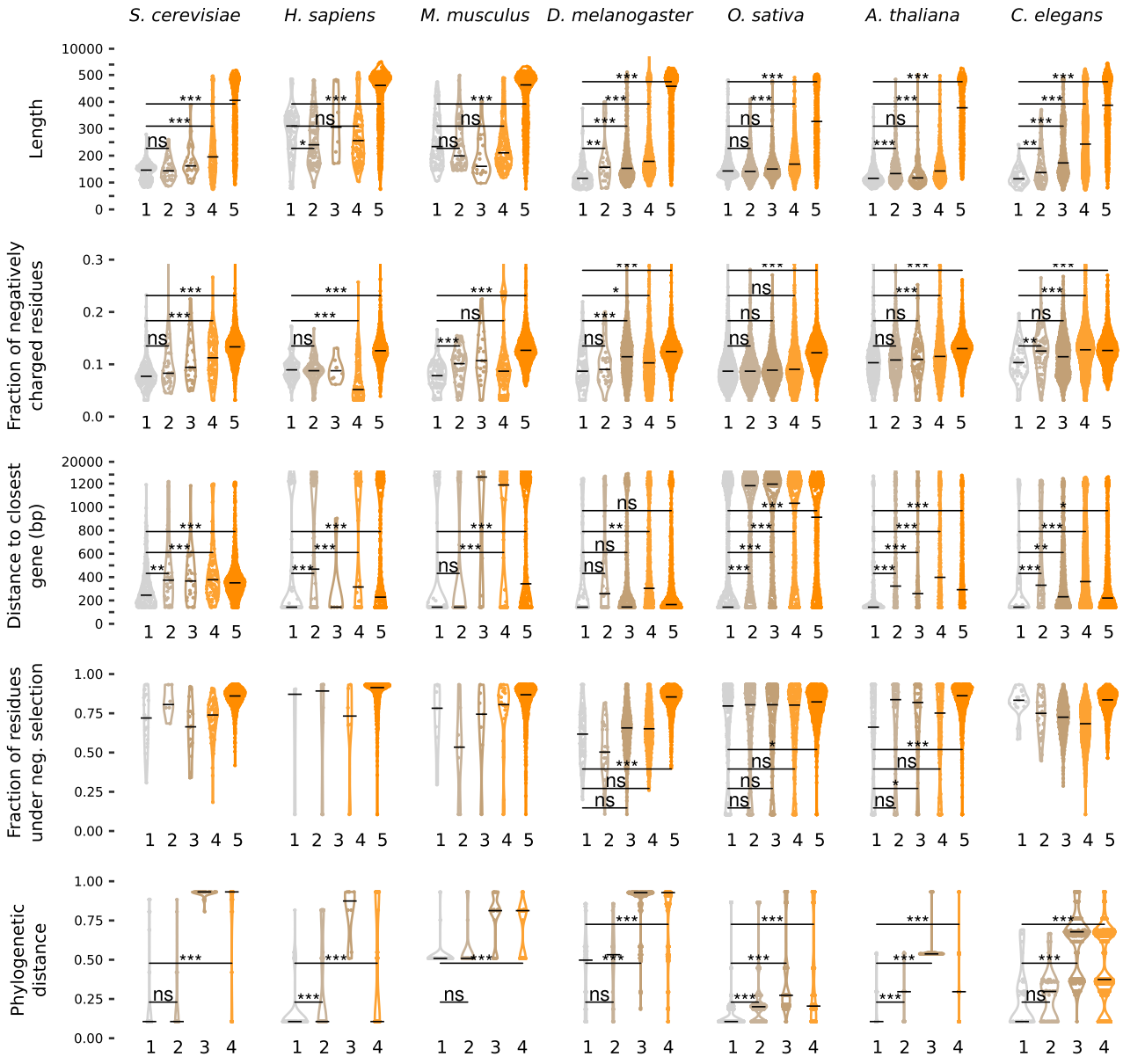

Fig. 6. Properties of candidates detected with different combinations of criteria for each species. For all properties except Phylogenetic distance, from left to right: (1) de novo emerged gene predicted with at least one syntenic noncoding hit in at least one outgroup (Strategy 1 with synteny); (2) de novo emerged gene candidates added when removing the synteny filter (Strategy 1 without synteny); (3) de novo emerged gene candidates added when removing the requirement of the noncoding hit to be in an outgroup species (Strategy 2 without synteny); (4) remaining TRGs; (5) and non-TRGs. Non-TRGs are absent from the Phylogenetic distance property. From top to bottom: length of the corresponding CDSs (in amino acids). Fraction of negatively charged residues. Distance to the closest neighboring gene (in base pairs). Fraction of positions under negative selection. Phylogenetic distance calculated with OrthoFinder between the focal species and the farthest species where the gene has been detected. *P*-values were computed with the Mann–Whitney *U* test (one-sided). Asterisks denote level of significance: *, **, *** for $P < 1 \times 10^{-1}$, 1×10^{-2}, 1×10^{-3}, respectively.

