

Origins, evolution, and physiological implications of de novo genes in yeast

Saurin B. Parikh  | Carly Houghton  | S. Branden Van Oss  |
Aaron Wacholder  | Anne-Ruxandra Carvunis 

Department of Computational and Systems Biology, School of Medicine, Pittsburgh Center for Evolutionary Biology and Evolution, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence

Anne-Ruxandra Carvunis, Department of Computational and Systems Biology, School of Medicine, Pittsburgh Center for Evolutionary Biology and Evolution, University of Pittsburgh, Pittsburgh, PA 15213, USA.
Email: anc201@pitt.edu

Funding information

NSF, Grant/Award Number: MCB-2144349

Abstract

De novo gene birth is the process by which new genes emerge in sequences that were previously noncoding. Over the past decade, researchers have taken advantage of the power of yeast as a model and a tool to study the evolutionary mechanisms and physiological implications of de novo gene birth. We summarize the mechanisms that have been proposed to explicate how noncoding sequences can become protein-coding genes, highlighting the discovery of pervasive translation of the yeast transcriptome and its presumed impact on evolutionary innovation. We summarize current best practices for the identification and characterization of de novo genes. Crucially, we explain that the field is still in its nascency, with the physiological roles of most young yeast de novo genes identified thus far still utterly unknown. We hope this review inspires researchers to investigate the true contribution of de novo gene birth to cellular physiology and phenotypic diversity across yeast strains and species.

KEYWORDS

de novo genes, evolutionary biology, genome biology, smORFs, systems biology

1 | A BRIEF HISTORY OF DE NOVO GENE BIRTH RESEARCH

De novo gene birth is the process by which new genes evolve from sequences that were previously noncoding (Tautz, 2014; Tautz & Domazet-Loso, 2011; Van Oss & Carvunis, 2019). Once thought to be exceedingly rare (Jacob, 1977), de novo gene birth has now been observed in a wide variety of taxa (Baalsrud et al., 2018; Cai et al., 2008; Chen et al., 2010; Heinen et al., 2009; Khan et al., 2020; Li et al., 2010; Reinhardt et al., 2013; Weisman, 2022; Xie et al., 2019). Several studies have described how young de novo genes that exist in only a single species can play important biological roles through species-specific molecular mechanisms

(Bungard et al., 2017; Cai et al., 2008; Li et al., 2010, 2014; Xie et al., 2019; Zhuang et al., 2019). The process of de novo gene birth has therefore received considerable recent attention as a major potential source of genetic, structural, and phenotypic novelty (Abrusan, 2013; Bornberg-Bauer et al., 2021; Capra et al., 2010; Chen et al., 2013; Knopp et al., 2019; Lee et al., 2019; McLysaght & Guerzoni, 2015; McLysaght & Hurst, 2016; Reinhardt et al., 2013; Schlotterer, 2015).

Yeasts have played a central role in the field of de novo gene birth since its inception. When the *Saccharomyces cerevisiae* genome was sequenced in 1996, approximately 6000 open reading frames (ORFs) longer than 300 nucleotides were predicted to be protein-coding genes (Goffeau et al., 1996). Of these, around 30%

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. Yeast published by John Wiley & Sons Ltd.

lacked identifiable homologs among known genes from other species—that is, they were “orphan genes” (Dujon, 1996). The sequencing of additional genomes over the course of the subsequent decades led to the identification of homologs for many of these orphans (Brachat et al., 2003; Cliften et al., 2003; Kellis et al., 2003; Riley et al., 2016; Shen et al., 2018; Weisman et al., 2020). Nevertheless, several hundred remained homolog-free, unable to be grouped into any gene family. Since de novo gene birth was considered highly implausible, such lack of cross-species conservation combined with the absence of experimental evidence was thought to indicate a lack of function. The remaining orphans were therefore initially presumed to correspond to mis-annotations, unlikely to encode functional protein-coding genes, and relegated to the status of “dubious” ORFs (Fisk et al., 2006). However, a 2008 survey of dubious ORFs showed that most were in fact expressed and detected in high-throughput functional genomics assays, suggesting that they did not correspond to mere mis-annotations but may encode *bona fide* orphan genes (Q. R. Li et al., 2008). The same year, Cai et al. (2008) demonstrated that the *S. cerevisiae* orphan gene *BSC4* was of de novo origin.

BSC4 was originally identified as a translated ORF exhibiting Sup35-dependent translational readthrough (Namy et al., 2003). Cai et al. (2008) then showed that *BSC4* evolved recently in the *S. cerevisiae* lineage via point mutations in a locus that was previously noncoding. This was the first demonstration that a full-length protein-coding gene can emerge de novo in any species. The authors showed that *BSC4* increases in expression level throughout the stationary phase and, based on synthetic lethal interactions with *RPN4* and *DUN1*, proposed that Bsc4 is involved in DNA repair during the stationary phase to enable the transition from nutrient-rich to nutrient-poor environments. A decade after this initial characterization, Bsc4 became the first protein encoded by a de novo gene to be structurally characterized in any species. Unlike typical conserved yeast proteins, it was found to exhibit a rudimentary “molten globule” fold with high beta-sheet content and a hydrophobic core (Bungard et al., 2017).

Shortly following the characterization of *BSC4*, Li et al. (2010, 2014) deployed an exhaustive set of experiments and analyses (Table 1) to generate the most complete characterization of a yeast de novo gene to date. Their studies showed that the *MDF1* ORF emerged de novo in *S. cerevisiae* in the previously noncoding sequence antisense to a conserved protein-coding gene, *ADF1* (Figure 1a). Interestingly, Adf1 represses transcription of *MDF1* by binding to its promoter such that sense and antisense expression at this locus have antagonistic physiological effects. When expressed, Mdf1 promotes fermentation and suppresses mating by physically interacting with Snf1 and Mata2 (Figure 1b). In other words, the young de novo gene *MDF1* mediates the crosstalk between reproduction and vegetative growth through a *S. cerevisiae*-specific molecular mechanism. The case of *MDF1* illustrates how, contrary to prior assumptions, young ORFs that emerged de novo in noncoding sequences and lack cross-species conservation can encode proteins with key cellular roles.

Take Aways

- Several yeast genes of recent de novo origin play important cellular roles.
- Yeasts express thousands of de novo sequences with unknown biology.
- Yeasts are well suited to address fundamental questions about de novo gene birth.

2 | METHODS FOR INFERRING DE NOVO ORIGIN

The most convincing evidence that an ORF originated de novo is the identification of a set of one or more “enabling mutations” that arose in previously noncoding sequences within the lineage resulting in a new ORF (e.g., mutation/s creating a new start codon) (McLysaght & Hurst, 2016). This is done by aligning the locus containing the ORF of interest with syntenic orthologous DNA regions in closely related species and showing that the enabling mutations are absent in these species, that is, showing that the orthologous DNA regions are truly noncoding (Vakirlis & McLysaght, 2019; Figure 2). A study applying this approach confirmed the de novo origin for 30 *Saccharomyces* ORFs (Vakirlis et al., 2018).

