PLOS GENETICS

TOPIC PAGE

De novo gene birth

Stephen Branden Van Oss, Anne-Ruxandra Carvunis*

Department of Computational and Systems Biology, Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States of America

* anc201@pitt.edu

***De novo* gene birth** is the process by which new genes evolve from DNA sequences that were ancestrally non-genic. *De novo* genes represent a subset of novel genes, and may be protein-coding or instead act as RNA genes [1]. The processes that govern *de novo* gene birth (**Fig 1A**) are not well understood, though several models exist that describe possible mechanisms by which *de novo* gene birth may occur. Although *de novo* gene birth may have occurred at any point in an organism's evolutionary history, ancient *de novo* gene birth events are difficult to detect. Most studies of *de novo* genes to date have thus focused on young genes, typically taxonomically-restricted genes (TRGs) that are present in a single species or lineage, including so-called orphan genes, defined as genes that lack any identifiable homolog. It is important to note, however, that not all orphan genes arise *de novo*, and instead may emerge through fairly well-characterized mechanisms such as gene duplication (including retroposition) or horizontal gene transfer followed by sequence divergence, or by gene fission/fusion [2, 3] (**Fig 2**) Though *de novo* gene birth was once viewed as a highly unlikely occurrence [4], there are now several unequivocal examples of the phenomenon that have been described. It furthermore has been advanced that *de novo* gene birth plays a major role in the generation of evolutionary innovation [5, 6].

Check for updates

OPEN ACCESS

Citation: Van Oss SB, Carvunis A-R (2019) *De novo* gene birth. PLoS Genet 15(5): e1008160. https://doi.org/10.1371/journal.pgen.1008160

Published: May 23, 2019

Copyright: © 2019 Van Oss, Carvunis. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Searle Scholars Program to A-RC, the National Institute of General Medical Sciences of the National Institutes of Health grants R00GM108865 to A-RC and F32GM129929 to BVO. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Wikipedia Version: https://en.wikipedia.org/wiki/de_novo_gene_birth.

1 History of the study of *de novo* gene birth

As early as the 1930s, J.B.S. Haldane and others suggested that copies of existing genes may lead to new genes with novel functions [3]. In 1970, Susumu Ohno published the seminal text *Evolution by Gene Duplication [9]*. For some time subsequently, the consensus view was that virtually all genes were derived from ancestral genes [10], with François Jacob famously remarking in a 1977 essay that "the probability that a functional protein would appear *de novo* by random association of amino acids is practically zero" [4]. In the same year, however, Pierre-Paul Grassé coined the term "overprinting" to describe the emergence of genes through the expression of alternative open reading frames (ORFs) that overlap preexisting genes [11] (**Fig 1B**). These new ORFs may be out of frame with or antisense to the preexisting gene. They may also be in frame with the existing ORF, creating a truncated version of the original gene, or represent 3' extensions of an existing ORF into a nearby ORF. The first two types of overprinting may be thought of as a particular subtype of *de novo* gene birth; although overlapping with a previously coding region of the genome, the primary amino-acid sequence of the newly encoded protein is entirely novel. The first examples of this phenomenon in bacteriophages were reported in a series of studies from 1976 to 1978 [12–14], and since then numerous other examples have been identified in viruses, bacteria, and several eukaryotic species [15–19]. The phenomenon of exonization also represents a special case of *de novo* gene birth, in which, for example, often-repetitive intronic sequences acquire splice sites through mutation, leading to *de novo* exons (**Fig 1C**). This was first described in 1994 in the context of *Alu* sequences found

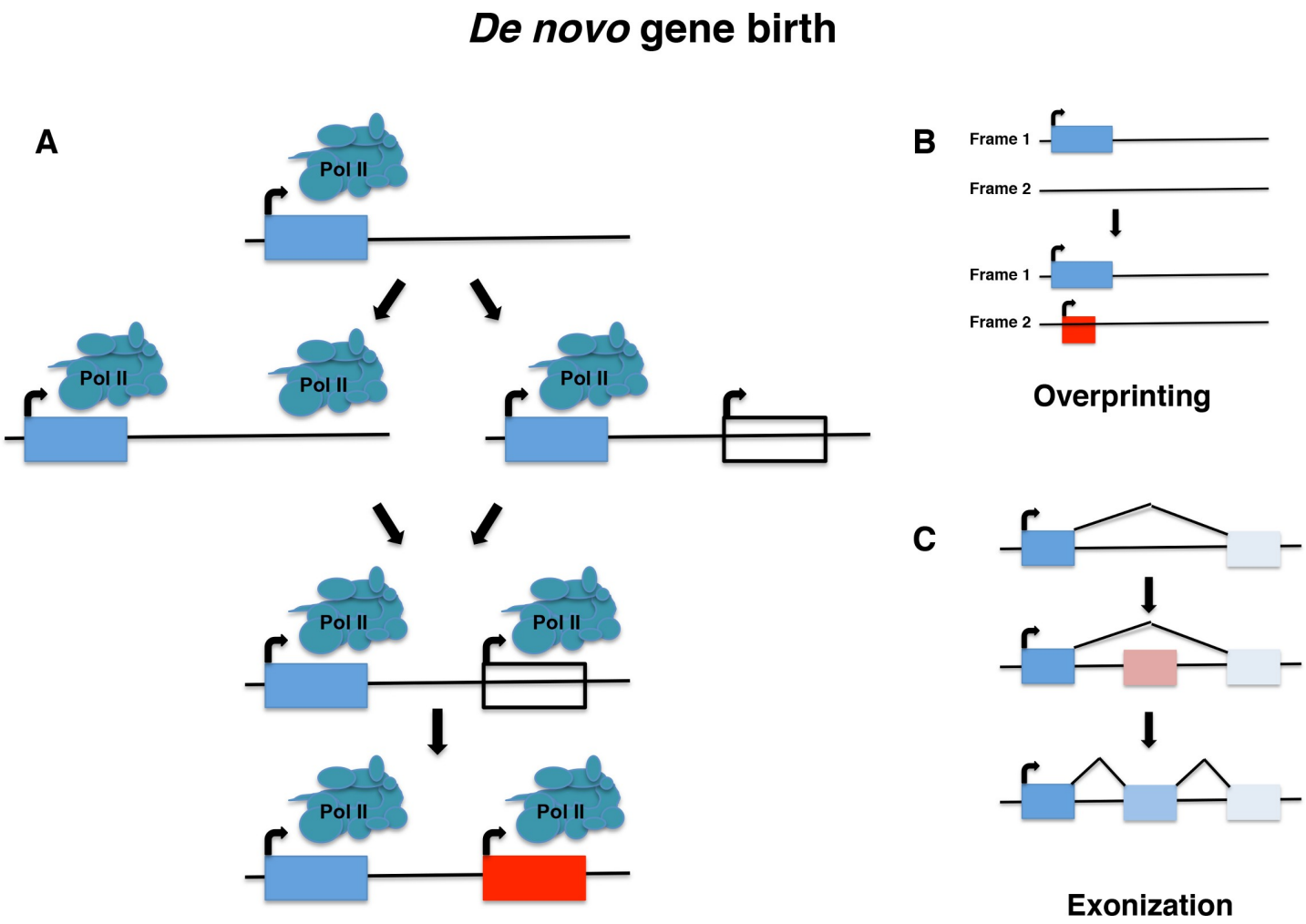

Fig 1. *De novo* gene birth. Novel genes can emerge from ancestrally non-genic regions through poorly understood mechanisms. **(A)** A non-genic region first gains transcription and an ORF, in either order, facilitating the birth of a *de novo* gene. The ORF is for illustrative purposes only, as *de novo* genes may also be multi-exonic, or lack an ORF, as with RNA genes. **(B)** Overprinting. A novel ORF is created that overlaps with an existing ORF, but in a different frame. **(C)** Exonization. A formerly intronic region becomes alternatively spliced as an exon, such as when repetitive sequences are acquired through retroposition and new splice sites are created through mutational processes. Overprinting and exonization may be considered as special cases of *de novo* gene birth.

https://doi.org/10.1371/journal.pgen.1008160.g001

in the coding regions of primate mRNAs [20]. Interestingly, such *de novo* exons are frequently found in minor splice variants, which may allow the evolutionary "testing" of novel sequences while retaining the functionality of the major splice variant(s) [21].