closest gene, and higher phylogenetic distance. We may hypothesize that some of these candidates correspond to de novo emerged genes, whose synteny with their orthologous noncoding region has been lost over time. In fact, although their phylogenetic distance with their farthest homolog is not significantly higher in four of the seven studied species, their conservation profiles reveal higher proportions of conserved de novo genes (supplementary fig. S1, Supplementary Material online). This observation further supports the hypothesis of a population including slightly older de novo emerged genes whose genomic regions are no longer in synteny.

We then evaluated the impact of the number of outgroups in which a noncoding hit is required to validate a

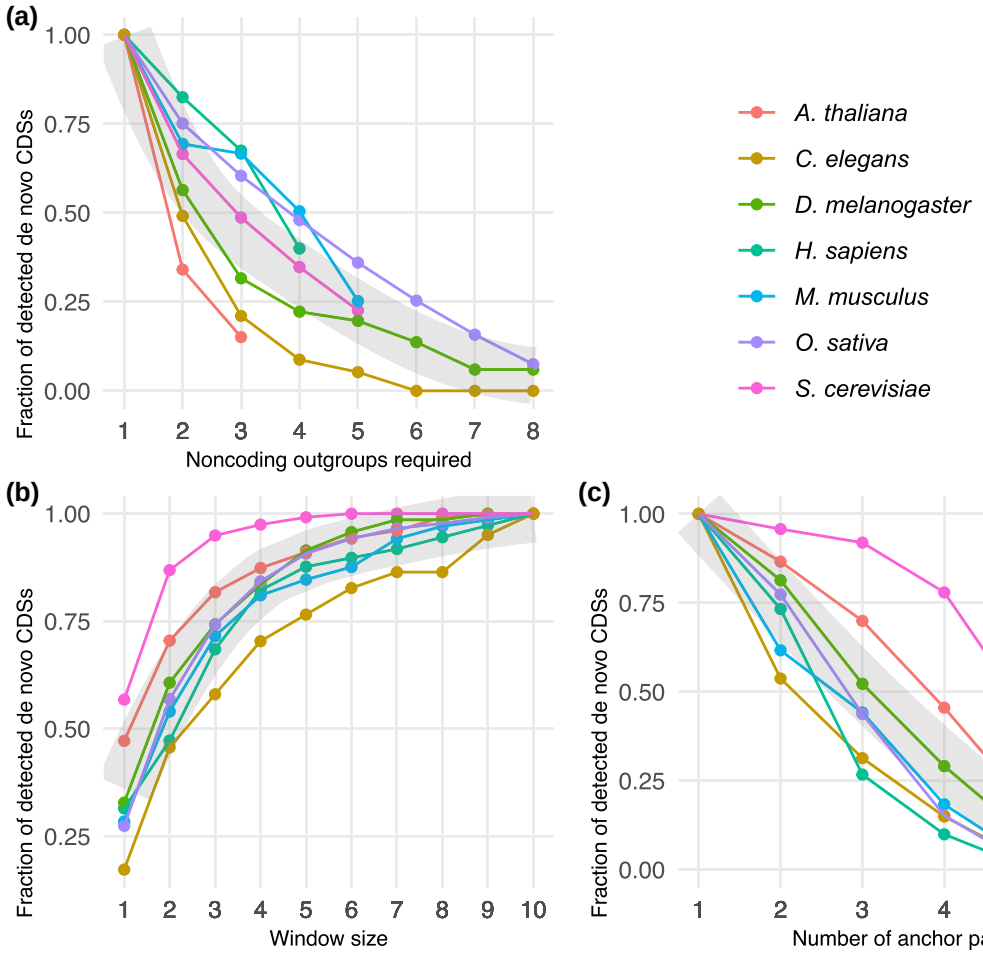

Fig. 7. Impact of the number of outgroups with noncoding hits, the window size, and the number of anchors used for the synteny search on the number of predicted de novo CDSs. a) Fraction of predicted de novo CDSs with respect to the number of outgroups with a noncoding hit(s). The fraction is calculated relative to the number of de novo CDSs predicted when requiring only one outgroup with a noncoding hit. b) Fraction of predicted de novo CDSs with respect to the size of the window (i.e. number of considered genes on each side of the focal de novo gene candidate). The fraction is calculated according to the number of de novo CDSs predicted with a size window of 10. The number of anchor pairs (i.e. number of anchors per window) has been set to one. c) Fraction of predicted de novo CDSs with respect to the number of required anchor pairs. The size of the window has been set to six. The fraction is calculated relative to the number of de novo CDSs predicted with a number of anchors set to one.