Such synteny analyses for de novo gene birth inference can be further refined using ancestral sequence reconstruction approaches. This was demonstrated for the first time in any species with the de novo *S. cerevisiae* ORF YBR196C-A (Vakirlis, Acar, et al., 2020). Ancestral reconstruction at this locus showed not only how enabling mutations conferred coding potential to an ancestrally noncoding DNA region, but also how subsequent frameshifts and substitutions have led to the rapid evolution of the initial ORF, leading to loss in some lineages and substantial changes in length and primary sequence in others. These mutational processes led to the emergence of a small species-specific transmembrane protein in *S. cerevisiae* that localizes at the endoplasmic reticulum and promotes larger colony growth when overexpressed. A subsequent study (Papadopoulos et al., 2021) used ancestral reconstruction to retrace the evolutionary history of 70 candidate de novo genes identified by previous studies (Carvunis et al., 2012; Lu et al., 2017; Vakirlis et al., 2018; Wu & Knudson, 2018), and reported that most de novo enabling mutations corresponded to frameshifts and loss-of-stop events leading to the merging of two small intergenic ORFs.

This complexity can obfuscate the identification of enabling mutations from syntenic alignment alone when one aims to use automated sequence analyses. To circumvent this challenge, a recent study (Wacholder et al., 2021) adapted a Reading Frame Conservation metric (Kellis et al., 2003) calculated from syntenic alignments to identify all ORFs in the *S. cerevisiae* genome with noncoding orthologous regions in other *Saccharomyces* species. These ORFs were then classified into candidate pseudogenes when distant homologs could be identified through sequence similarity searches

TABLE 1 Applying the evolutionary systems biology approach to the investigation of *MDF1*

Categories	Test for evidence	Results
Sequence	Comparative genomics	It is under positive selection.
	PSI-BLAST	There are no significantly homologous ORFs in all of the other organisms examined beyond two short, truncated ORFs in the close relatives <i>Saccharomyces bayanus</i> and <i>Saccharomyces mikatae</i> .
	Syteny	The intergenic region between flanking genes could not encode a protein in other species due to the presence of multiple stop codons.
Expression	Strand-specific RT-PCR	<i>MDF1</i> is only expressed in <i>Saccharomyces cerevisiae</i> .
	Western blot	Positive signal for the protein.
Structure	Structure prediction server–PORTER	Mdf1 mimics Mata1 in having a three-helix-domain that can bind to Mata2.
Localization	Fluorescent tagging	Mdf1 exists in the cytoplasm and nucleus.
Interaction and mechanism	Chromatin immunoprecipitation	Adf1 binds to the upstream region of <i>MDF1</i> . Mdf1 binds haploid-specific genes (<i>MATα1</i> , <i>STE4</i> , <i>STE5</i> , <i>FUS1</i> , <i>FUS2</i> , <i>FUS3</i> , <i>GPA1</i> , <i>SST2</i> , and <i>RME1</i>).
	Gel electrophoresis	<i>ATP1</i> , <i>PGK1</i> , <i>MDH1</i> , and <i>SAM1</i> expression is increased in <i>MDF1 ADF1Δ</i> strains.
	Microarray	Downregulation of mating pathway (MAPK).
	Semi-quantitative RT-PCR	MAPK pathway genes (<i>STE3</i> , <i>STE12</i> , <i>FUS1</i> , <i>FUS3</i>) are downregulated.
	Complementation assay	Overexpression of <i>MATα1</i> gene rescues the mating ability of an <i>mdf1Δ</i> mutant.
	Yeast two-hybrid assay	Mdf1 interacts with Mata2.
	Pull-down assay	Mdf1 interacts with Mata2.
	Electrophoretic mobility shift assays	Mdf1 and Mata2 are bound to each other and function in a mutually dependent manner.
Phenotype and fitness	Competition experiment	<i>MDF1 ADF1Δ</i> strain grows more quickly than the wild-type strain.
	Growth rate analyses	
	Mating assay	<i>MDF1 ADF1Δ</i> is less successful at mating. No such effect is seen in closely related species.

Abbreviations: MAPK, mitogen-activated protein kinase; ORF, open reading frame; RT-PCR, reverse transcription polymerase chain reaction.

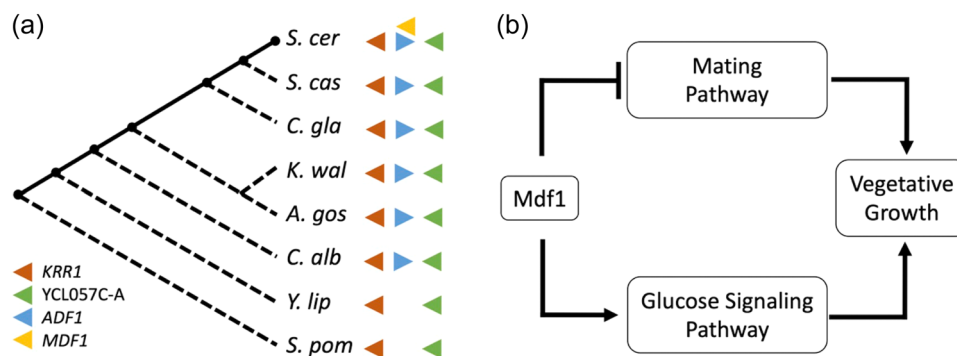


FIGURE 1 *MDF1*: A de novo-evolved gene that integrates into essential biological pathways. (a) Phylogeny- and synteny-based analysis of various fungi revealed that *MDF1* emerged specifically in *S. cer* subsequent to its split from *S. cas*. At the same time, *ADF1*, an antisense gene to *MDF1*, is conserved in all but the most distant member of the hemiascomycete subdivision of fungi. The *MDF1* syntenic block is shown to the right of the phylogenetic tree. (Li et al., 2010) (b) *Mdf1* promotes vegetative growth by suppressing the mating pathway and enhancing the glucose signaling pathway (Li et al., 2014). *A. gos*, *Ashbya gossypii*; *C. alb*, *Candida albicans*; *C. gla*, *Candida glabrata*; *S. cas*, *Saccharomyces castellii*; *S. cer*, *Saccharomyces cerevisiae*; *S. pom*, *Schizosaccharomyces pombe*; *Y. lip*, *Yarrowia lipolytica*.