Still, it was thought by some that most or all eukaryotic proteins were constructed from a constrained pool of "starter type" exons [22]. Using the sequence data available at the time, a 1991 review estimated the number of unique, ancestral eukaryotic exons to be < 60,000 [22], while in 1992 a piece was published estimating that the vast majority of proteins belonged to no more than 1,000 families [23]. Around the same time, however, the sequence of chromosome III of the budding yeast *Saccharomyces cerevisiae* was released [24], representing the first time an entire chromosome from any eukaryotic organism had been sequenced. Sequencing of the entire yeast nuclear genome was then completed by early 1996 through a massive, collaborative international effort [25]. In his review of the yeast genome project, Bernard Dujon

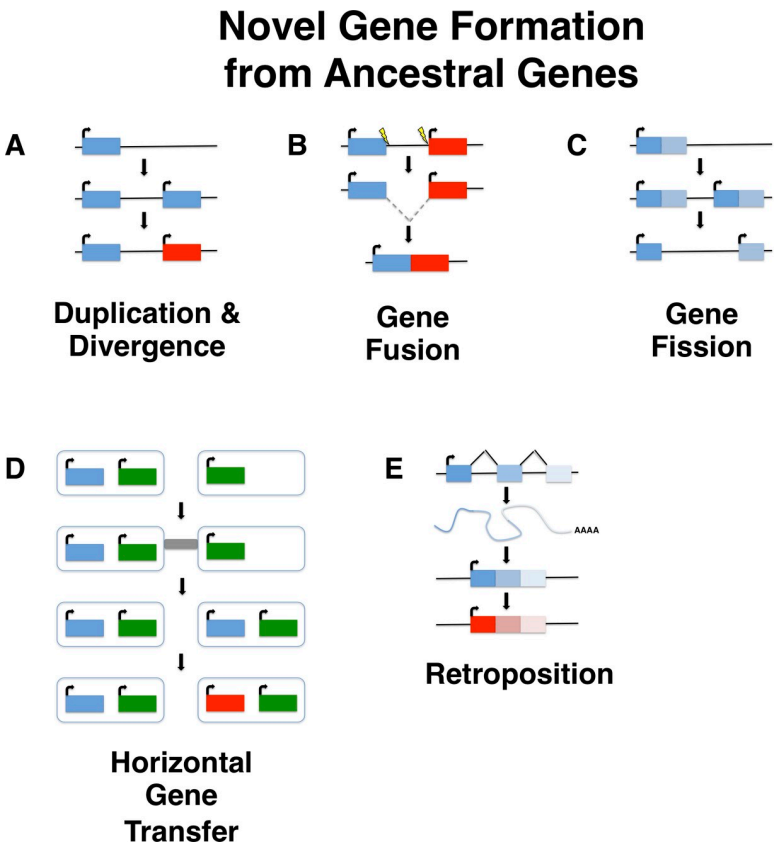

Fig 2. Novel gene formation from ancestral genes. Novel genes can be formed from ancestral genes through a variety of mechanisms. Inspired by Table 1 from [7]. **(A)** Duplication and divergence. Following duplication, one copy experiences relaxed selection and gradually acquires novel function(s). **(B)** Gene fusion. A hybrid gene formed from some or all of two previously separate genes. Gene fusions can occur by different mechanisms; shown here is an interstitial deletion. **(C)** Gene fission. A single gene separates to form two distinct genes, such as by duplication and differential degeneration of the two copies [8]. **(D)** Horizontal gene transfer. Genes acquired from other species by horizontal transfer undergo divergence and neofunctionalization. **(E)** Retroposition. Transcripts may be reverse transcribed and integrated as an intronless gene elsewhere in the genome. This new gene may then undero divergence.

https://doi.org/10.1371/journal.pgen.1008160.g002

noted that the unexpected abundance of genes lacking any known homologs was perhaps the most striking finding of the entire project [25].

In 2006 and 2007, a series of studies provided arguably the first documented examples of full-length *de novo* gene birth [26–28]. An analysis of the accessory gland transcriptomes of *Drosophila yakuba* and *Drosophila erecta* first identified 20 putative lineage-restricted genes that appeared unlikely to have resulted from gene duplication [28]. Levine and colleagues then confirmed the *de novo* origination of five genes specific to *Drosophila* melanogaster and/or the closely related *Drosophila simulans* through a rigorous pipeline that combined bioinformatic and experimental techniques [27]. These genes were identified by combining BLAST search-based and synteny-based approaches (see below), which demonstrated the absence of the genes in closely-related species [27]. Despite their recent evolution, all five genes appear fixed in *D. melanogaster*, and the presence of paralogous non-coding sequences that are absent in close relatives suggests that four of the five genes may have arisen through a recent intrachromosomal duplication event *[27]*. Interestingly, all five were preferentially expressed in the testes of male flies [27] (see below). The three genes for which complete ORFs exist in both *D. melanogaster* and *D. simulans* showed evidence of rapid evolution and positive selection [27].

This is consistent with a recent emergence of these genes, as it is typical for young, novel genes to undergo adaptive evolution [29–31]. A subsequent study using methods similar to Levine *et al*. and an expressed sequence tag library derived from *D*. *yakuba* testes identified seven genes derived from six unique *de novo* gene birth events in *D*. *yakuba* and/or the closely related *D*. *erecta* [26]. Three of these genes are extremely short (<90 bp), suggesting that they may be RNA genes [26], although several examples of very short functional peptides have also been documented [32–35]. Around the same time as these studies in *Drosophila* were published, a homology search of genomes from all domains of life, including 18 fungal genomes, identified 132 fungal-specific proteins, 99 of which were unique to *S*. *cerevisiae* [36].

Since these initial studies, many groups have identified specific cases of *de novo* gene birth events in diverse organisms [37]. The *BSC4* gene in *S*. *cerevisiae*, identified in 2008, shows evidence of purifying selection, is expressed at both the mRNA and protein levels, and when deleted is synthetically lethal with two other yeast genes, all of which indicate a functional role for the *BSC4* gene product [38]. Historically, one argument against the notion of widespread *de novo* gene birth is the evolved complexity of protein folding. Interestingly, Bsc4 was later shown to adopt a partially folded state that combines properties of native and non-native protein folding [39]. Another well-characterized example in yeast is *MDF1*, which both represses mating efficiency and promotes vegetative growth, and is intricately regulated by a conserved antisense ORF [40, 41]. In plants, the first *de novo* gene to be functionally characterized was *QQS*, an *Arabidopsis thaliana* gene identified in 2009 that regulates carbon and nitrogen metabolism [42]. The first functionally characterized *de novo* gene identified in mice, a non-coding RNA gene, was also described in 2009 [43]. In primates, a 2008 informatic analysis estimated that 15/270 primate orphan genes had been formed *de novo [44]*. A 2009 report identified the first three *de novo* human genes, one of which is a therapeutic target in chronic lymphocytic leukemia [45]. Since this time, a plethora of genome-level studies have identified large numbers of orphan genes in many organisms (**Table 1**), although the extent to which they arose *de novo* remains debated.

2 Identification of *de novo* genes

2.1 Identification of *de novo* emerging sequences

There are two major approaches to the systematic identification of novel genes: genomic phylostratigraphy [46] and synteny-based methods. Both approaches are widely used, individually or in a complementary fashion (**Table 1**).

2.1.1 Genomic phylostratigraphy. Genomic phylostratigraphy involves examining each gene in a focal species and inferring the presence or absence of ancestral homologs through the use of the BLAST sequence alignment algorithms [47] or related tools. Each gene in the focal species can be assigned an "age" (aka "conservation level" or "genomic phylostrata") that is based on a predetermined phylogeny, with the age corresponding to the most distantly related species in which a homolog is detected [46]. When a gene lacks any detectable homolog outside of its own genome, or close relatives, it is said to be a novel, taxonomically-restricted or orphan gene, although such a designation is of course dependent on the group of species being searched against.

Phylogenetic trees are limited by the set of closely related genomes that are available, and results are dependent on BLAST search criteria [48]. Because it is based on sequence similarity, it is often difficult for phylostratigraphy to determine whether a novel gene has emerged *de novo* or has diverged from an ancestral gene beyond recognition, for instance following a duplication event. This was pointed out by a study that simulated the evolution of genes of equal age and found that distant orthologs can be undetectable for the most rapidly evolving

Table 1. Genome-scale studies of orphan and *de novo* genes in various lineages. For purposes of this table, genes are defined as **orphan genes** (when species-specific) or **TRGs** (when limited to a closely related group of species) when the mechanism of origination has not been investigated, and as ***de novo* genes** when *de novo* origination has been inferred, irrespective of method of inference. The designation of *de novo* genes as "candidates" or "proto-genes" reflects the language used by the authors of the respective studies.

Organism /Lineage	Homology Detection Method(s)	Evidence of Expression?	Evidence of Selection?	Evidence of Physiological Role?	# Orphan/*De Novo* Genes	Notes	Ref.
Arthropods	BLASTP for all 30 species against each other, TBLASTN for *Formicidae* only, searched by synteny for unannotated orthologs in *Formicidae* only	ESTs, RNA-seq; RT-PCR on select candidates	37 *Formicidae*-restricted orthologs appear under positive selection (M1a to M2a and M7 to M8 models using likelihood ratio tests); as a group, *Formicidae*-restricted orthologs have a significantly higher K_a/K_srate than non-restricted orthologs	Prediction of signal peptides and subcellular localization for subset of orphans	~65,000 orphan genes across 30 species	Abundance of orphan genes dependent on time since emergence from common ancestor; >40% of orphans from intergenic matches indicating possible *de novo* origin	[80]
Arabidopsis thaliana	BLASTP against 62 species, PSI-BLAST against NCBI nonredundant protein database, TBLASTN against PlantGDB-assembled unique transcripts database; searched syntenic region of two closely related species	Transcriptomic and translatomic data from multiple sources	Allele frequencies of *de novo* genes correlated with their DNA methylation levels	None	782 *de novo* genes	Also assessed DNA methylation and histone modifications	[62]
Bombyx mori	BLASTP against four lepidopterans, TBLASTN against lepidopteran EST sequences, BLASTP against NCBI nonredundant protein database	Microarray, RT-PCR	None	RNAi on five *de novo* genes produced no visible phenotypes	738 orphan genes	Five orphans identified as *de novo* genes	[87]
Brassicaceae	BLASTP against NCBI nonredundant protein database, TBLASTN against NCBI nucleotide database, TBLASTN against NCBI EST database, PSI-BLAST against NCBI nonredundant protein database, InterProScan [145]	Microarray	None	TRGs enriched for expression changes in response to abiotic stresses compared to other genes	1761 nuclear TRGs; 28 mitochondrial TRGs	~2% of TRGs thought to be *de novo* genes	[88]
Drosophila melanogaster	BLASTN of query cDNAs against *D. melanogaster*, *D. simulans* and *D. yakuba* genomes; also performed check of syntenic region in sister species	cDNA/ expressed sequence tags (ESTs)	K_a/K_s ratios calculated between retained new genes and their parental genes are significantly >1, indicating most new genes are functionally constrained	List includes several genes with characterized molecular roles	72 orphan genes; 2 *de novo* genes	Gene duplication dominant mechanism for new genes; 7/59 orphans specific to *D. melanogaster* species complex identified as *de novo*	[65]