de novo gene candidate. Figure 7a shows that increasing this number significantly reduces the number of de novo candidates in *A. thaliana* and *C. elegans*. It is worth noting that the phylogenetic tree of *A. thaliana* includes a maximum of three potential outgroups, which rapidly limits the detection of de novo genes when the number of required outgroups with noncoding hits increases. In contrast, *C. elegans* has many potential outgroups but exhibits a very low detectability profile, challenging the identification of noncoding hits in multiple outgroups. Interestingly, the effect is significantly less pronounced for *H. sapiens*, *M. musculus*, and *O. sativa*, where over 75% of the de novo genes detected when requiring only one outgroup with a noncoding hit, are still identified when requiring noncoding hits in two outgroups. These species are characterized by high detectability profiles, which likely facilitate the detection of noncoding hits across the phylogenetic tree. Overall, the de novo candidates detected using different outgroup thresholds are comparable yet distinguishable from canonical CDSs, revealing a relatively homogeneous population (supplementary fig. S3, Supplementary Material online). This result suggests that, while requiring higher numbers of outgroups with noncoding hits can enhance confidence in the noncoding status of the ancestor and, ultimately, in de novo gene detection, it might also be too restrictive, particularly for species associated with low detectability profiles.

Candidates identified using Strategy 2 (i.e. whose noncoding hit(s) are not required to be in an outgroup species, Fig. 2a) without synteny display a more distinct separation from those identified with Strategy 1. They generally exhibit wider distributions than the de novo genes predicted with

high confidence, unveiling a heterogeneous population of genes with various ages and probably different origins (Fig. 6). These genes may include fast-evolving genes whose orthologs would have been lost in some neighboring species, explaining the detection of a noncoding hit in the noncoding regions of some neighbors. However, we do not exclude that it also comprises genes that emerged de novo million years ago, and for which homology in the noncoding orthologous regions of the outgroup species is no longer detectable. Finally, as expected, TRGs consist of a highly heterogeneous population of genes with wider distributions for all considered features, reinforcing the idea that TRGs encompass genes with various origins and ages. It is interesting to note, that as filters are removed, the phylogenetic distance of the detected candidates increases, probably reflecting false positives with different ages, but also older de novo emerged genes that no longer satisfy the filter criteria.

Ultimately, we studied the influence of two parameters of the synteny filter on the number of validated de novo gene candidates: (i) the size of the window of genes considered on each side of the focal de novo gene candidate and of the noncoding hit in the outgroup species, and (ii) the number of anchor pairs required in the vicinity of the noncoding hit (Fig. 2c). In Fig. 7b, a candidate is classified as de novo emerged, if at least one anchor pair is found in the vicinity of its homologous noncoding hit in an outgroup species. This vicinity is defined by a window size that varies from one to ten genes (see Fig. 2c for more details). For all focal species, a window of four genes enables the detection of >84% of the candidates detected with the largest window. This proportion significantly drops with smaller windows, reaching only 35% of the candidates detected with a window of one. Conversely, beyond four genes, the proportion of validated candidates increases gradually with the size of the window, suggesting that a window of four genes offers a suitable compromise. Figure 7c illustrates the impact of the number of anchor pairs, using a window of six, on the number of validated candidates. As expected, this number sharply diminishes with the number of anchor pairs, regardless of the focal species, reflecting the high speed at which the synteny is altered in related genomes.

DENSE Online

The filtering part of the DENSE workflow, including all criteria combinations (Fig. 1c and d) is available through a web server once users provide their own list of TRGs (https://bioi2.i2bc.paris-saclay.fr/django/denovodb/dense-run/). Since calculating the phylostratigraphy can be highly time-consuming, and as many studies involve the same model organisms, we also provide the DENSE calculations for the seven model organisms of the present study through a public database (https://bioi2.i2bc.paris-saclay.fr/django/denovodb/). These calculations include the phylostratigraphy calculated from the nr database (downloaded on 2022 March 23) and the predictions obtained from most of the DENSE available combinations of criteria.