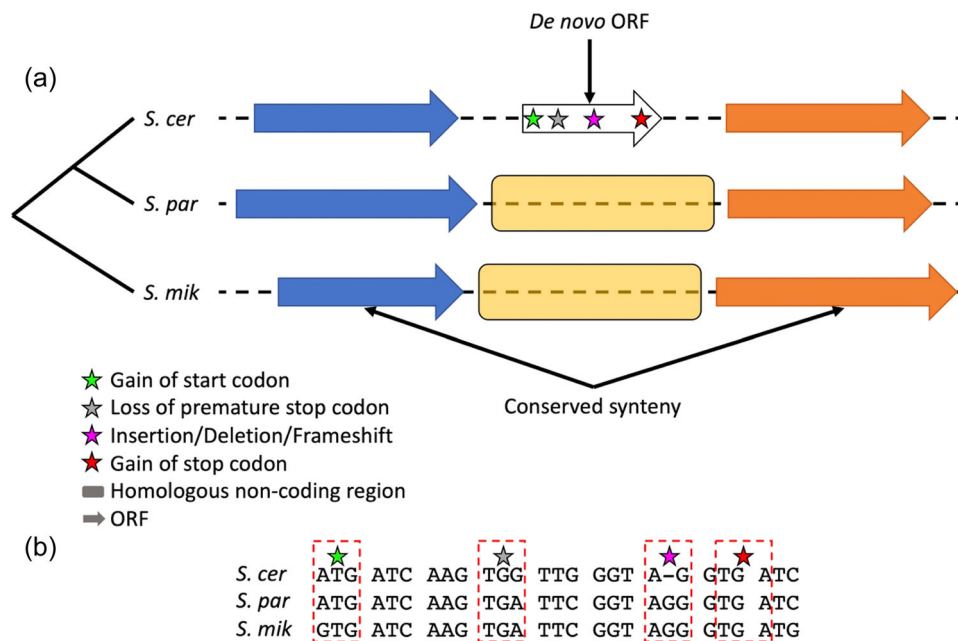


FIGURE 2 Pictographic representation of a hypothetical de novo ORF in *Saccharomyces cerevisiae*. (a) A combination of conserved synteny and phylostratigraphy is used to identify the homologous region of interest (highlighted in yellow) in the closely related species. This region of interest can be used to identify enabling mutations across the lineage that led to the de novo ORF in the focal species (*S. cerevisiae* in this case). The enabling mutations can include but are not limited to a gain of the start codon (green star), loss of premature stop codon (gray star), insertion–deletion and/or a frameshift (pink star) and a gain of stop codon (red star). Figure inspired by Vakirlis and McLysaght (2019). (b) A hypothetical example of enabling mutations that occurred along the lineage to result in a de novo ORF in the focal genome. Changes highlighted within boxes are possible enablers. Identification of one or more of such mutations (example gain of the start codon) are needed to provide convincing evidence of de novo ORF emergence. ORF, open reading frame; *S. cer*, *Saccharomyces cerevisiae*; *S. mik*, *Saccharomyces mikatae*; *S. par*, *Saccharomyces paradoxus*.

in the fungal lineage, or candidate de novo ORFs when the corresponding protein sequence was lineage-specific. This analysis estimated that 251 annotated *S. cerevisiae* ORFs (7 verified, 96 uncharacterized, 148 dubious) emerged de novo.

While approaches based on synteny and enabling mutations are now considered standard, phylostratigraphy-based approaches have been widely used in earlier studies of gene birth. In phylostratigraphy, the origin of a new gene is inferred in the most recent common ancestor of all species with a homolog identified by sequence similarity searches (Domazet-Lošo et al., 2007). Three groups have performed phylostratigraphy analyses on the *S. cerevisiae* genome, providing lists of hundreds of *S. cerevisiae* orphan genes (Carvunis et al., 2012; Lu et al., 2017; Vakirlis et al., 2018); such analyses have also been conducted on *Lachancea* yeasts (Vakirlis et al., 2016). However, these results must be interpreted with caution because orphan genes can originate via several evolutionary mechanisms other than de novo gene birth, including lateral transfer and duplication followed by extreme sequence divergence (Long et al., 2003; Van Oss & Carvunis, 2019). When an orphan gene identified by phylostratigraphy is present in at least two taxa, it is possible to estimate the likelihood that it has acquired a unique sequence through extreme sequence divergence by extrapolating an estimate of its evolutionary rate (Weisman et al., 2020). However, when a gene is found only in one species with no detectable homolog at all,

analyses of synteny and enabling mutations are required to infer the mechanism of origin. A recent analysis (Vakirlis, Carvunis, et al., 2020) showed that sequence divergence is not the main source of orphan genes in *S. cerevisiae*, suggesting that de novo gene birth may play a major role in generating molecular novelty.

A fundamentally different strategy for de novo gene birth inference consists in comparing genome expression patterns, rather than ORF sequences, between yeast strains and species. Indeed, de novo gene birth can take place when a pre-existing noncoding RNA acquires a novel ORF and becomes translated (“transcription first”), or when a pre-existing ORF becomes transcribed and translated (“ORF first”) (Schlotterer, 2015). In this latter case, the “enabling mutations” would be those that lead to a novel transcription or translation event rather than those that lead to a novel ORF. These are harder to identify by DNA sequence analysis than mutations enabling the emergence of an ORF, as the genetic determinants of expression changes over evolutionary time are not as well understood. It is however well established that the yeast lineage undergoes substantial evolutionary transcriptional turnover (H. Li & Johnson, 2010). Lu et al. (2017) identified 4340 putative *S. cerevisiae*-specific de novo genes that are transcribed but share no orthologues in other *Saccharomycetaceae*, most of which were inferred to have arisen from transcript isoforms of ancient genes. By comparing the transcriptomes of *S. cerevisiae* and 10 other yeast species, Blevins

et al. (2021) identified 213 de novo-originated transcripts in *S. cerevisiae*, half of which were in the antisense orientation of other genes and many of which appeared to be translated. At a finer evolutionary scale, Durand et al. (2019) analyzed the turnover of ribosome-associated transcripts among wild *Saccharomyces paradoxus* strains and identified 447 lineage-specific translation events (Durand et al., 2019). While most were attributable to lineage-specific ORF gains and losses, several instances appeared to have been potentiated by lineage-specific increases in expression level.

The prevalence of de novo gene birth in yeast is supported by overwhelming comparative genomic evidence from the aforementioned studies. Yet, there is currently no definitive, community-approved list of which yeast genes have originated de novo. Different approaches yield different—though overlapping—results (Blevins et al., 2021; Papadopoulos et al., 2021). The issue lies, in part, in that there is no consensus operational definition of what constitutes a “gene” in the context of de novo gene birth, where the signatures of evolutionary conservation typically relied on to predict functionality are absent (Keeling et al., 2019). Further developments of computational methods for the detection of de novo-originated genes are also much needed for the advancement of the field (Li et al., 2022). Such advances are more challenging to attain in the yeast lineage than in other eukaryotic lineages whose genomes tend to evolve more slowly. Yet, as a plethora of yeast genomes have now been sequenced (Kurtzman et al., 2011; Peter et al., 2018; Shen et al., 2018; Vakirlis et al., 2016), exciting opportunities for large-scale comparative and evolutionary studies of de novo gene emergence in yeasts are arising.

The increasing availability of intraspecies genome sequences in yeast has also revealed substantial genetic diversity, distinguishing between the “core” and the “accessory” genomes, the latter containing genes specific to sets of isolates or individual strains (McCarthy & Fitzpatrick, 2019). Exploring the extent to which the accessory genome is comprised of de novo genes will likely shed light on the mechanisms by which rapid genetic evolution mediates rapid phenotypic and ecological adaptation. Along these lines, a recent study found that only 41% of young de novo ORFs identified in the S288C reference annotation were fixed in the *S. cerevisiae* species, while most were still segregating (Vakirlis, Acar, et al., 2020). Future studies integrating evolutionary dynamics of sequence and expression variation at the population level will be instrumental in deriving models of ORF and transcript evolution in real time, to shed light on the full extent of de novo gene emergence and its impacts on the diversity of the yeast lineage.

3 | THE “NONCANONICAL TRANSLATOME” AS A RESERVOIR FOR DE NOVO GENE BIRTH

The first unbiased genome-scale transcriptomic studies reported that most of the *S. cerevisiae* genome is transcribed (David et al., 2006; Nagalakshmi et al., 2008). Soon after, the first ribosome profiling

studies revealed widespread translation outside of annotated *S. cerevisiae* genes (Brar et al., 2012; Carvunis et al., 2012; Ingolia et al., 2009; Wilson & Masel, 2011). Shortly following these discoveries in yeast, the phenomenon was also reported in other taxa spanning the tree of life (Ruiz-Orera & Alba, 2019). All these “noncanonical” translated elements had been missed by genome annotations because they tend to be short and rapidly evolving. The “translatome” is much larger, and much more diverse, than currently reflected in genome annotation databases.