(*Continued*)

Table 1. (Continued)

Organism /Lineage	Homology Detection Method(s)	Evidence of Expression?	Evidence of Selection?	Evidence of Physiological Role?	# Orphan/*De Novo* Genes	Notes	Ref.
Drosophila melanogaster	Presence or absence of orthologs in other *Drosophila* species inferred by synteny based on UCSC genome alignments and FlyBase protein-based synteny; TBLASTN against *Drosophila* subgroup	Indirect (RNAi)	Youngest essential genes show signatures of positive selection ($\alpha = 0.25$ as a group)	Knockdown with constitutive RNAi lethal for 59 TRGs	195 "young" (>35myo) TRGs; 16 *de novo* genes	Gene duplication dominant mechanism for new genes	[63]
Drosophila melanogaster	RNA-seq in *D. melanogaster* and close relatives; syntenic alignments with *D. simulans* and *D. yakuba*; BLASTP against NCBI nonredundant protein database	RNA-seq	Nucleotide diversity lower in non-expressing relatives; Hudson-Kreitman-Aguade-like statistic lower in fixed *de novo* genes than in intergenic regions	Structural features of *de novo* genes (e.g. enrichment of long ORFs) suggestive of function	106 fixed and 142 segregating *de novo* genes	Specifically expressed in testes	[64]
Homo sapiens	BLASTP against other primates; BLAT against chimpanzee and orangutan genomes, manual check of syntenic regions in chimpanzee and orangutan	RNA-seq	Substitution rate provides some evidence for weak selection; 59/60 *de novo* genes are fixed	None	60 *de novo* genes	Enabling mutations identified; highest expression seen in brain and testes	[66]
Homo sapiens	BLASTP against chimpanzee, BLAT and Ssearch of syntenic region in chimpanzee, manual check of syntenic regions in chimpanzee and macaque	EST/cDNA	No evidence of selective constraint seen by nucleotide divergence	One of the genes identified has a known role in leukemia	3 *de novo* genes	Estimated that human genome contains ~ 18 human-specific *de novo* genes	[45]
Lachancea and *Saccharomyces*	BLASTP of all focal species against each other, BLASTP against NCBI nonredundant protein database, PSI-BLAST against NCBI nonredundant protein database, HMM Profile-Profile of TRG families against each other; families then merged and searched against four profile databases	Mass Spectrometry (MS)	K_a/K_s ratios across *Saccharomyces*indicate that candidates are under weak selection that increases with gene age; in *Lachancea* species with multiple strains, pN/pS ratios are lower for *de novo* candidates than for "spurious TRGs"	None	288 candidate *de novo* genes	MS evidence of translation for 25 candidates	[90]

(*Continued*)

Table 1. (Continued)

Organism /Lineage	Homology Detection Method(s)	Evidence of Expression?	Evidence of Selection?	Evidence of Physiological Role?	# Orphan/*De Novo* Genes	Notes	Ref.
Mus musculus and *Rattus norvegicus*	BLASTP of rat and mouse against each other, BLASTP against Ensembl compara database; searched syntenic regions in rat and mouse	UniGene Database	Subset of genes shows low nucleotide diversity and high ORF conservation across 17 strains	Two mouse genes cause morbidity when knocked out	69 *de novo* genes in mouse and 6 "de novo" genes in ra	Enabling mutations identified for 9 mouse genes	[146]
Mus musculus	BLASTP against NCBI nonredundant protein database	Microarray	None	None	781 orphan genes	Age-dependent features of genes compatible with *de novo* emergence of many orphans	[76]
Oryza	Protein-to-protein and nucleotide-to-nucleotide BLAT against eight *Oryza* species and two outgroup species; searched syntenic regions of these species for coding potential	RNA-seq (all *de novo* TRGs); Ribosome Profiling and targeted MS (some *de novo* TRGs)	22 *de novo* candidates appear under negative selection, and six under positive selection, as measured by K_a/K_s rate	Expression of *de novo* TRGs is tissue-specific	175 *de novo* TRGs	~57% of *de novo* genes have translational evidence; transcription predates coding potential in most cases	[147]
Primates	BLASTP against 15 eukaryotes, BLASTN against human genome, analysis of syntenic regions	ESTs	K_a/K_s ratios for TRGs below one but higher than established genes; coding scores consistent with translated proteins	Several genes have well-characterized cellular roles	270 TRGs	~5.5% of TRGs estimated to have originated *de novo*	[44]
Rodentia	BLASTP against NCBI nonredundant protein database	None	Mouse genes share 50% identity with rat ortholog	None	84 TRGs	Species-specific genes excluded from analysis; results robust to evolutionary rate	[98]
Saccharomyces cerevisiae	BLASTP and PSI-BLAST against 18 fungal species, HMMER and HHpred against several databases, TBLASTN against three close relatives	None	None	Majority of orphans have characterized fitness effects	188 orphan genes	Ages of genes determined at level of individual residues	[83]
Saccharomyces cerevisiae	BLASTP, TBLASTX, and TBLASTN against 14 other yeast species, BLASTP against NCBI nonredundant protein database	Ribosome Profiling	All 25 *de novo* genes, 115 proto-genes under purifying selection ($pN/pS < 1$)	None	25 *de novo* genes; 1,891 "proto-genes"	*De novo* gene birth more common than new genes from duplication; proto-genes are unique to *Saccharomyces sensu strictu* yeasts	[75]

(*Continued*)

Table 1. (Continued)

Organism /Lineage	Homology Detection Method(s)	Evidence of Expression?	Evidence of Selection?	Evidence of Physiological Role?	# Orphan/*De Novo* Genes	Notes	Ref.
Saccharomyces sensu strictu	BLASTP against NCBI nonredundant protein database, TBLASTN against ten outgroup species; BLASTP and phmmer against 20 yeast species reannotated using syntenic alignments	Transcript isoform sequencing (TIF-seq), Ribosome Profiling	Most genes weakly constrained but a subset under strong selection, according to Neutrality Index, Direction of Selection, K_a/K_s, and McDonald-Kreitman tests	Subcellular localization demonstrated for five genes	~13,000 *de novo* genes	>65% of *de novo* transcripts are isoforms of ancient genes; >97% from TIF-seq dataset	[61]

https://doi.org/10.1371/journal.pgen.1008160.t001

genes [49]. When accounting for changes in the rate of evolution to portions of young genes that acquire selected functions, a phylostratigraphic approach was much more accurate at assigning gene ages in simulated data [50]. A subsequent pair of studies using simulated evolution found that phylostratigraphy failed to detect an ortholog in the most distantly related species for 13.9% of *D. melanogaster* genes and 11.4% of *S. cerevisiae* genes [51, 52]. Similarly, a spurious relationship between a gene's age and its likelihood to be involved in a disease process was claimed to be detected in the simulated data [52]. However, a reanalysis of studies that used phylostratigraphy in yeast, fruit flies and humans found that even when accounting for such error rates and excluding difficult-to-stratify genes from the analyses, the qualitative conclusions were unaffected for all three studies [53]. The impact of phylostratigraphic bias on studies examining various features of *de novo* genes (see below) remains debated.

To increase the detectability of ancestral homologues, sensitive sequence-based similarity searches, such as CS-BLAST and Hidden Markov Model (HMM)-based searches, may also be used, alone or in combination with BLAST-based phylostratigraphy analysis, to identify *de novo* genes. The PSI-BLAST technique [54] is particularly useful for detecting ancient homologs. A benchmarking study found that some of these "profile-based" analyses were more accurate than conventional pairwise tools [55]. The impact of false positives, when genes are incorrectly inferred to have an ancestral homolog when they are new in reality, on our understanding of *de novo* gene birth has not yet been specifically assessed.

It is important to disentangle the technical difficulties associated with detection of the oldest ancestor of a gene, and estimates of how old a gene is (the ultimate goal of phylostratigraphy), from challenges linked to inferring the mechanisms by which a gene has evolved. Young and ancestral genes can all have evolved *de novo*, or through other mechanisms. The current approach of choice to determine whether a gene has emerged *de novo* is synteny, and can generally only be applied to young genes.