Discussion

In this study, we investigated the influence of different parameters in de novo emerged gene detection. Notably, we demonstrated the significant impact of the phylogenetic distance separating the focal species and its closest neighbors on the ability to detect de novo emerged genes, with the homology signal in noncoding regions decreasing rapidly as this distance increases. The decrease in the number of predicted candidates in *O. sativa* from 2,455 to 1,515, upon removing neighboring species with <1.5 million years of divergence, raises questions about the existence of a population of young candidates being missed in the other focal species. Additionally, we showed that the distance between the focal species and its nearest neighbors impacts the nature of the detected candidates. Indeed, using very close neighbors enables the detection of very young de novo emerged genes, such as the 1,881 candidates specific to *O. sativa*. The latter, while displaying similar properties to the older de novo emerged genes, may not share the same fate. In particular, an important fraction of them might have a limited lifespan in evolutionary history, as suggested by the low amounts of de novo emerged genes shared across closely related neighbors. Furthermore, we found that removing the synteny filter leads to a significant increase in the number of predicted de novo emerged genes. These candidates, overall, resemble those detected with a high degree of confidence. It can be, therefore, hypothesized that the latter include a population of de novo emerged genes whose synteny has been lost over time, being subsequently excluded with the strictest combination of criteria. Finally, omitting the requirement of a noncoding hit in an outgroup species led to a heterogeneous population of candidates with properties slightly different from those of candidates detected with high confidence. This population may encompass de novo emerged genes but also TRGs with other origins, underlining the difficulty of finding the set of parameters that minimizes the number of false positives and false negatives in the context of de novo emerged gene prediction.

In fact, for each species, there is a specific window of time during which de novo emerged gene candidates can be identified confidently and beyond which, the signals used for reliable prediction are no longer detectable. As de novo emerged genes get older, their genetic environment is likely to have evolved, and homology traces within orthologous noncoding regions are no longer detectable. Determining whether their ancestor was noncoding,

hence, becomes challenging, and classifying them as de novo emerged rapidly becomes impossible. Although these older de novo genes are challenging to detect, they remain too young to resemble old, canonical CDSs, and overall exhibit intermediate properties between recently emerged de novo genes and canonical CDSs. These properties could have been helpful in identifying them. However, it would require being able to discriminate them from pseudogenizing genes or fast-evolving duplicated genes, which, unfortunately, are also expected to exhibit intermediate properties (Vakirlis et al. 2020; Montañés et al. 2023). Determining the noncoding status of the ancestor is, therefore, a key and methods based on ancestral sequence reconstruction appear as a promising approach to distinguish emerging genes from those undergoing free-fall evolution (Vakirlis et al. 2024). While requiring noncoding hits in multiple outgroup species may increase confidence in the noncoding status of the ancestor, young de novo genes are expected to be associated with high turnover, and pseudogenes may be resurrected. These potential back-and-forth events reveal a complex and intertwined evolutionary landscape, further complicating the detection of de novo emerged genes, and again underscoring the challenges of identifying them beyond a specific time window. The time window for the efficient detection of de novo emerged genes is not universal, and determining the right one is not trivial. Species evolution is not linear, and species are associated with different generation times. In other words, species have their own evolutionary time. In this work, we showed that characterizing the conservation profiles or intactness of both CDSs and noncoding segments across the tree offers a valuable route to delineating the boundaries of the evolutionary signal resulting from either selection (i.e. CDSs under selective pressure) or neutral evolution (i.e. noncoding segments). These profiles provide useful landmarks to estimate the upper and lower bounds of intact ORFs or detectable evolutionary traces expected in neighboring species associated with different divergence times, regardless of the knowledge of the generation time and evolutionary history of the considered species. Consequently, they can assist the user in adjusting the list of neighboring species or in identifying configurations where the nearest neighbors are too divergent for the accurate detection of de novo emerged genes.