Noncanonical translation was originally predicted by early models of de novo gene birth that were largely built on data from yeast (Cai et al., 2008; Carvunis et al., 2012; Masel, 2006; Wilson & Masel, 2011). These models postulated that some of the hundreds of thousands of short ORFs that appear and disappear continuously during the evolution of noncoding sequences could, if transcribed, become translated and expose new genetic variation to the action of natural selection. Those “proto-genes” (Carvunis et al., 2012) with deleterious translation products would be purged away, while those with nearly neutral or adaptive effects would constitute a reservoir for de novo gene emergence. Multiple studies have now uncovered that many, if not most, noncanonical translated elements in yeast are of de novo origin (Durand et al., 2019; Spealman et al., 2018; Wacholder et al., 2021). Most recently, Wacholder et al. (2021) combined Ribo-seq data from 42 published studies and identified strong translation evidence for almost 20,000 noncanonical *S. cerevisiae* ORFs, including 12,129 of apparent de novo origin based on Reading Frame Conservation analyses (Wacholder et al., 2021).

Future empirical studies are needed to estimate the true size of the yeast translatome, considering the expanding genetic diversity of the yeast genome and pan-genome. Computational predictions are not yet possible, as the molecular signals governing which noncanonical ORFs are translated in vivo remain unknown. A small number of proteomic and microscopy studies have detected the protein products of some of these noncanonical translation events in yeast cells (He et al., 2018; Lu et al., 2017; Yagoub et al., 2015), but the vast majority remain undetected. The de novo-translated ORFs include upstream ORFs and downstream ORFs (translated ORFs located upstream and downstream of annotated coding sequences, respectively) as well as ORFs translated from transcripts containing no annotated gene (Blevins et al., 2021; Carvunis et al., 2012; Durand et al., 2019; Li et al., 2021; Smith et al., 2014; Wacholder et al., 2021; Wilson & Masel, 2011). To what extent the noncanonical translatome generates an entirely novel proteome or yields rapidly degraded products that serve to regulate translation and transcript stability remains an open research question. Unknown too is the proportion of the noncanonical translatome that corresponds to translation “noise” and does not contribute to fitness. The fraction of de novo emerged noncanonical translated ORFs that become fixed into de novo genes maintained by selection is estimated to be low (Carvunis et al., 2012; Vakirlis et al., 2018; Wacholder et al., 2021). The fact that cells exert a considerable amount of energy to translate so many novel ORFs raises the question of whether such pervasive translation confers an adaptive fitness advantage, beyond providing the raw material for gene birth.

4 | INSIGHTS INTO THE FEATURES OF DE NOVO GENES AND MECHANISMS OF DE NOVO EMERGENCE

Yeast de novo ORFs, whether annotated or not, tend to share general characteristics: Their primary sequences tend to be very similar to those of intergenic ORFs, and they tend to be short, rapidly evolving, and often expressed in both lineage-specific and condition-specific manners (Basile et al., 2017; Blevins et al., 2021; Carvunis et al., 2012; Durand et al., 2019; Ekman & Elofsson, 2010; Li et al., 2021; Papadopoulos et al., 2021; Vakirlis et al., 2018; Wu & Knudson, 2018). These characteristics are thought to derive directly from their de novo emergence and to be associated with possible physiological corollaries. For example, condition-specific expression of yeast de novo ORFs has been reported in the context of various stresses (Blevins et al., 2021; Carvunis et al., 2012; Doughty et al., 2020; Li et al., 2021; Wacholder et al., 2021). Could these species-specific translated elements represent a rapidly evolving part of the cell's response to stress?

Vakirlis, Acar et al. (2020) provided some evidence to this question by showing that overexpression of young *S. cerevisiae* de novo ORFs with predicted transmembrane domains can increase colony growth under nitrogen or carbon limitation. Transmembrane domains are overrepresented among annotated de novo ORFs in *S. cerevisiae* (Carvunis et al., 2012; Vakirlis, Acar, et al., 2020), but the cellular mechanisms by which increased expression of species-specific transmembrane domains would allow cells to adapt to starvation stress remain to be elucidated. Interestingly, Vakirlis, Acar et al. (2020) did elucidate the evolutionary mechanisms giving rise to the de novo origination of ORFs with transmembrane domains as a direct result of codon biases in the genetic code, whereby transmembrane residues tend to be encoded by thymine-rich codons. A "transmembrane-first" model was therefore proposed whereby translation of intergenic sequences that are rich in thymine have a high propensity to generate transmembrane peptides, which in turn are more likely to be adaptive and retained by natural selection. The transmembrane-first model is, to date, the only proposed model that directly ties molecular mechanisms of de novo gene birth to a specific biophysical protein property associated with an adaptive fitness advantage.

Several studies, however, have identified additional properties of yeast de novo ORFs that are also linked to their evolutionary trajectories and possibly to their physiological roles. In particular, the specific genomic location where de novo emergence takes place appears to greatly influence primary sequence, transcriptional regulation, and evolutionary rate. Vakirlis et al. (2018) reported a strong overrepresentation of de novo ORFs at GC-rich loci across multiple yeast lineages. These loci are depleted in stop codons and often correspond to divergent gene promoters, suggesting a regulatory relationship between these de novo ORFs and their conserved neighbors. Blevins et al. (2021) identified many de novo transcripts located on the opposite strand of conserved genes and coregulated with their overlapping counterparts in response to stress. Loci opposite

protein-coding genes also tend to be depleted in stop codons. An over-representation of yeast de novo ORFs has been reported in rapidly evolving subtelomeric regions (Carvunis et al., 2012) and recombination hot spots (Vakirlis et al., 2018). It is tempting to speculate that genomic regions that are transcriptionally active, fast-evolving, or depleted in stop codons favor not only the emergence but also the functional evolution and retention of de novo genes.

Collectively, these studies suggest the existence of diverse evolutionary avenues for de novo gene birth, each possibly associated with different biophysical protein properties and phenotypic impacts. It is unclear how young de novo ORFs change over time, although some evidence suggests a trend towards increasing foldability (Papadopoulos et al., 2021). It may be that distinct selective pressures favor the emergence of distinct types of proteins in different environments. For example, while intrinsic disorder is predicted to be rare among yeast de novo genes in general (Basile et al., 2017; Carvunis et al., 2012; Ekman & Elofsson, 2010; Vakirlis et al., 2018; Vakirlis, Acar, et al., 2020), it is observed in high excess in older de novo genes from the *Lachancea* lineage (Vakirlis et al., 2018). It is thought that de novo genes increase in length, expression level, and cellular interactivity over time (Abrusan, 2013; Carvunis et al., 2012; Lu et al., 2017; Tautz, 2014), but more mechanistic research is needed to fully understand the long-term evolutionary dynamics of de novo gene origination and evolution. The physiological implications of de novo gene emergence are in dire need of further study as well. No noncanonical de novo-translated ORFs, and very few annotated de novo genes, have been deeply characterized to date.

5 | PROPOSED EVOLUTIONARY SYSTEMS BIOLOGY FRAMEWORK FOR FUTURE INVESTIGATIONS OF DE NOVO GENE BIRTH IN YEAST

The study of de novo gene birth offers an unprecedented paradigm to understand the role of genetic novelty in the emergence of novel protein structures, functions, and phenotypes. By studying genetic elements that are transitioning from noncoding to protein-coding, we can unravel how novelty arises on multiple scales, from the DNA sequence to integration into cellular networks and the possible emergence of new phenotypic traits. Given that novel genes, in general, have been shown to rapidly integrate into cellular networks (Abrusan, 2013; Tsai et al., 2012), network-based approaches may turn out to be very fruitful for understanding what makes a strain or species unique from a molecular standpoint. The example of *MDF1* demonstrates how an emergent de novo protein can rapidly integrate into an existing cellular network and evolve a critical biological role (Figure 1, Table 1) (Li et al., 2010, 2014). For future studies in the emerging field of de novo gene research, we propose a novel framework guided by an evolutionary systems biology approach to utilize yeast's potential as a model and a tool for this field of study (Figure 3). Guided by this framework, related levels of evidence and function (Keeling et al., 2019) can be investigated to characterize de novo ORFs.