2.1.2 Synteny-based approaches. Approaches based on the analysis of syntenic sequences in outgroups–blocks of sequence in which the order and relative positioning of features has been maintained–allow for the identification of non-genic ancestors of candidate *de novo* genes [6, 48]. Syntenic alignments are anchored by short, conserved "markers." Genes are the most common marker in defining syntenic blocks, although k-mers and exons are also used [56, 57]. Assuming that a high-quality syntenic alignment can be obtained, confirmation that the syntenic region lacks coding potential in outgroup species allows a *de novo* origin to be asserted with higher confidence [48]. The strongest possible evidence for *de novo* emergence is the inference of the specific mutation(s) that created coding potential, typically through the analysis of microsyntenic regions of closely related species.

One challenge in applying synteny-based methods is the fact that synteny can be difficult to detect across longer timescales. To address this, various techniques have been tried, such as using exons clustered irrespective of their specific order to define syntenic blocks [57] or algorithms that use well-conserved genomic regions to expand microsyntenic blocks [58]. There are also difficulties associated with applying synteny-based approaches to genome assemblies that are fragmented [59] or in lineages with high rates of chromosomal rearrangements, as is common in insects [60]. Although synteny-based approaches have conventionally been lower-throughput in nature, they are now being applied to genome-wide surveys of *de novo* genes [44, 45, 61–66] and represent a promising area of algorithmic development for gene birth dating. Some have used synteny-based approaches in combination with similarity searches in an attempt to develop standardized, stringent pipelines [67] that can be applied to any group of genomes in an attempt to address discrepancies in the various lists of *de novo* genes that have been generated (see below).

2.2. Determination of *de novo* gene status

Even when the evolutionary origin of a particular sequence has been rigorously established computationally, it is important to note that there is a lack of consensus about what constitutes a genuine *de novo* gene birth event. One reason for this is a lack of agreement on whether or not the entirety of the newly genic sequence must be non-genic in origin. With respect to protein-coding *de novo* genes, it has been proposed that *de novo* genes be divided into subtypes corresponding to the proportion of the ORF in question that was derived from previously non-coding sequence [48]. Furthermore, for *de novo* gene birth to occur, the sequence in question must not just have emerged *de novo* but must in fact be a gene. Accordingly, the discovery of *de novo* gene birth has also led to a questioning of what constitutes a gene, with some models establishing a strict dichotomy between genic and non-genic sequences, and others proposing a more fluid continuum (see below). All definitions of genes are linked to the notion of function, as it is generally agreed that a genuine gene should encode a functional product, be it RNA or protein. There are, however, different views of what constitutes function, depending in part on whether a given sequence is assessed using genetic, biochemical, or evolutionary approaches [48, 68, 69].

It is generally accepted that a genuine *de novo* gene is expressed in at least some context [2], allowing selection to operate, and many studies use evidence of expression as an inclusion criterion in defining *de novo* genes. The expression of sequences at the mRNA level may be confirmed individually through conventional techniques such as quantitative PCR, or globally through more modern techniques such as RNA sequencing (RNA-seq). Similarly, expression at the protein level can be determined with high confidence for individual proteins using techniques such as mass spectrometry or western blotting, while ribosome profiling (Ribo-seq) provides a global survey of translation in a given sample. Ideally, to confirm that the gene in question arose *de novo*, a lack of expression of the syntenic region of outgroup species would also be demonstrated [70].

Confirmation of gene expression is only one approach to infer function. Genetic approaches, where one seeks to detect a specific phenotype or change in fitness upon disruption of a particular sequence, are considered by some to be the gold standard [69]; however, for large-scale analyses of entire genomes, obtaining such evidence is often not feasible. Other experimental approaches, including screens for protein-protein and/or genetic interactions, may also be employed to confirm a biological effect for a particular *de novo* ORF. As more is learned about a particular locus, standard molecular biology techniques can be applied to dissect its specific cellular role. Alternatively, evolutionary approaches may be employed to infer

the existence of a molecular function from computationally-derived signatures of selection. In the case of TRGs, one common signature of selection is the ratio of nonsynonymous to synonymous substitutions (K_a/K_s ratio), calculated from different species from the same taxon. This ratio indicates that the sequence in question is either evolving neutrally, or under either positive or negative selection. Evolutionary biologists tend to view only those sequences under selective constraint as being functional in the strict sense of the word [68]. Similarly, in the case of species-specific genes, polymorphism data may be used to calculate a pN/pS ratio from different strains or populations of the focal species. Given that young, species-specific *de novo* genes lack deep conservation by definition, detecting such signatures can be difficult without a large number of sequenced strains/populations. An example of this can be seen in *Mus musculus*, where three very young *de novo* genes lack signatures of selection despite well-demonstrated physiological roles [71]. Other signatures of selection, such as the degree of nucleotide divergence within syntenic regions, conservation of ORF boundaries, or for protein-coding genes, a coding score based on nucleotide hexamer frequencies, have instead been employed [72]. Despite these and other challenges in the identification of *de novo* gene birth events, there is now abundant evidence indicating that the phenomenon is not simply possible, but has occurred in every lineage systematically examined thus far.

3 Prevalence of *de novo* gene birth

3.1 Estimates of *de novo* gene numbers

Estimates regarding the frequency of *de novo* gene birth and the number of *de novo* genes in various lineages vary widely and are highly dependent on methodology. Studies may identify *de novo* genes by phylostratigraphy/BLAST-based methods alone, or may employ a combination of computational techniques (see above), and may or may not assess experimental evidence for expression and/or biological role. Furthermore, genome-scale analyses may consider all or most ORFs in the genome, or may instead limit their analysis to already annotated genes.

The *D. melanogaster* lineage is illustrative of these differing approaches. An early survey using a combination of BLAST searches performed on cDNA sequences along with manual searches and synteny information identified 72 new genes specific to *D. melanogaster* and 59 new genes specific to three of the four species in the *D. melanogaster* species complex. This report found that only 2/72 (~2.8%) of *D. melanogaster*-specific new genes and 7/59 (~11.9%) of new genes specific to the species complex were derived *de novo* [65], with the remainder arising via duplication/retroposition. Similarly, an analysis of 195 young (<35 million years old) *D. melanogaster* genes identified from syntenic alignments found that only 16 had arisen *de novo* [63]. In contrast, an analysis focused on transcriptomic data from the testes of six *D. melanogaster* strains identified 106 fixed and 142 segregating *de novo* genes [64]. For many of these, ancestral ORFs were identified but were not expressed. Highlighting the differences between inter- and intra-species comparisons, a study in natural *Saccharomyces paradoxus* populations found that the number of *de novo* polypeptides identified more than doubled when considering intra-species diversity [73]. In primates, one early study identified 270 orphan genes (unique to humans, chimpanzees, and macaques), of which 15 were thought to have originated *de novo* [44], while a later report identified 60 de novo genes in humans alone that are supported by transcriptional and proteomic evidence [66]. Studies in other lineages/organisms have also reached different conclusions with respect to the number of *de novo* genes present in each organism, as well as the specific sets of genes identified. A sample of these large-scale studies is described in **Table 1**.

A reanalysis of three such studies in murines that identified between 69 and 773 candidate *de novo* genes argued that the various estimates included many genes that were not in fact *de*

novo genes [74]. Many candidates were excluded on the basis of no longer being annotated in the major databases. A conservative approach was applied to the remaining genes, which excluded candidates with paralogs, distantly related homologs or conserved domains, or that lacked syntenic sequence information in non-rodents. This approach validated ~40% of candidate *de novo* genes, resulting in an upper estimate of only 11.6 *de novo* genes formed (and retained) per million years, a rate ~5–10 times slower than what was estimated for novel genes formed by duplication [74]. It is notable that even after application of this stringent pipeline, the 152 validated *de novo* genes that remained still represents a significant fraction of the mouse genome likely to have originated *de novo*. Generally speaking, however, it remains debated whether duplication and divergence or *de novo* gene birth represent the dominant mechanism for the emergence of new genes [63, 65, 73, 75–77], in part due to the fact that *de novo* genes are likely both to emerge and to be lost more frequently than other young genes (see below).

3.2. Dynamics of *de novo* gene birth

It is important to distinguish between the frequency of *de novo* gene birth and the number of *de novo* genes in a given lineage. If *de novo* gene birth is frequent, it might be expected that genomes would tend to grow in their gene content over time; however, the gene content of genomes is usually relatively stable [6]. This implies that a frequent gene death process must balance *de novo* gene birth, and indeed, *de novo* genes are distinguished by their rapid turnover relative to established genes. In support of this notion, recently emerged *Drosophila* genes are much more likely to be lost, primarily through pseudogenization, with the youngest orphans being lost at the highest rate [78]; this despite the fact that some *Drosophila* orphan genes have been shown to rapidly become essential [63]. A similar trend of frequent loss among young gene families was observed in nematode genus *Pristionchus [79]*. In wild *S. paradoxus* populations, *de novo* ORFs emerge and are lost at similar rates [73]. Similarly, an analysis of five mammalian transcriptomes found that most ORFs in mice were either very old or species specific, implying frequent birth and death of *de novo* transcripts [77]. Nevertheless, there remains a positive correlation between the number of species-specific genes in a genome and the evolutionary distance from its most recent ancestor [80]. In addition to the birth and death of *de novo* genes at the level of the ORF, mutational and other processes also subject genomes to constant "transcriptional turnover". One study in murines found that while all regions of the ancestral genome were transcribed at some point in at least one descendent, the portion of the genome under active transcription in a given strain or subspecies is subject to rapid change [81]. The "transcriptional turnover" of noncoding RNA genes is particularly fast as compared to that of coding genes [82].