It should be noted that the number of de novo emerged genes that can be detected is also bounded by the number of TRGs, since, by definition, the number of predicted de novo emerged genes cannot exceed the count of TRGs. This number is, in turn, directly affected by the taxonomic sampling around the focal species, the phylostratum used to define the TRGs, the heterogeneity in genome annotations across neighbors, and the sensitivity of the homology search when dating a species' genome. Undersampling is likely to underestimate gene age by lacking the evolutionary relays that could connect them to their homologs in remote species, thus potentially leading to the misclassification of old genes as TRGs. In addition, the sensitivity of the homology search during the phylostratigraphy stage may impact the number of detected TRGs, as a lack of sensitivity can lead to underestimating gene age (Domazet-Lošo et al. 2017). On the other hand, heterogeneity in genome annotation among the compared species is also expected to lead to gene age underestimation by incorrectly categorizing genes that have been overlooked in other genomes as orphans. However, we showed that the characterization of the intactness of gene ORFs constitutes an efficient proxy to control this potential bias. It is also important to mention that genome annotations for the same species may exhibit significant disparities, especially for young genes, which are difficult to annotate. Notably, supplementary fig. S4, Supplementary Material online shows the number of *O. sativa* CDSs from the Ensembl Genomes 53 annotation (Cunningham et al. 2022) that are also present in the annotation published by Stein et al. (2018). While most CDSs are shared between the two annotated proteomes, the majority of those belonging to the youngest phylostrata are not, likely contributing to variations in de novo gene lists across different studies. Finally, choosing a phylostratum for the definition of TRGs that is too young inherently results in reduced lists of TRGs, causing users to miss the genes whose emergence predates this phylostratum. If the genus level appears to be an effective phylostratum for most species in this study, the phylostratum threshold for *H. sapiens*, and *M. musculus* had to be adjusted to *Hominidae* and *Murinae*, respectively, to get sufficient numbers of outgroup species. Again, the conservation profiles of a subset of noncoding segments across the species associated with a given phylostratum can help users define the appropriate threshold. In any case, although defining the right phylostratum seems nontrivial in theory and directly depends on the species under consideration, in practice, false positives in TRG detection should be eliminated through the requirement of homology traces in the orthologous syntenic noncoding region of outgroup species. This strict combination of criteria strongly supports the noncoding status of the ancestor. However, we do not exclude the possibility that a small fraction of genes that meet these criteria may consist of fast-evolving genes, which would have been recently lost in the sister lineages, thereby explaining the homology trace(s) detected in these species.

Finally, the main difficulty may stem from the term "de novo" genes itself, which, in fact, refers to the mechanism by which these genes have emerged. This semantic confusion, conflating the process with the product, may inaccurately impart that de novo emerged genes are uniformly young and constitute a single and cohesive gene category, whereas, in reality, the population of de novo emerged genes is continuous and heterogeneous. It encompasses

recently emerged genes but also, genes that appeared very early during evolution and whose origin is now unpredictable. This continuum implies a wide diversity of properties, functions, and trajectories, with recently emerged genes probably associated with uncertain fates, as suggested by the numerous young genes detected in *O. sativa*. In contrast, older de novo emerged genes have diversified over time, forging their unique trajectory during evolution. Consequently, all these genes, in all their diversity, share no more in common than their mechanism of emergence, and ultimately, their origin. Thus, which de novo emerged genes are we looking for? If the question is "How do we pass from the noncoding to the coding world," part of the answer lies in the transition between these two worlds; specifically, in the study of genes that still bear the footprints of this transition. Precisely, the combination of criteria offered by DENSE is tailored for the identification of these young genes that have recently emerged de novo.

Conclusion

We introduced DENSE, a user-friendly Nextflow pipeline designed to seamlessly execute the entire protocol required for detecting de novo emerged genes from genomic data. This process encompasses the identification of TRGs through phylostratigraphy, along with their filtering according to various combinations of criteria. For higher specificity, we recommend employing the strictest protocol that relies on the filtering of candidates that exhibit homology traces in a syntenic noncoding region of an outgroup species. The latter, applicable genome-wide, stands out as the most promising for confirming the noncoding status of the ancestor using genomic data. It is important to note, however, that DENSE offers different combinations of strategies and parameters, empowering users to adapt to specific situations or explore new combinations.

The filtering step of DENSE is accessible to the scientific community through a web server, should users provide their own list of TRGs. Furthermore, as most studies focus on a limited set of model organisms, we have precalculated phylostratigraphies and executed the different DENSE strategies for the seven model organisms studied in this work. The associated results are available through a requestable database, that is to the best of our knowledge, the first public database of predicted de novo emerged genes. This unique dataset, encompassing seven model organisms and calculated with a consistent protocol, provides the scientific community with a valuable resource for cross-species analyses and large-scale studies. We plan to extend this database to other organisms and hope that it will serve as a reference for de novo emerged gene lists generated with specific combinations of criteria. The integration of DENSE into a fully automated pipeline and the modularity of its framework embedding different strategies and parameters should enable users to establish rational protocols for de novo emerged gene detection. This, in turn, should promote enhanced protocol communication, effective interoperability, and improved reproducibility across studies. While we anticipate that protocols will continue to evolve, we hope that this work, along with the rationality and interoperability facilitated by DENSE, will stimulate fruitful discussions and lead to further enhancements of protocols. Precisely, implemented through a Nextflow pipeline, DENSE is perfectly suited to these collaborative goals.