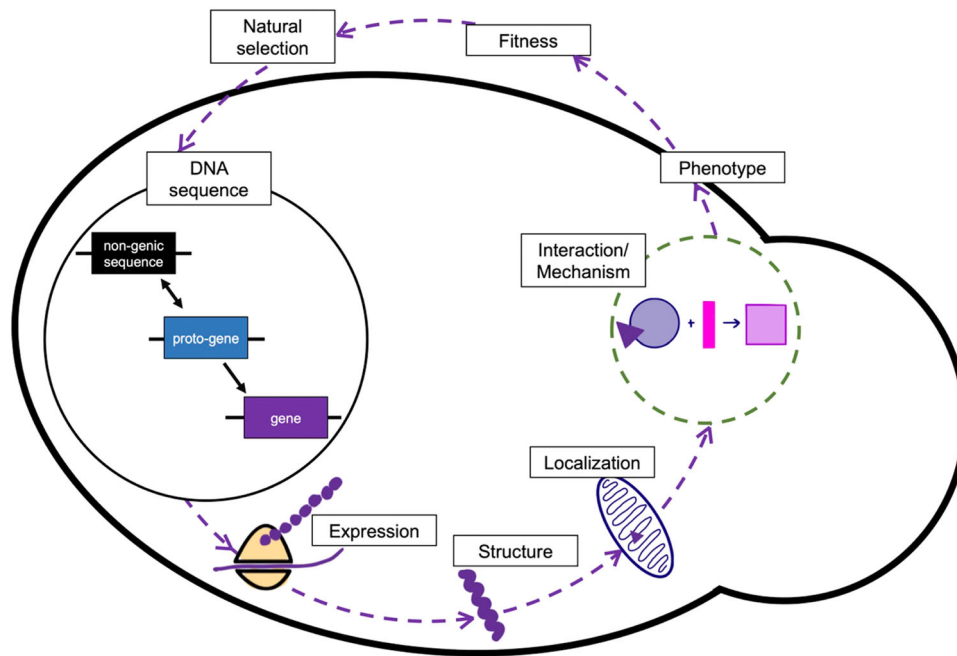


FIGURE 3 Evolutionary systems biology approach for characterizing the biological role of a candidate de novo gene. The framework proposes a combination of evolutionary and molecular approaches that may be used to identify and investigate a candidate de novo gene. Insights drawn from these varied approaches can then be put together to provide a holistic understanding of the ORF's biology. Overall, this framework represents a circular continuum that is under the influence of natural selection. ORF, open reading frame.

For a given de novo candidate, some key questions would be: In what context is it transcribed and translated? When and how did it acquire the regulatory signals controlling expression? Does it participate in genetic or protein–protein interactions? Does it stably localize to a specific subcellular compartment? When and how did its protein sequence acquire the necessary residues or domains to specify its localization and/or interactions? Does its expression impact fitness in a particular biological context? Concomitantly researching a de novo candidate's characteristics along with when and how these characteristics arose is expected to yield insights into the interrelated evolutionary and physiological forces at play. A particular candidate may be required for survival in a specific context, or it may modulate traits that do not impact fitness. As more proto-genes and de novo genes are discovered, the wealth of resources and the repertoire of techniques available to researchers working in yeast combined with our proposed framework (Figure 3) offer a unique opportunity to explore this untapped font of molecular diversity.

Yeasts are established as an exceptional model for molecular genetics, cell biology, and biochemistry due to their ease of culture, simple life cycles, short generation times, a paucity of multi-intronic genes, and their relatively small genomes (~10–20 Mbp.) Genome-wide deletion and overexpression libraries have been developed for multiple yeast strains, particularly in *S. cerevisiae* (Alberti et al., 2007; Brachmann et al., 1998; Douglas et al., 2012; Fasanello et al., 2020; Gelperin et al., 2005; Giaever et al., 2002; Mclsaac et al., 2013; Sopko et al., 2006), enabling advanced, high-throughput approaches that can be expanded to characterize phenotypes for de novo candidates (Costanzo et al., 2010, 2016; Douglas et al., 2012; Parsons et al., 2006; Piotrowski et al., 2017;

Vizeacoumar et al., 2010). Once a phenotype is detected with confidence, mechanisms can be inferred with many tools, for example, with deep mutational scanning (Fowler & Fields, 2014) or network-based computational approaches (Li et al., 2021). As de novo ORFs often overlap with noncoding sequences that may function as regulatory elements or noncoding RNAs, it can be important to experimentally dissect which aspects of null mutant phenotypes are truly caused by loss of translation or loss of the protein product. This can be achieved with single nucleotide genome editing of the translation start site, for example (Wacholder et al., 2021).

Yeasts have also long been at the forefront of the “omics” revolution, offering the opportunity to conduct systems-level studies that can investigate the genome, transcriptome, translome, interactome, metabolome, and phenome (Yu & Nielsen, 2019). Yeast offers yet another advantage over other systems as not only are more and more strains being sequenced every day (Libkind et al., 2020) but such strains are also being used for “comparative-phenomics” in the laboratory setting (Robinson et al., 2021). The exploitation of natural yeast isolates and diverse experimental conditions that attempt to recreate their natural environment may shed light on why so many de novo translated elements exist, and why they evolve so rapidly. It will be informative to compare the tolerance of de novo ORF expression in wild strains and natural environments with that of commonly used laboratory strains and experimental settings.

Yeasts are also well-positioned as a model system for addressing the “holy grail” of de novo gene birth: the opportunity to observe the phenomenon in real-time. While this is not a trivial endeavor, yeasts are amenable to experimental evolution (Voordeckers & Verstrepen,

2015). One can imagine applying selective pressure to an experimentally evolving population and combining it with sequencing to observe the order of events that lead to the formation of de novo genes, and indeed, if a particular path or order is “preferred.” The ability to control this phenomenon opens the possibility that applying appropriate selective pressures may lead to evolved populations with unique genes to overcome current limitations in using yeasts for agricultural, industrial, or medical purposes.

6 | CONCLUSION

Yeasts are well suited to address some of the fundamental questions and promising opportunities in the de novo gene birth field. The pliability of the system allows us to ask nearly any question: How do de novo ORFs acquire the signals necessary for expression? How are new genetic elements integrated into the vast pre-existing *S. cerevisiae* transcriptional and protein-protein interaction networks? How can we perturb these networks to dissect the function (s) of these novel genetic elements? The rapidly evolving field of de novo gene birth can shed new light on our understanding of genes, proteins, and how they evolve. Furthermore, it opens the door for exciting medical and industrial possibilities. Yeasts are uniquely situated to exploit these opportunities.

AUTHOR CONTRIBUTIONS

Saurin Bipin Parikh and Anne-Ruxandra Carvunis: Conceptualization. **Saurin B. Parikh, S. Branden Van Oss, Carly Houghton, Aaron Wacholder:** Writing – original draft. **Saurin B. Parikh, Carly Houghton, S. Branden Van Oss, Aaron Wacholder, Anne-Ruxandra Carvunis:** Writing – review and editing. **Supervision:** Anne-Ruxandra Carvunis

ACKNOWLEDGMENTS

The authors thank the reviewers for their constructive comments during the preparation of this manuscript.