4 Features of *de novo* genes

Recently emerged *de novo* genes differ from established genes in a number of ways. Across a broad range of species, young and/or taxonomically restricted genes or ORFs have been reported to be shorter in length than established genes, to evolve more rapidly, and to be less expressed [44, 75, 78, 79, 83–90]. Some of these reports, however, may have been partially influenced by the choice of homology-detection methods (see Genomic phylostratigraphy section). Their expression has also been found to be more tissue- or condition-specific than that of established genes [26, 28, 44, 64, 66, 75, 88, 91–93]. In particular, relatively high expression of *de novo* genes was observed in male reproductive tissues in *Drosophila*, mice, and humans (see below), and, in humans, in the cerebral cortex or the brain more generally [66, 94]. In animals with adaptive immune systems, higher expression in the brain and testes may at least in

part be a function of the immune-privileged nature of these tissues. An analysis in mice found specific expression of intergenic transcripts in the thymus and spleen (in addition to the brain and testes), and it has been proposed that in vertebrates *de novo* transcripts must first be expressed in these tissues before they can be expressed in tissues subject to surveillance by immune cells [93].

4.1 Lineage-dependent features

Other general features of *de novo* genes appear dependent on the species or lineage being examined. This appears to partly be a result of the fact that genomes vary in their GC content, and young genes bear more similarity to non-genic sequences from the genome in which they arose than do established genes [95]. Features such as predicted intrinsic structural disorder (ISD), the percentage of transmembrane residues, and the relative frequency of various predicted secondary structural features all show a strong GC dependency in orphan genes, whereas in more ancient genes these features are only weakly influenced by GC content [95]. This is exemplified by the fact that in organisms with relatively high GC content, ranging from *D. melanogaster* to the parasite *Leishmania major*, young genes have high ISD [96, 97], while in a low GC genome such as budding yeast, young genes have low ISD [75, 83, 90, 95]. It is noteworthy, however, that the most ancestral budding yeast genes display smaller ISD than genes of intermediate age [75, 98].

4.2 Role of epigenetic modifications

An examination of *de novo* genes in *A. thaliana* found that they are both hypermethylated and generally depleted of histone modifications *[62]*. In agreement with the proto-gene model (see below), methylation levels of *de novo* genes were intermediate between established genes and intergenic regions. The methylation patterns of these *de novo* genes are stably inherited, and methylation levels were highest, and most similar to established genes, in *de novo* genes with verified protein-coding ability [62]. In the pathogenic fungus *Magnaporthe oryzae*, less conserved genes tend to have methylation patterns associated with low levels of transcription [99]. A study in yeasts also found that *de novo* genes are enriched at recombination hotspots, which tend to be nucleosome-free regions [90].

In *Pristionchus pacificus*, orphan genes with confirmed expression display chromatin states that differ from those of similarly expressed established genes [89]. Orphan gene start sites have epigenetic signatures that are characteristic of enhancers, in contrast to conserved genes that exhibit classical promoters [89]. Many unexpressed orphan genes are decorated with repressive histone modifications, while a lack of such modifications facilitates transcription of an expressed subset of orphans, supporting the notion that open chromatin promotes the formation of novel genes [89].

5 Models and mechanisms of *de novo* gene birth

Several theoretical models and possible mechanisms of *de novo* gene birth have been described. The models are generally not mutually exclusive, and it is possible to imagine a number of plausible ways in which a *de novo* gene might emerge.

5.1 Order of events

5.1.1 ORF first vs. transcription first. For birth of a *de novo* protein-coding gene to occur, a non-genic sequence must both be transcribed and acquire an ORF before becoming translated (**Fig 1A**). These events may in theory occur in either order, and there is evidence

supporting both an "ORF first" and a "transcription first" model [2]. An analysis of *de novo* genes that are segregating in *D. melanogaster* with respect to their expression found that sequences that are transcribed had similar coding potential to the orthologous sequences from lines lacking evidence of transcription [64], supporting the notion that many ORFs, at least, exist prior to being expressed. The antifreeze glycoprotein gene *AFGP*, which emerged *de novo* in Arctic codfishes, provides a more definitive example in which the *de novo* emergence of the ORF was shown to precede that of the promoter region [100]. Furthermore, putatively non-genic ORFs long enough to encode functional peptides are numerous in eukaryotic genomes, and expected to occur at high frequency by chance [64, 75]. At the same time, transcription of eukaryotic genomes is far more extensive than previously thought, and documented examples also exist of genomic regions that were transcribed prior to the appearance of an ORF that became a *de novo* gene [101]. The proportion of *de novo* genes that are protein-coding is unknown, but the appearance of "transcription first" has led some to posit that protein-coding *de novo* genes may first exist as RNA gene intermediates. The case of bifunctional RNAs, which are both translated and function as RNA genes, shows that such a mechanism is plausible [102].

5.1.2 "Out of Testis" hypothesis. An early case study of *de novo* gene birth, which identified five *de novo* genes in *D. melanogaster*, noted preferential expression of these genes in the testes [27], and several additional *de novo* genes were identified using transcriptomic data derived from the testes and male accessory glands of *D. yakuba* and *D. erecta* [26, 28] (see above). This was in keeping with the rapid evolution of genes related to reproduction that has been observed across a range of lineages [103–105], suggesting that sexual selection may play a key role in adaptive evolution and *de novo* gene birth. A subsequent large-scale analysis of six *D. melanogaster* strains identified 248 testis-expressed *de novo* genes, of which ~57% were not fixed [64]. It has been suggested that the large number of *de novo* genes with male-specific expression identified in *Drosophila* is likely due to the fact that such genes are preferentially retained relative to other *de novo* genes, for reasons that are not entirely clear [78]. Interestingly, two putative *de novo* genes in *Drosophila* (*Goddard* and *Saturn*) were shown to be required for normal male fertility [106].

In humans, a study that identified 60 human-specific *de novo* genes found that their average expression, as measured by RNA-seq, was highest in the testes [66]. Another study looking at mammalian-specific genes more generally also found enriched expression in the testes [107]. Transcription in mammalian testes is thought to be particularly promiscuous, due in part to elevated expression of the transcription machinery [108, 109] and an open chromatin environment [110]. Along with the immune-privileged nature of the testes (see above), this promiscuous transcription is thought to create the ideal conditions for the expression of non-genic sequences required for *de novo* gene birth. Testes-specific expression seems to be a general feature of all novel genes, as an analysis of *Drosophila* and vertebrate species found that young genes showed testes-biased expression regardless of their mechanism of origination [91].

5.2 Pervasive expression

With the development and wide use of technologies such as RNA-seq and Ribo-seq, eukaryotic genomes are now known to be pervasively transcribed [111–114] and translated [115]. Many ORFs that are either unannotated, or annotated as long non-coding RNAs (lncRNAs), are translated at some level, under at least some condition, or in a particular tissue [75, 115–118]. Though infrequent, these translation events expose non-genic sequence to selection. This pervasive expression forms the basis for several theoretical models describing *de novo* gene birth.

It has been speculated that the epigenetic landscape of *de novo* genes in the early stages of formation may be particularly variable between and among populations, resulting in variable levels of gene expression and thereby allowing young genes to explore the "expression landscape" [119]. The *QQS* gene in *A*. *thaliana* is one example of this phenomenon; its expression is negatively regulated by DNA methylation that, while heritable for several generations, varies widely in its levels both among natural accessions and within wild populations [119]. Epigenetics are also largely responsible for the permissive transcriptional environment in the testes, particularly through the incorporation into nucleosomes of non-canonical histone variants that are replaced by histone-like protamines during spermatogenesis [120].

5.2.1 Proto-gene model. The proto-gene model proposes that de novo gene birth is mediated by a reservoir of "proto-genes" generated by pervasive expression of non-genic sequences [75]. It asserts that some of the proto-genes thereby exposed to the action of natural selection are occasionally retained and subsequently evolve the characteristics of genes. Proto-genes are thus expected to exhibit features intermediate between genes and non-genes. This model considers the genome as a spectrum ranging from non-genic to genic sequences, as opposed to the conventional binary classification scheme of gene vs. non-gene. The model makes use of the observation that in *S*. *cerevisiae*, several features of ORFs (see above) correlate with ORF age as determined by phylostratigraphic analysis [75]. A similar continuum with respect to gene age was seen for ORF features in a wide range of organisms (see above).

Most non-genic ORFs that are translated appear to be evolving neutrally [73, 75, 116]. The proto-gene model predicts, however, that expression of non-genic ORFs will occasionally provide an adaptive advantage to the cell. Adaptive proto-genes will gradually mature under selection, eventually leading to *de novo* gene birth. Differential translation of proto-genes in stress conditions, as well as an enrichment near proto-genes of binding sites for transcription factors involved in regulating stress response [75], support the adaptive potential of proto-genes. Furthermore, it is known that novel, functional proteins can be experimentally evolved from random amino acid sequences [121]. Random sequences are generally well-tolerated *in vivo*; many readily form secondary structures, and even highly disordered proteins may take on important biological roles [122–124]. The pervasive nature of translation suggests that new proto-genes emerge frequently, usually returning to the non-genic state.