Materials and Methods

Identification of the De Novo Emerged Genes in the Seven Studied Organisms

All de novo emerged genes were predicted using the full pipeline of DENSE with default parameters (Strategy 1 with synteny, a noncoding hit in at least one outgroup, one anchor pair, and a window size of 4). For TRG detection, we used the nr database downloaded on 2022 March 23. The phylostratum threshold used for TRG detection was set to the genus level, except for *M. musculus* and *H. sapiens*. For these two latter, to ensure a sufficient number of outgroup species, the threshold was extended to *Murinae* and *Hominidae*, respectively. It is worth noting that the neighboring genomes used for detecting coding and noncoding hits during the TRG filtering process (Fig. 1c and d) may not correspond to all genomes included in the phylostratum selected for TRG detection. Notably, only species belonging to the *D. melanogaster* subgroup were considered for the filtering step. The list of all genomes used in this study, along with the links to their sequence and annotation files, are available in supplementary table S1, Supplementary Material online. For each focal species, the associated local trees (focal and neighboring species) were generated using OrthoFinder (Emms and Kelly 2019) with default parameters, except for the "msa" method that was used for gene tree inference ("-M" option).

ORF Properties

Except for the calculation of distance to the closest neighboring gene, all studied properties were computed from the CDSs of the genes, including the de novo genes, orphan de novo genes, nonorphan de novo genes, TRGs, and non-TRGs. In cases where genes were associated with multiple isoforms, the evaluated properties were calculated on all their corresponding CDSs. The phylogenetic distance was directly extracted from the tree computed by OrthoFinder (Emms and Kelly 2019). The fraction of residues under negative selection was calculated with codeml (Yang 2007). Therefore, for each considered gene in the focal species, we searched for its orthologs within the neighboring species using the Reciprocal Best Hits method (*E*-value:

10^{-3}, coverage: 70%). The corresponding CDSs were aligned with MAFFT (Katoh and Standley 2013) and provided to codeml. We employed the model assuming a fixed Ω value along the branches and three states per site (negative, neutral, and positive). The distance to the closest neighboring gene was calculated using BedTools (Quinlan and Hall 2010).

Homology Detection

The conservation of the de novo genes across the neighboring species (Fig. 3) and the intactness of the three ORF categories (Fig. 4b) were assessed using blastp (*E*-value: 10^{-3}, coverage: 50%). The detectability profiles were calculated for each focal species based on a similarity search of 1,000 randomly selected intergenic segments of 300 nucleotides across its neighboring species. The similarity search was conducted using tblastn (*E*-value: 10^{-3}, coverage: 50%).

Statistical Analyses

All statistical tests were performed in R (4.3.2) (R Core Team 2021). When samples were >500 individuals, tests were performed 100 times on random subsets of 500 individuals chosen from the initial sample, and the averaged *P*-value was subsequently calculated.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

The authors thank those associated with the *Caenorhabditis* Genomes Project for prepublication access to genome data. They thank Ambre Baumann and Simon Herman for their assistance in testing the DENSE pipeline.

Author Contributions

Conceptualization: P.R., A.L.; development and investigations: P.R., A.G., C.Q., and A.L.; writing: P.R., A.G., and A.L.; supervision: A.L.

Funding

P.R.'s work was supported by a French government fellowship. A.G.'s work was supported by the Deutsche Forschungsgemeinschaft priority program "Genomic Basis of Evolutionary Innovations" (SPP 2349) BO 2544/20-1.

Conflict of Interest

The authors declare that they have no competing interests.

Data Availability

The complete list of genomes used for this study is provided in the supplementary table S1, Supplementary Material online. All the custom scripts used in this study for the generation of figures, along with the associated data are available as supplemental data files accessible at: http://bim.i2bc.paris-saclay.fr/anne-lopes/data_Roginski_GBE2024/.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990:215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Arendsee Z, Li J, Singh U, Bhandary P, Seetharam A, Wurtele ES. Fagin: synteny-based phylostratigraphy and finer classification of young genes. BMC Bioinformatics. 2019:20(1):440. https://doi.org/10.1186/s12859-019-3023-y.

Arendsee Z, Li J, Singh U, Seetharam A, Dorman K, Wurtele ES. Phylostratr: a framework for phylostratigraphy. Bioinforma Oxf Engl. 2019:35(19):3617–3627. https://doi.org/10.1093/bioinformatics/btz171.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. Progressive *Cactus* is a multiple-genome aligner for the thousand-genome era. Nature. 2020:587(7833):246–251. https://doi.org/10.1038/s41586-020-2871-y.