CONFLICT OF INTEREST

Anne-Ruxandra Carvunis is a member of the scientific advisory board for Flagship Labs 69 Inc. (ProFound Therapeutics).

ORCID

Saurin B. Parikh  <https://orcid.org/0000-0002-8167-5768>

Carly Houghton  <https://orcid.org/0000-0003-0158-3044>

S. Branden Van Oss  <https://orcid.org/0000-0002-2933-7083>

Aaron Wacholder  <https://orcid.org/0000-0001-8739-0029>

Anne-Ruxandra Carvunis  <http://orcid.org/0000-0002-6474-6413>

REFERENCES

- Abrusan, G. (2013). Integration of new genes into cellular networks, and their structural maturation. *Genetics*, 195(4), 1407–1417. <https://doi.org/10.1534/genetics.113.152256>
- Alberti, S., Gitler, A. D., & Lindquist, S. (2007). A suite of gateway cloning vectors for high-throughput genetic analysis in *Saccharomyces cerevisiae*. *Yeast*, 24(10), 913–919. <https://doi.org/10.1002/yea.1502>
- Baalsrud, H. T., Torresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., & Jentoft, S. (2018). De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Molecular Biology and Evolution*, 35(3), 593–606. <https://doi.org/10.1093/molbev/msx311>
- Basile, W., Sachenkova, O., Light, S., & Elofsson, A. (2017). High GC content causes orphan proteins to be intrinsically disordered. *PLoS Computational Biology*, 13(3), e1005375. <https://doi.org/10.1371/journal.pcbi.1005375>
- Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Canas, J. L., Espinar, L., Diez, J., Carey, L. B., & Alba, M. M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, 12(1), 604. <https://doi.org/10.1038/s41467-021-20911-3>
- Bornberg-Bauer, E., Hlouchova, K., & Lange, A. (2021). Structure and function of naturally evolved de novo proteins. *Current Opinion in Structural Biology*, 68, 175–183. <https://doi.org/10.1016/j.sbi.2020.11.010>
- Brachat, S., Dietrich, F. S., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T., & Philippsen, P. (2003). Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biology*, 4(7), R45. <https://doi.org/10.1186/gb-2003-4-7-r45>
- Brachmann, C. B., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P., & Boeke, J. D. (1998). Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, 14(2), 115–132. [https://doi.org/10.1002/\(SICI\)1097-0061\(19980130\)14:2<115::AID-YEA204>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2)
- Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., & Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, 335(6068), 552–557. <https://doi.org/10.1126/science.1215110>
- Bungard, D., Copple, J. S., Yan, J., Chhun, J. J., Kumirov, V. K., Foy, S. G., Masel, J., Wysocki, V. H., & Cordes, M. H. J. (2017). Foldability of a natural de novo evolved protein. *Structure*, 25(11), 1687–1696. <https://doi.org/10.1016/j.str.2017.09.006>
- Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*, 179(1), 487–496. <https://doi.org/10.1534/genetics.107.084491>
- Capra, J. A., Pollard, K. S., & Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biology*, 11(12), R127. <https://doi.org/10.1186/gb-2010-11-12-r127>
- Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charlotheaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, 487(7407), 370–374. <https://doi.org/10.1038/nature11184>
- Chen, S., Krinsky, B. H., & Long, M. (2013). New genes as drivers of phenotypic evolution. *Nature Reviews Genetics*, 14(9), 645–660. <https://doi.org/10.1038/nrg3521>
- Chen, S., Zhang, Y. E., & Long, M. (2010). New genes in *Drosophila* quickly become essential. *Science*, 330(6011), 1682–1685. <https://doi.org/10.1126/science.1196380>
- Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., & Johnston, M. (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, 301(5629), 71–76. <https://doi.org/10.1126/science.1084337>
- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S.,

- Prinz, J., Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., ... Boone, C. (2010). The genetic landscape of a cell. *Science*, 327(5964), 425–431. <https://doi.org/10.1126/science.1180823>
- Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z. Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., ... Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):aaf1420. <https://doi.org/10.1126/science.aaf1420>
- David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., & Steinmetz, L. M. (2006). A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America*, 103(14), 5320–5325. <https://doi.org/10.1073/pnas.0601091103>
- Domazet-Loso, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, 23(11), 533–539. <https://doi.org/10.1016/j.tig.2007.08.014>
- Doughty, T. W., Domenzain, I., Millan-Oropeza, A., Montini, N., de Groot, P. A., Pereira, R., Nielsen, J., Henry, C., Daran, J. G., Siewers, V., & Morrissey, J. P. (2020). Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nature Communications*, 11(1), 2144. <https://doi.org/10.1038/s41467-020-16073-3>
- Douglas, A. C., Smith, A. M., Sharifpoor, S., Yan, Z., Durbic, T., Heisler, L. E., Lee, A. Y., Ryan, O., Gottert, H., Surendra, A., van Dyk, D., Giaever, G., Boone, C., Nislow, C., & Andrews, B. J. (2012). Functional analysis with a barcoder yeast gene overexpression system. *G3*, 2(10), 1279–1289. <https://doi.org/10.1534/g3.112.003400>
- Dujon, B. (1996). The yeast genome project: What did we learn. *Trends in Genetics*, 12(7), 263–270. [https://doi.org/10.1016/0168-9525\(96\)10027-5](https://doi.org/10.1016/0168-9525(96)10027-5)
- Durand, E., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dube, A. K., Nielly-Thibault, L., Namy, O., & Landry, C. R. (2019). Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Research*, 29(6), 932–943. <https://doi.org/10.1101/gr.239822.118>
- Ekman, D., & Elofsson, A. (2010). Identifying and quantifying orphan protein sequences in fungi. *Journal of Molecular Biology*, 396(2), 396–405. <https://doi.org/10.1016/j.jmb.2009.11.053>
- Fasanello, V. J., Liu, P., Botero, C. A., & Fay, J. C. (2020). High-throughput analysis of adaptation using barcoded strains of *Saccharomyces cerevisiae*. *PeerJ*, 8, e10118. <https://doi.org/10.7717/peerj.10118>
- Fisk, D. G., Ball, C. A., Dolinski, K., Engel, S. R., Hong, E. L., Issel-Tarver, L., Schwartz, K., Sethuraman, A., Botstein, D., & Cherry, J. M. (2006). *Saccharomyces cerevisiae* S288C genome annotation: A working hypothesis. *Yeast*, 23(12), 857–865. <https://doi.org/10.1002/yea.1400>
- Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nature Methods*, 11(8), 801–807. <https://doi.org/10.1038/nmeth.3027>
- Gelperin, D. M., White, M. A., Wilkinson, M. L., Kon, Y., Kung, L. A., Wise, K. J., Lopez-Hoyo, N., Jiang, L., Piccirillo, S., Yu, H., Gerstein, M., Dumont, M. E., Phizicky, E. M., Snyder, M., & Grayhack, E. J. (2005). Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. *Genes and Development*, 19(23), 2816–2826. <https://doi.org/10.1101/gad.1362105>
- Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., ... Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896), 387–391. <https://doi.org/10.1038/nature00935>
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287), 563–567. <https://doi.org/10.1126/science.274.5287.546>
- He, C., Jia, C., Zhang, Y., & Xu, P. (2018). Enrichment-based proteogenomics identifies microproteins, missing proteins, and novel smORFs in *Saccharomyces cerevisiae*. *Journal of Proteome Research*, 17(7), 2335–2344. <https://doi.org/10.1021/acs.jproteome.8b00032>
- Heinen, T. J., Staubach, F., Haming, D., & Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Current Biology*, 19(18), 1527–1531. <https://doi.org/10.1016/j.cub.2009.07.049>
- Ingolia, N. T., Ghaemmghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924), 218–223. <https://doi.org/10.1126/science.1168978>
- Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295), 1161–1166. <https://doi.org/10.1126/science.860134>
- Keeling, D. M., Garza, P., Nartey, C. M., & Carvunis, A. R. (2019). The meanings of ‘function’ in biology and the problematic case of de novo gene emergence. *eLife*, 8, e47014. <https://doi.org/10.7554/eLife.47014>
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937), 241–254. <https://doi.org/10.1038/nature01644>
- Khan, Y. A., Jungreis, I., Wright, J. C., Mudge, J. M., Choudhary, J. S., Firth, A. E., & Kellis, M. (2020). Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genetics*, 21(1), 25. <https://doi.org/10.1186/s12863-020-0828-7>
- Knopp, M., Gudmundsdottir, J. S., Nilsson, T., König, F., Warsi, O., Rajer, F., Ädelroth, P., & Andersson, D. I. (2019). De novo emergence of peptides that confer antibiotic resistance. *mBio*, 10(3), e00837–00819. <https://doi.org/10.1128/mBio.00837-19>
- Kurtzman, C., Fell, J. W., Boekhout, T., Kurtzman, C. P., Fell, J. W., & Boekhout, T. (2011). *The Yeasts: a Taxonomic Study* (5th ed.). Elsevier Science & Technology.
- Lee, Y. C. G., Ventura, I. M., Rice, G. R., Chen, D. Y., Colmenares, S. U., & Long, M. (2019). Rapid evolution of gained essential developmental functions of a young gene via interactions with other essential genes. *Molecular Biology and Evolution*, 36(10), 2212–2226. <https://doi.org/10.1093/molbev/msz137>
- Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., & Wang, W. (2010). A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research*, 20(4), 408–420. <https://doi.org/10.1038/cr.2010.31>
- Li, D., Yan, Z., Lu, L., Jiang, H., & Wang, W. (2014). Pleiotropy of the de novo-originated gene MDF1. *Scientific Reports*, 4, 7280. <https://doi.org/10.1038/srep07280>
- Li, H., & Johnson, A. D. (2010). Evolution of transcription networks—Lessons from yeasts. *Current Biology*, 20(17), R746–R753. <https://doi.org/10.1016/j.cub.2010.06.056>
- Li, J., Singh, U., Arendsee, Z., & Wurtele, E. S. (2021). Landscape of the dark transcriptome revealed through re-mining massive RNA-Seq data. *Frontiers in Genetics*, 12, 722981. <https://doi.org/10.3389/fgene.2021.722981>
- Li, J., Singh, U., Bhandary, P., Campbell, J., Arendsee, Z., Seetharam, A. S., & Wurtele, E. S. (2022). Foster thy young: Enhanced prediction of orphan genes in assembled genomes. *Nucleic Acids Research*, 50(7), e37. <https://doi.org/10.1093/nar/gkab1238>
- Li, Q. R., Carvunis, A. R., Yu, H., Han, J. D., Zhong, Q., Simonis, N., Tam, S., Hao, T., Klitgord, N. J., Dupuy, D., Mou, D., Wapinski, I., Regev, A., Hill, D. E., Cusick, M. E., & Vidal, M. (2008). Revisiting the