Consistent with the notion that various features of ORFs exhibit a continuum that reflects their evolutionary age, a subsequent analysis, also in *S*. *cerevisiae*, found that ORF regulation by transcription factors, indicative of their integration into larger molecular networks, displays a similar continuum. Similarly, the likelihood of physical interactions, as well as the likelihood and strength of genetic interactions, is correlated with ORF age as determined by phylostratigraphy [125]. In contrast, with respect to certain predicted structural features such as β-strand content and aggregation propensity, the putative peptides encoded by proto-genes are similar to non-genic sequences and categorically distinct from canonical genes [125].

5.2.2 Preadaptation model. The preadaptation model of *de novo* gene birth uses mathematical modeling to argue that when standing genetic variation that is normally hidden is exposed to weak or shielded selection, the resulting pool of "cryptic" variation is purged of "self-evidently deleterious" sequences, such as those prone to lead to protein aggregation, and enriched in potential adaptations relative to completely non-expressed sequences [126]. This revealing of cryptic variation and purging of deleterious non-genic sequences, which may be considered as proto-genes under the above model, is a byproduct of pervasive transcription and translation of intergenic sequences [118]. Beyond such purging, selection is thought to operate on non-genic sequences that already contain gene-like properties. Using the evolutionary definition of function (i.e. a gene is by definition under purifying selection), the preadaptation model asserts that "gene birth is a sudden transition to functionality [98]" that occurs as

soon as an ORF acquires a selected effect. In contrast to the proto-gene model, recently emerged genes are expected to display exaggerated genic features, rather than features intermediate between old genes and non-genes [98]. In support of this, an analysis of ISD in mice found that young genes have higher ISD than old genes, while random non-genic sequences tend to show the lowest levels of ISD [98]. Although the observed trend may have partly resulted from a subset of young genes derived by overprinting [74], higher ISD in young genes was also seen among overlapping gene pairs [127]. Whether this trend holds over shorter timescales is debated [77, 128]. In wild *S. paradoxus* populations, ORFs with exaggerated gene-like features are found among the pool of translated intergenic polypeptides [73]. It is not clear whether such ORFs are preferentially retained.

The preadaptation model also proposes that in order to avoid the deleterious consequences associated with molecular errors, populations may either evolve local solutions, in which selection operates on each individual locus and a relatively high error rate is maintained, or global solutions that select for a low error rate and permit the accumulation of deleterious cryptic variation [126]. *De novo* gene birth is thought to be favored in populations that evolve local solutions, as the relatively high error rate will result in a pool of cryptic variation that is "preadapted" through the purging of deleterious sequences.

5.2.3 Grow slow and moult model. The "grow slow and moult" model describes a potential mechanism of *de novo* gene birth, particular to protein-coding genes. In this scenario, existing protein-coding ORFs expand at their ends, especially their 3' ends, leading to the creation of novel N- and C-terminal domains [129]. Novel C-terminal domains may first evolve under weak selection via occasional expression through read-through translation, as in the preadaptation model, only later becoming constitutively expressed through a mutation that disrupts the stop codon [126, 129]. Genes experiencing high translational readthrough tend to have intrinsically disordered C-termini [130]. Furthermore, existing genes are often close to repetitive sequences that encode disordered domains. These novel, disordered domains may initially confer some non-specific binding capability that becomes gradually refined by selection. Sequences encoding these novel domains may occasionally separate from their parent ORF, leading or contributing to the creation of a *de novo* gene [129]. Interestingly, an analysis of 32 insect genomes found that novel domains (i.e. those unique to insects) tend to evolve fairly neutrally, with only a few sites under positive selection, while their host proteins remain under purifying selection, suggesting that functional new domains emerge gradually and somewhat stochastically [131].

6 *De novo* gene birth and human health

In addition to its significance for the field of evolutionary biology, *de novo* gene birth has implications for human health. It has been speculated that novel genes, including *de novo* genes, may play an outsized role in species-specific traits [6, 37, 132]; however, many species-specific genes lack functional annotation [107]. Nevertheless, there is evidence to suggest that human-specific *de novo* genes are involved in disease processes such as cancer. *NYCM*, a *de novo* gene unique to humans and chimpanzees, regulates the pathogenesis of neuroblastomas in mouse models [133], and the primate-specific *PART1*, an lncRNA gene, has been identified as both a tumor suppressor and an oncogene in different contexts [44, 134, 135]. Several other human- or primate-specific *de novo* genes, including *PBOV1 [136]*, *GR6 [137, 138]*, *MYEOV [139]*, *ELFN1-AS1 [140],* and *CLLU1 [45]*, are also linked to cancer. Some have even suggested considering tumor-specifically expressed, evolutionary novel genes as their own class of genetic elements, noting that many such genes are under positive selection and may be neofunctionalized in the context of tumors [140].

The specific expression of many *de novo* genes in the human brain [66] also raises the intriguing possibility that *de novo* genes influence human cognitive traits. One such example is *FLJ33706*, a *de novo* gene that was identified in GWAS and linkage analyses for nicotine addiction and shows elevated expression in the brains of Alzheimer's patients [141]. Generally speaking, expression of young, primate-specific genes is enriched in the fetal human brain relative to the expression of similarly young genes in the mouse brain [142]. Most of these young genes, several of which originated *de novo*, are expressed in the neocortex, which is thought to be responsible for many aspects of human-specific cognition. Many of these young genes show signatures of positive selection, and functional annotations indicate that they are involved in diverse molecular processes, and are specifically enriched for genes involved in transcriptional regulation relative to other functional classes [142].

In addition to their roles in cancer processes, *de novo* originated human genes have been implicated in the maintenance of pluripotency [143] and in immune function [44, 107, 144]. The preferential expression of *de novo* genes in the testes (see above) is also suggestive of a role in reproduction. Given that the function of many *de novo* human genes remains uncharacterized, it seems likely that an appreciation of their contribution to human health and development will continue to grow.

Supporting information

S1 Text. Version history of the text file.
(XML)

S2 Text. Peer reviews and response to reviews.
(XML)

Acknowledgments

We thank Aaron Wacholder for his careful reading of and feedback on draft manuscripts.

References

1. Schmitz JF, Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. F1000Res. 2017; 6:57. https://doi.org/10.12688/f1000research.10079.1 PMID: 28163910; PubMed Central PMCID: PMCPMC5247788.
2. Schlotterer C. Genes from scratch—the evolutionary fate of de novo genes. Trends Genet. 2015; 31 (4):215–9. https://doi.org/10.1016/j.tig.2015.02.007 PMID: 25773713; PubMed Central PMCID: PMCPMC4383367.
3. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010; 20 (10):1313–26. https://doi.org/10.1101/gr.101386.109 PMID: 20651121; PubMed Central PMCID: PMCPMC2945180.
4. Jacob F. Evolution and tinkering. Science. 1977; 196(4295):1161–6. PMID: 860134.
5. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009; 25(9):404–13. Epub 2009/09/01. https://doi.org/10.1016/j.tig.2009.07.006 PMID: 19716618.
6. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011; 12 (10):692–702. https://doi.org/10.1038/nrg3053 PMID: 21878963.
7. Long M, Betran E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003; 4(11):865–75. Epub 2003/11/25. https://doi.org/10.1038/nrg1204 PMID: 14634634.
8. Wang W, Yu H, Long M. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. Nat Genet. 2004; 36(5):523–7. Epub 2004/04/06. https://doi.org/10.1038/ng1338 PMID: 15064762.
9. Ohno S. Evolution by gene duplication. London, New York,: Allen & Unwin; Springer-Verlag; 1970. xv, 160 p. p.

10. Tautz D. The discovery of de novo gene evolution. Perspect Biol Med. 2014; 57(1):149–61. Epub 2014/10/28. https://doi.org/10.1353/pbm.2014.0006 PMID: 25345708.

11. Grassé P-P. Evolution of living organisms: evidence for a new theory of transformation. New York: Academic Press; 1977. x, 297 p. p.

12. Barrell BG, Air GM, Hutchison CA, 3rd. Overlapping genes in bacteriophage phiX174. Nature. 1976; 264(5581):34–41. PMID: 1004533.

13. Shaw DC, Walker JE, Northrop FD, Barrell BG, Godson GN, Fiddes JC. Gene K, a new overlapping gene in bacteriophage G4. Nature. 1978; 272(5653):510–5. PMID: 692656.

14. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977; 265(5596):687–95. PMID: 870828.

15. Keese PK, Gibbs A. Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci U S A. 1992; 89(20):9489–93. https://doi.org/10.1073/pnas.89.20.9489 PMID: 1329098; PubMed Central PMCID: PMCPMC50157.

16. Ohno S. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci U S A. 1984; 81(8):2421–5. https://doi.org/10.1073/pnas.81.8.2421 PMID: 6585807; PubMed Central PMCID: PMCPMC345072.

17. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012; 29(12):3767–80. https://doi.org/10.1093/molbev/mss179 PMID: 22821011; PubMed Central PMCID: PMCPMC3494269.

18. Makalowska I, Lin CF, Hernandez K. Birth and death of gene overlaps in vertebrates. BMC Evol Biol. 2007; 7:193. https://doi.org/10.1186/1471-2148-7-193 PMID: 17939861; PubMed Central PMCID: PMCPMC2151771.