Barrera-Redondo J, Lotharukpong JS, Drost H-G, Coelho SM. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. Genome Biol. 2023:24(1):54. https://doi.org/10.1186/s13059-023-02895-z.

Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. Uncovering de novo gene birth in yeast using deep transcriptomics. Nat Commun. 2021:12(1):604. https://doi.org/10.1038/s41467-021-20911-3.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021:18(4):366–368. https://doi.org/10.1038/s41592-021-01101-x.

Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, Masel J, Wysocki VH, Cordes MHJ. Foldability of a natural de novo evolved protein. Struct Lond Engl. 2017:25:1687–1696.e4. https://doi.org/10.1016/j.str.2017.09.006.

Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. Genetics. 2008:179(1):487–496. https://doi.org/10.1534/genetics.107.084491.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. Proto-genes and de novo gene birth. Nature. 2012:487(7407):370–374. https://doi.org/10.1038/nature11184.

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020:367(6482):1140–1146. https://doi.org/10.1126/science.aay0262.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. *Saccharomyces* genome database: the genomics resource of budding yeast. Nucleic Acids Res. 2012:40(D1):D700–D705. https://doi.org/10.1093/nar/gkr1029.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al. The ecoresponsive

genome of *Daphnia pulex*. Science. 2011:331(6017):555–561. https://doi.org/10.1126/science.1197761.
Couso J-P, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. 2017:18(9):575–589. https://doi.org/10.1038/nrm.2017.58.
Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, et al. Ensembl 2022. Nucleic Acids Res. 2022:50(D1):D988–D995. https://doi.org/10.1093/nar/gkab1049.
Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017:35(4):316–319. https://doi.org/10.1038/nbt.3820.
Domazet-Loso T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet TIG. 2007:23(11):533–539. https://doi.org/10.1016/j.tig.2007.08.014.
Domazet-Lošo T, Carvunis A-R, Albà MM, Šestak MS, Bakaric R, Neme R, Tautz D. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Mol Biol Evol. 2017:34:843–856. https://doi.org/10.1093/molbev/msw284.
Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of *Brassicaceae* specific genes in *Arabidopsis thaliana*. BMC Evol Biol. 2011:11(1):47. https://doi.org/10.1186/1471-2148-11-47.
Doolittle WF. We simply cannot go on being so vague about "function". Genome Biol. 2018:19(1):223. https://doi.org/10.1186/s13059-018-1600-4.
Doolittle WF, Brunet TDP, Linquist S, Gregory TR. Distinguishing between "function" and "effect" in genome biology. Genome Biol Evol. 2014:6(5):1234–1237. https://doi.org/10.1093/gbe/evu098.
Elghraoui A, Mirarab S, Swenson KM, Valafar F. Evaluating impacts of syntenic block detection strategies on rearrangement phylogeny using *Mycobacterium tuberculosis* isolates. Bioinformatics. 2023:39(1):btad024. https://doi.org/10.1093/bioinformatics/btad024.
Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019:20(1):238. https://doi.org/10.1186/s13059-019-1832-y.
Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 2006:4(1):41. https://doi.org/10.1186/1741-7007-4-41.
Gotea V, Petrykowska HM, Elnitski L. Bidirectional promoters as important drivers for the emergence of species-specific transcripts. PLoS One. 2013:8(2):e57323. https://doi.org/10.1371/journal.pone.0057323.
Grandchamp A, Kühl L, Lebherz M, Brüggemann K, Parsch J, Bornberg-Bauer E. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in *Drosophila melanogaste*. Genome Res. 2023:33(6):872–890. https://doi.org/10.1101/gr.277482.122.
Jacob F. Evolution and tinkering. Science. 1977:196(4295):1161–1166. https://doi.org/10.1126/science.860134.
Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013:30(4):772–780. https://doi.org/10.1093/molbev/mst010.
Keeling DM, Garza P, Nartey CM, Carvunis A-R. The meanings of 'function' in biology and the problematic case of de novo gene emergence. eLife. 2019:8:e47014. https://doi.org/10.7554/eLife.47014.
Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009:19(10):1752–1759. https://doi.org/10.1101/gr.095026.109.
Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol. 2022:39(8):msac174. https://doi.org/10.1093/molbev/msac174.
Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. Structural and functional characterization of a putative de novo gene in *Drosophila*. Nat Commun. 2021:12(1):1667. https://doi.org/10.1038/s41467-021-21667-6.
Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 2006:103(26):9935–9939. https://doi.org/10.1073/pnas.0509809103.
Liu D, Hunt M, Tsai IJ. Inferring synteny between genome assemblies: a systematic evaluation. BMC Bioinformatics. 2018:19(1):26. https://doi.org/10.1186/s12859-018-2026-4.
McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016:17(9):567–578. https://doi.org/10.1038/nrg.2016.78.
Montañés JC, Huertas M, Messeguer X, Albà MM. Evolutionary trajectories of new duplicated and putative de novo genes. Mol Biol Evol. 2023:40(5):msad098. https://doi.org/10.1093/molbev/msad098.
Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 2015:32(1):258–267. https://doi.org/10.1093/molbev/msu286.
Papadopoulos C, Arbes H, Chevrollier N, Blanchet S, Cornu D, Roginski P, Rabier C, Atia S, Lespinet O, Namy O, et al. The Ribosome Profiling landscape of yeast reveals a high diversity in pervasive translation. bioRxiv 2023. 2023.03.16.532990. https://doi.org/10.1101/2023.03.16.532990.
Papadopoulos C, Callebaut I, Gelly J-C, Hatin I, Namy O, Renard M, Lespinet O, Lopes A. Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. Genome Res. 2021:31(12):2303–2315. https://doi.org/10.1101/gr.275638.121.
Peng J, Zhao L. The origin and structural evolution of de novo genes in *Drosophila*. Nat Commun. 2024:15(1):810. https://doi.org/10.1038/s41467-024-45028-1.
Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz-Orera J, Bassani-Sternberg M, Moritz RL, Deutsch EW, van Heesch S. What can Ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? Mol Cell Proteomics. 2023:22(9):100631. https://doi.org/10.1016/j.mcpro.2023.100631.
Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010:26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.
Ranz JM, Casals F, Ruiz A. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. Genome Res. 2001:11(2):230–239. https://doi.org/10.1101/gr.162901.
R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 2013:9(10):e1003860. https://doi.org/10.1371/journal.pgen.1003860.
Schlötterer C. Genes from scratch—the evolutionary fate of de novo genes. Trends Genet. 2015:31(4):215–219. https://doi.org/10.1016/j.tig.2015.02.007.
Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018:2(10):1626–1632. https://doi.org/10.1038/s41559-018-0639-7.
Stein JC, Yu Y, Copetti D, Zwickl DJ., Zhang L, Zhang C, Chougule K, Gao D, Iwata A, Goicoechea JL, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and