- Saccharomyces cerevisiae* predicted ORFeome. *Genome Research*, 18(8), 1294–1303. <https://doi.org/10.1101/gr.076661.108>
- Libkind, D., Peris, D., Cubillos, F. A., Steenwyk, J. L., Opulente, D. A., Langdon, Q. K., Rokas, A., & Hittinger, C. T. (2020). Into the wild: New yeast genomes from natural environments and new tools for their analysis. *FEMS Yeast Research*, 20(2):foaa008. <https://doi.org/10.1093/femsyr/foaa008>
- Long, M., Betran, E., Thornton, K., & Wang, W. (2003). The origin of new genes: Glimpses from the young and old. *Nature Reviews Genetics*, 4(11), 865–875. <https://doi.org/10.1038/nrg1204>
- Lu, T. C., Leu, J. Y., & Lin, W. C. (2017). A comprehensive analysis of transcript-supported de novo genes in *Saccharomyces sensu stricto* yeasts. *Molecular Biology and Evolution*, 34(11), 2823–2838. <https://doi.org/10.1093/molbev/msx210>
- Masel, J. (2006). Cryptic genetic variation is enriched for potential adaptations. *Genetics*, 172(3), 1985–1991. <https://doi.org/10.1534/genetics.105.051649>
- McCarthy, C. G. P., & Fitzpatrick, D. A. (2019). Pan-genome analyses of model fungal species. *Microbial Genomics*, 5(2):e000243. <https://doi.org/10.1099/mgen.0.000243>
- McIsaac, R. S., Oakes, B. L., Wang, X., Dummit, K. A., Botstein, D., & Noyes, M. B. (2013). Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Research*, 41(4), e57. <https://doi.org/10.1093/nar/gks1313>
- McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 370(1678), 20140332. <https://doi.org/10.1098/rstb.2014.0332>
- McLysaght, A., & Hurst, L. D. (2016). Open questions in the study of de novo genes: What, how and why. *Nature Reviews Genetics*, 17(9), 567–578. <https://doi.org/10.1038/nrg.2016.78>
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881), 1344–1349. <https://doi.org/10.1126/science.1158441>
- Namy, O., Duchateau-Nguyen, G., Hatin, I., Hermann-Le Denmat, S., Termier, M., & Rousset, J. P. (2003). Identification of stop codon readthrough genes in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 31(9), 2289–2296. <https://doi.org/10.1093/nar/gkg330>
- Papadopoulos, C., Callebaut, I., Gelly, J. C., Hatin, I., Namy, O., Renard, M., Lespinet, O., & Lopes, A. (2021). Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. *Genome Research*, 31, 2303–2315. <https://doi.org/10.1101/gr.275638.121>
- Parsons, A. B., Lopez, A., Givoni, I. E., Williams, D. E., Gray, C. A., Porter, J., Chua, G., Sopko, R., Brost, R. L., Ho, C. H., Wang, J., Ketela, T., Brenner, C., Brill, J. A., Fernandez, G. E., Lorenz, T. C., Payne, G. S., Ishihara, S., Ohya, Y., ... Boone, C. (2006). Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell*, 126(3), 611–625. <https://doi.org/10.1016/j.cell.2006.06.040>
- Peter, J., De Chiara, M., Friedrich, A., Yue, J. X., Pflieger, D., Bergstrom, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J. M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., ... Schacherer, J. (2018). Genome evolution across 1011 *Saccharomyces cerevisiae* isolates. *Nature*, 556(7701), 339–344. <https://doi.org/10.1038/s41586-018-0030-5>
- Piotrowski, J. S., Li, S. C., Deshpande, R., Simpkins, S. W., Nelson, J., Yashiroda, Y., Barber, J. M., Safizadeh, H., Wilson, E., Okada, H., Gebre, A. A., Kubo, K., Torres, N. P., LeBlanc, M. A., Andrusiak, K., Okamoto, R., Yoshimura, M., DeRango-Adem, E., van Leeuwen, J., ... Boone, C. (2017). Functional annotation of chemical libraries across diverse biological processes. *Nature Chemical Biology*, 13(9), 982–993. <https://doi.org/10.1038/nchembio.2436>
- Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., & Jones, C. D. (2013). De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genetics*, 9(10), e1003860. <https://doi.org/10.1371/journal.pgen.1003860>
- Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Goker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., Aerts, A. L., Barry, K. W., Choi, C., Clum, A., Coughlan, A. Y., Deshpande, S., Douglass, A. P., Hanson, S. J., Klenk, H. P., ... Jeffries, T. W. (2016). Comparative genomics of biotechnologically important yeasts. *Proceedings of the National Academy of Sciences of the United States of America*, 113(35), 9882–9887. <https://doi.org/10.1073/pnas.1603941113>
- Robinson, D., Place, M., Hose, J., Jochem, A., & Gasch, A. P. (2021). Natural variation in the consequences of gene overexpression and its implications for evolutionary trajectories. *bioRxiv* 2021.2005(2019), 444863. <https://doi.org/10.1101/2021.05.19.444863>
- Ruiz-Orera, J., & Alba, M. M. (2019). Translation of small open reading frames: Roles in regulation and evolutionary innovation. *Trends in Genetics*, 35(3), 186–198. <https://doi.org/10.1016/j.tig.2018.12.003>
- Schlotterer, C. (2015). Genes from scratch—the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4), 215–219. <https://doi.org/10.1016/j.tig.2015.02.007>
- Shen, X. X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R. I., Cadez, N., Libkind, D., ... Rokas, A. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, 175(6), 1533–1545. <https://doi.org/10.1016/j.cell.2018.10.023>
- Smith, J. E., Alvarez-Dominguez, J. R., Kline, N., Huynh, N. J., Geisler, S., Hu, W., Coller, J., & Baker, K. E. (2014). Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Reports*, 7(6), 1858–1866. <https://doi.org/10.1016/j.celrep.2014.05.023>
- Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S. G., Cyert, M., Hughes, T. R., Boone, C., & Andrews, B. (2006). Mapping pathways and phenotypes by systematic gene overexpression. *Molecular Cell*, 21(3), 319–330. <https://doi.org/10.1016/j.molcel.2005.12.011>
- Spealman, P., Naik, A. W., May, G. E., Kuersten, S., Freeberg, L., Murphy, R. F., & McManus, J. (2018). Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Research*, 28(2), 214–222. <https://doi.org/10.1101/gr.221507.117>
- Tautz, D. (2014). The discovery of de novo gene evolution. *Perspectives in Biology and Medicine*, 57(1), 149–161. <https://doi.org/10.1353/pbm.2014.0006>
- Tautz, D., & Domazet-Loso, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10), 692–702. <https://doi.org/10.1038/nrg3053>
- Tsai, Z. T., Tsai, H. K., Cheng, J. H., Lin, C. H., Tsai, Y. F., & Wang, D. (2012). Evolution of cis-regulatory elements in yeast de novo and duplicated new genes. *BMC Genomics*, 13, 717. <https://doi.org/10.1186/1471-2164-13-717>
- Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S. B., Wacholder, A., Medetgul-Ernar, K., Bowman, R. W., Hines, C. P., Iannotta, J., Parikh, S. B., McLysaght, A., Camacho, C. J., O'Donnell, A. F., Ideker, T., & Carvunis, A. R. (2020). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature Communications*, 11(1), 781. <https://doi.org/10.1038/s41467-020-14500-z>