19. Samandi S, Roy AV, Delcourt V, Lucier JF, Gagnon J, Beaudoin MC, et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. Elife. 2017; 6. Epub 2017/10/31. https://doi.org/10.7554/eLife.27860 PMID: 29083303; PubMed Central PMCID: PMCPMC5703645.

20. Makalowski W, Mitchell GA, Labuda D. Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 1994; 10(6):188–93. Epub 1994/06/01. PMID: 8073532.

21. Sorek R. The birth of new exons: mechanisms and evolutionary consequences. RNA. 2007; 13 (10):1603–8. Epub 2007/08/22. https://doi.org/10.1261/rna.682507 PMID: 17709368; PubMed Central PMCID: PMCPMC1986822.

22. Dorit RL, Gilbert W. The limited universe of exons. Curr Opin Genet Dev. 1991; 1(4):464–9. PMID: 1822278.

23. Chothia C. Proteins. One thousand families for the molecular biologist. Nature. 1992; 357(6379):543–4. Epub 1992/06/18. https://doi.org/10.1038/357543a0 PMID: 1608464.

24. Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. The complete DNA sequence of yeast chromosome III. Nature. 1992; 357(6373):38–46. https://doi.org/10.1038/357038a0 PMID: 1574125.

25. Dujon B. The yeast genome project: what did we learn? Trends Genet. 1996; 12(7):263–70. PMID: 8763498.

26. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics. 2007; 176(2):1131–7. https://doi.org/10.1534/genetics.106.069245 PubMed Central PMCID: PMCPMC1894579. PMID: 17435230

27. 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):9935–9. https://doi.org/10.1073/pnas.0509809103 PubMed Central PMCID: PMCPMC1502557. PMID: 16777968

28. Begun DJ, Lindfors HA, Thompson ME, Holloway AK. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics. 2006; 172 (3):1675–81. https://doi.org/10.1534/genetics.105.050336 PubMed Central PMCID: PMCPMC1456303. PMID: 16361246

29. Betran E, Long M. Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics. 2003; 164(3):977–88. PubMed Central PMCID: PMCPMC1462638. PMID: 12871908

30. Jones CD, Begun DJ. Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci U S A. 2005; 102 (32):11373–8. https://doi.org/10.1073/pnas.0503528102 PubMed Central PMCID: PMCPMC1183565. PMID: 16076957

31. Long M, Langley CH. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science. 1993; 260(5104):91–5. PMID: 7682012.

32. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol. 2007; 5(5):e106. Epub 2007/04/19. https://doi.org/10.1371/journal.pbio.0050106 PubMed Central PMCID: PMCPMC1852585. PMID: 17439302

33. Hsu PY, Benfey PN. Small but Mighty: Functional Peptides Encoded by Small ORFs in Plants. Proteomics. 2018; 18(10):e1700038. Epub 2017/08/02. https://doi.org/10.1002/pmic.201700038 PMID: 28759167.

34. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science. 2016; 351(6270):271–5. Epub 2016/01/28. https://doi.org/10.1126/science.aad4076 PubMed Central PMCID: PMCPMC4892890. PMID: 26816378

35. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 2014; 15(3):193–204. Epub 2014/02/12. https://doi.org/10.1038/nrg3520 PMID: 24514441.

36. Nishida H. Detection and characterization of fungal-specific proteins in Saccharomyces cerevisiae. Biosci Biotechnol Biochem. 2006; 70(11):2646–52. https://doi.org/10.1271/bbb.60251 PMID: 17090923.

37. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci. 2015; 370 (1678):20140332. https://doi.org/10.1098/rstb.2014.0332 PMID: 26323763; PubMed Central PMCID: PMCPMC4571571.

38. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 2008; 179(1):487–96. https://doi.org/10.1534/genetics.107.084491 PMID: 18493065; PubMed Central PMCID: PMCPMC2390625.

39. Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, et al. Foldability of a Natural De Novo Evolved Protein. Structure. 2017; 25(11):1687–96 e4. https://doi.org/10.1016/j.str.2017.09.006 PMID: 29033289; PubMed Central PMCID: PMCPMC5677532.

40. Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. 2010; 20 (4):408–20. https://doi.org/10.1038/cr.2010.31 PMID: 20195295.

41. Li D, Yan Z, Lu L, Jiang H, Wang W. Pleiotropy of the de novo-originated gene MDF1. Sci Rep. 2014; 4:7280. https://doi.org/10.1038/srep07280 PMID: 25452167; PubMed Central PMCID: PMCPMC4250933.

42. Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, et al. Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. Plant J. 2009; 58(3):485–98. https://doi.org/10.1111/j.1365-313X.2009.03793.x PMID: 19154206.

43. Heinen TJ, Staubach F, Haming D, Tautz D. Emergence of a new gene from an intergenic region. Curr Biol. 2009; 19(18):1527–31. https://doi.org/10.1016/j.cub.2009.07.049 PMID: 19733073.

44. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, et al. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 2009; 26(3):603–12. https://doi.org/10.1093/molbev/msn281 PMID: 19064677.

45. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009; 19(10):1752–9. https://doi.org/10.1101/gr.095026.109 PMID: 19726446; PubMed Central PMCID: PMCPMC2765279.

46. Domazet-Loso T, Brajkovic J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 2007; 23(11):533–9. https://doi.org/10.1016/j.tig.2007.08.014 PMID: 18029048.

47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712.

48. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016; 17(9):567–78. https://doi.org/10.1038/nrg.2016.78 PMID: 27452112.

49. Elhaik E, Sabath N, Graur D. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 2006; 23(1):1–3. https://doi.org/10.1093/molbev/msj006 PMID: 16151190.

50. Alba MM, Castresana J. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 2007; 7:53. https://doi.org/10.1186/1471-2148-7-53 PMID: 17408474; PubMed Central PMCID: PMCPMC1855329.

51. Moyers BA, Zhang J. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. Mol Biol Evol. 2016; 33(5):1245–56. https://doi.org/10.1093/molbev/msw008 PMID: 26758516; PubMed Central PMCID: PMCPMC5010002.

52. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 2015; 32(1):258–67. https://doi.org/10.1093/molbev/msu286 PMID: 25312911; PubMed Central PMCID: PMCPMC4271527.

53. Domazet-Loso T, Carvunis AR, Alba MM, Sestak MS, Bakaric R, Neme R, et al. No Evidence for Phylostratigraphic Bias Impacting Inferences on Patterns of Gene Emergence and Evolution. Mol Biol Evol. 2017; 34(4):843–56. https://doi.org/10.1093/molbev/msw284 PMID: 28087778; PubMed Central PMCID: PMCPMC5400388.

54. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. Epub 1997/09/01. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694; PubMed Central PMCID: PMCPMC146917.

55. Saripella GV, Sonnhammer EL, Forslund K. Benchmarking the next generation of homology inference tools. Bioinformatics. 2016; 32(17):2636–41. https://doi.org/10.1093/bioinformatics/btw305 PMID: 27256311; PubMed Central PMCID: PMCPMC5013910.

56. Ghiurcuta CG, Moret BM. Evaluating synteny for improved comparative studies. Bioinformatics. 2014; 30(12):i9–18. https://doi.org/10.1093/bioinformatics/btu259 PMID: 24932010; PubMed Central PMCID: PMCPMC4058928.

57. Gehrmann T, Reinders MJ. Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. Bioinformatics. 2015; 31(21):3437–44. https://doi.org/10.1093/bioinformatics/btv389 PMID: 26116928; PubMed Central PMCID: PMCPMC4612220.

58. Jean G, Nikolski M. SyDiG: uncovering Synteny in Distant Genomes. Int J Bioinform Res Appl. 2011; 7(1):43–62. https://doi.org/10.1504/IJBRA.2011.039169 PMID: 21441096.

59. 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 PMID: 29382321; PubMed Central PMCID: PMCPMC5791376.

60. 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):230–9. Epub 2001/02/07. https://doi.org/10.1101/gr.162901 PMID: 11157786; PubMed Central PMCID: PMCPMC311025.

61. Lu TC, Leu JY, Lin WC. A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. Mol Biol Evol. 2017; 34(11):2823–38. https://doi.org/10.1093/molbev/msx210 PMID: 28981695; PubMed Central PMCID: PMCPMC5850716.

62. Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, et al. On the Origin of De Novo Genes in Arabidopsis thaliana Populations. Genome Biol Evol. 2016; 8(7):2190–202. https://doi.org/10.1093/gbe/evw164 PMID: 27401176; PubMed Central PMCID: PMCPMC4987118.

63. Chen S, Zhang YE, Long M. New genes in Drosophila quickly become essential. Science. 2010; 330 (6011):1682–5. https://doi.org/10.1126/science.1196380 PMID: 21164016.

64. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 2014; 343(6172):769–72. https://doi.org/10.1126/science.1248286 PMID: 24457212; PubMed Central PMCID: PMCPMC4391638.

65. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, et al. On the origin of new genes in Drosophila. Genome Res. 2008; 18(9):1446–55. https://doi.org/10.1101/gr.076588.108 PMID: 18550802; PubMed Central PMCID: PMCPMC2527705.

66. Wu DD, Irwin DM, Zhang YP. De novo origin of human protein-coding genes. PLoS Genet. 2011; 7 (11):e1002379. https://doi.org/10.1371/journal.pgen.1002379 PMID: 22102831; PubMed Central PMCID: PMCPMC3213175.