innovation across the genus *Oryza*. Nat Genet. 2018:50(2): 285–296. https://doi.org/10.1038/s41588-018-0040-0.
Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011:12(10):692–702. https://doi.org/10.1038/nrg3053.
Vakirlis N, Acar O, Cherupally V, Carvunis A-R. Ancestral sequence reconstruction as a tool to detect and study de novo gene emergence. bioRxiv 2024. 2024.01.02.573862. https://doi.org/10.1101/2024.01.02.573862.
Vakirlis N, Carvunis A-R, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. Elife. 2020:9:e53500. https://doi.org/10.7554/eLife.53500.
Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. A molecular portrait of de novo genes in yeasts. Mol Biol Evol. 2018:35(3):631–645. https://doi.org/10.1093/molbev/msx315.
Vakirlis N, McLysaght A. Computational prediction of de novo emerged protein-coding genes. In: Sikosek T, editor. Computational methods in protein evolution. New York (NY): Springer; 2019. p. 63–81 (Methods in Molecular Biology). https://doi.org/10.1007/978-1-4939-8736-8_4.
Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019:15(5): e1008160. https://doi.org/10.1371/journal.pgen.1008160.
Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, Carvunis A-R. A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Syst. 2023:14(5):363–381.e8. https://doi.org/10.1016/j.cels.2023.04.002.
Weisman CM. The origins and functions of de novo genes: against all odds? J Mol Evol. 2022:90(3–4):244–257. https://doi.org/10.1007/s00239-022-10055-3.
Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020:18(11):e3000862. https://doi.org/10.1371/journal.pbio.3000862.
Wu X, Sharp PA. Divergent transcription: a driving force for new gene origination? Cell. 2013:155(5):990–996. https://doi.org/10.1016/j.cell.2013.10.048.
Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007:24(8):1586–1591. https://doi.org/10.1093/molbev/msm088.
Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. Nat Ecol Evol. 2019:3(4):679–690. https://doi.org/10.1038/s41559-019-0822-5.

Associate editor: Claudia Alvarez Carreño