- Vakirlis, N., Carvunis, A. R., & McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife*, 9, e53500. <https://doi.org/10.7554/eLife.53500>
- Vakirlis, N., Hebert, A. S., Oplente, D. A., Achaz, G., Hittinger, C. T., Fischer, G., Coon, J. J., & Lafontaine, I. (2018). A molecular portrait of de novo genes in yeasts. *Molecular Biology and Evolution*, 35(3), 631–645. <https://doi.org/10.1093/molbev/msx315>
- Vakirlis, N., & McLysaght, A. (2019). Computational prediction of de novo emerged protein-coding genes. *Methods in Molecular Biology*, 1851, 63–81. https://doi.org/10.1007/978-1-4939-8736-8_4
- Vakirlis, N., Sarilar, V., Drillon, G., Fleiss, A., Agier, N., Meyniel, J. P., Blanpain, L., Carbone, A., Devillers, H., Dubois, K., Gillet-Markowska, A., Graziani, S., Huu-Vang, N., Poirer, M., Reisser, C., Schott, J., Schacherer, J., Lafontaine, I., Llorente, B., ... Fischer, G. (2016). Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Research*, 26(7), 918–932. <https://doi.org/10.1101/gr.204420.116>
- Van Oss, S. B., & Carvunis, A. R. (2019). De novo gene birth. *PLoS Genetics*, 15(5), e1008160. <https://doi.org/10.1371/journal.pgen.1008160>
- Vizeacoumar, F. J., van Dyk, N., F, S. V., Cheung, V., Li, J., Sydorsky, Y., Case, N., Li, Z., Datti, A., Nislow, C., Raught, B., Zhang, Z., Frey, B., Bloom, K., Boone, C., & Andrews, B. J. (2010). Integrating high-throughput genetic interaction mapping and high-content screening to explore yeast spindle morphogenesis. *Journal of Cell Biology*, 188(1), 69–81. <https://doi.org/10.1083/jcb.200909013>
- Voordeckers, K., & Verstrepen, K. J. (2015). Experimental evolution of the model eukaryote *Saccharomyces cerevisiae* yields insight into the molecular mechanisms underlying adaptation. *Current Opinion in Microbiology*, 28, 1–9. <https://doi.org/10.1016/j.mib.2015.06.018>
- Wacholder, A., Acar, O., & Carvunis, A.-R. (2021). A reference translome map reveals two modes of protein evolution. *bioRxiv*. <https://doi.org/10.1101/2021.07.17.452746>
- Weisman, C. M. (2022). The origins and functions of de novo genes: Against all odds. *Journal of Molecular Evolution*, 90, 244–257. <https://doi.org/10.1007/s00239-022-10055-3>
- Weisman, C. M., Murray, A. W., & Eddy, S. R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biology*, 18(11), e3000862. <https://doi.org/10.1371/journal.pbio.3000862>
- Wilson, B. A., & Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biology and Evolution*, 3, 1245–1252. <https://doi.org/10.1093/gbe/evr099>
- Wu, B., & Knudson, A. (2018). Tracing the de novo origin of protein-coding genes in yeast. *mBio*, 9(4):e01024-18. <https://doi.org/10.1128/mBio.01024-18>
- Xie, C., Bekpen, C., Kunzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K., & Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife*, 8, e44392. <https://doi.org/10.7554/eLife.44392>
- Yagoub, D., Tay, A. P., Chen, Z., Hamey, J. J., Cai, C., Chia, S. Z., Hart-Smith, G., & Wilkins, M. R. (2015). Proteogenomic discovery of a small, novel protein in yeast reveals a strategy for the detection of unannotated short open reading frames. *Journal of Proteome Research*, 14(12), 5038–5047. <https://doi.org/10.1021/acs.jproteome.5b00734>
- Yu, R., & Nielsen, J. (2019). Big data in yeast systems biology. *FEMS Yeast Research*, 19(7):foz070. <https://doi.org/10.1093/femsyr/foz070>
- Zhuang, X., Yang, C., Murphy, K. R., & Cheng, C. C. (2019). Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in Northern gadids. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10), 4400–4405. <https://doi.org/10.1073/pnas.1817138116>

How to cite this article: Parikh, S. B., Houghton, C., Van Oss, S. B., Wacholder, A., & Carvunis, A.-R. (2022). Origins, evolution, and physiological implications of de novo genes in yeast. *Yeast*, 39, 471–481. <https://doi.org/10.1002/yea.3810>