67. Vakirlis N, McLysaght A. Computational Prediction of De Novo Emerged Protein-Coding Genes. Methods Mol Biol. 2019; 1851:63–81. https://doi.org/10.1007/978-1-4939-8736-8_4 PMID: 30298392.

68. Doolittle WF, Brunet TD, Linquist S, Gregory TR. Distinguishing between "function" and "effect" in genome biology. Genome Biol Evol. 2014; 6(5):1234–7. https://doi.org/10.1093/gbe/evu098 PMID: 24814287; PubMed Central PMCID: PMCPMC4041003.

69. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014; 111(17):6131–8. Epub 2014/04/23. https://doi.org/10.1073/pnas.1318948111 PMID: 24753594; PubMed Central PMCID: PMCPMC4035993.

70. Andersson DI, Jerlstrom-Hultqvist J, Nasvall J. Evolution of new functions de novo and from preexisting genes. Cold Spring Harb Perspect Biol. 2015; 7(6). https://doi.org/10.1101/cshperspect.a017996 PMID: 26032716; PubMed Central PMCID: PMCPMC4448608.

71. Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, et al. Studying the dawn of de novo gene emergence in mice reveals fast integration of new genes into functional networks. bioRxiv. 2019.

72. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R, et al. Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet. 2015; 11(12):e1005721. Epub 2016/01/01. https://doi.org/10.1371/journal.pgen.1005721 PMID: 26720152; PubMed Central PMCID: PMCPMC4697840.

73. Durand É, Gagnon-Arsenault I, Hatin I, Nielly-Thibault L, Namy O, Landry CR. The high turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. bioRxiv. 2018. https://doi.org/10.1101/329730

74. Casola C. From de novo to "de nono": most novel protein coding genes identified with phylostratigraphy represent old genes or recent duplicates. bioRxiv. 2018. https://doi.org/10.1101/287193

75. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012; 487(7407):370–4. https://doi.org/10.1038/nature11184 PMID: 22722833; PubMed Central PMCID: PMCPMC3401362.

76. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics. 2013; 14:117. https://doi.org/10.1186/1471-2164-14-117 PMID: 23433480; PubMed Central PMCID: PMCPMC3616865.

77. 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):1626–32. https://doi.org/10.1038/s41559-018-0639-7 PMID: 30201962.

78. Palmieri N, Kosiol C, Schlotterer C. The life cycle of Drosophila orphan genes. Elife. 2014; 3:e01311. https://doi.org/10.7554/eLife.01311 PMID: 24554240; PubMed Central PMCID: PMCPMC3927632.

79. Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rodelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in Pristionchus nematodes. Genome Res. 2018; 28(11):1664–74. Epub 2018/09/21. https://doi.org/10.1101/gr.234971.118 PMID: 30232197.

80. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013; 5(2):439–55. https://doi.org/10.1093/gbe/evt009 PMID: 23348040; PubMed Central PMCID: PMCPMC3590893.

81. Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife. 2016; 5:e09977. https://doi.org/10.7554/eLife.09977 PMID: 26836309; PubMed Central PMCID: PMCPMC4829534.

82. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 2012; 8(7):e1002841. https://doi.org/10.1371/journal.pgen.1002841 PMID: 22844254; PubMed Central PMCID: PMCPMC3406015.

83. Ekman D, Elofsson A. Identifying and quantifying orphan protein sequences in fungi. J Mol Biol. 2010; 396(2):396–405. https://doi.org/10.1016/j.jmb.2009.11.053 PMID: 19944701.

84. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 2003; 13(10):2213–9. https://doi.org/10.1101/gr.1311003 PMID: 14525923; PubMed Central PMCID: PMCPMC403679.

85. Guo WJ, Li P, Ling J, Ye SP. Significant comparative characteristics between orphan and nonorphan genes in the rice (Oryza sativa L.) genome. Comp Funct Genomics. 2007:21676. https://doi.org/10.1155/2007/21676 PMID: 18273382; PubMed Central PMCID: PMCPMC2216055.

86. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A. 2009; 106(18):7273–80. https://doi.org/10.1073/pnas.0901808106 PMID: 19351897; PubMed Central PMCID: PMCPMC2666616.

87. Sun W, Zhao XW, Zhang Z. Identification and evolution of the orphan genes in the domestic silkworm, Bombyx mori. FEBS Lett. 2015; 589(19 Pt B):2731–8. https://doi.org/10.1016/j.febslet.2015.08.008 PMID: 26296317.

88. Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC Evol Biol. 2011; 11:47. https://doi.org/10.1186/1471-2148-11-47 PMID: 21332978; PubMed Central PMCID: PMCPMC3049755.

89. Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Res. 2018; 28(11):1675–87. Epub 2018/09/21. https://doi.org/10.1101/gr.234872.118 PMID: 30232198.

90. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A Molecular Portrait of De Novo Genes in Yeasts. Mol Biol Evol. 2018; 35(3):631–45. https://doi.org/10.1093/molbev/msx315 PMID: 29220506; PubMed Central PMCID: PMCPMC5850487.

91. Zhou Q, Zhang J-y. On the regulatory evolution of new genes throughout their life history. bioRxiv. 2018. https://doi.org/10.1101/276667

92. Wu B, Knudson A. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. MBio. 2018; 9(4). Epub 2018/08/02. https://doi.org/10.1128/mBio.01024-18 PMID: 30065088; PubMed Central PMCID: PMCPMC6069113.

93. Bekpen C, Xie C, Tautz D. Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. BMC Evol Biol. 2018; 18(1):121. Epub 2018/08/05. https://doi.org/10.1186/s12862-018-1232-z PMID: 30075701; PubMed Central PMCID: PMCPMC6091031.

94. Pertea M, Shumate A, Pertea G, Varabyou A, Chang Y-C, Madugundu AK, et al. Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. bioRxiv. 2018. https://doi.org/10.1101/332825

95. Basile W, Sachenkova O, Light S, Elofsson A. High GC content causes orphan proteins to be intrinsically disordered. PLoS Comput Biol. 2017; 13(3):e1005375. https://doi.org/10.1371/journal.pcbi.1005375 PMID: 28355220; PubMed Central PMCID: PMCPMC5389847.

96. Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis". Biochimie. 2015; 119:244–53. https://doi.org/10.1016/j.biochi.2015.02.019 PMID: 25736992.

97. Mukherjee S, Panda A, Ghosh TC. Elucidating evolutionary features and functional implications of orphan genes in Leishmania major. Infect Genet Evol. 2015; 32:330–7. https://doi.org/10.1016/j.meegid.2015.03.031 PMID: 25843649.

98. Wilson BA, Foy SG, Neme R, Masel J. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. Nat Ecol Evol. 2017; 1(6):0146–146. https://doi.org/10.1038/s41559-017-0146 PMID: 28642936; PubMed Central PMCID: PMCPMC5476217.

99. Jeon J, Choi J, Lee GW, Park SY, Huh A, Dean RA, et al. Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, Magnaporthe oryzae. Sci Rep. 2015; 5:8567. https://doi.org/10.1038/srep08567 PMID: 25708804; PubMed Central PMCID: PMCPMC4338423.

100. Zhuang X, Yang C, Murphy KR, Cheng CC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proc Natl Acad Sci U S A. 2019. Epub 2019/02/16. https://doi.org/10.1073/pnas.1817138116 PMID: 30765531.

101. 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 PMID: 24146629; PubMed Central PMCID: PMCPMC3798262.

102. Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol. 2008; 4(11):e1000176. https://doi.org/10.1371/journal.pcbi.1000176 PMID: 19043537; PubMed Central PMCID: PMCPMC2518207.

103. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nat Rev Genet. 2002; 3(2):137–44. https://doi.org/10.1038/nrg733 PMID: 11836507.

104. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005; 437(7062):1153–7. https://doi.org/10.1038/nature04240 PMID: 16237444.

105. Clark NL, Aagaard JE, Swanson WJ. Evolution of reproductive proteins from animals and plants. Reproduction. 2006; 131(1):11–22. https://doi.org/10.1530/rep.1.00357 PMID: 16388004.

106. Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, et al. The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. Mol Biol Evol. 2017; 34(5):1066–82. Epub 2017/01/21. https://doi.org/10.1093/molbev/msx057 PMID: 28104747; PubMed Central PMCID: PMCPMC5400382.

107. Luis Villanueva-Canas J, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Alba MM. New Genes and Functional Innovation in Mammals. Genome Biol Evol. 2017; 9(7):1886–900. https://doi.org/10.1093/gbe/evx136 PMID: 28854603; PubMed Central PMCID: PMCPMC5554394.

108. Schmidt EE. Transcriptional promiscuity in testes. Curr Biol. 1996; 6(7):768–9. PMID: 8805310.

109. White-Cooper H, Davidson I. Unique aspects of transcription regulation in male germ cells. Cold Spring Harb Perspect Biol. 2011; 3(7). https://doi.org/10.1101/cshperspect.a002626 PMID: 21555408; PubMed Central PMCID: PMCPMC3119912.

110. Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev. 2001; 106(1–2):3–23. PMID: 11472831.

111. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, et al. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A. 2006; 103(14):5320–5. https://doi.org/10.1073/pnas.0601091103 PMID: 16569694; PubMed Central PMCID: PMCPMC1414796.

112. Tisseur M, Kwapisz M, Morillon A. Pervasive transcription—Lessons from yeast. Biochimie. 2011; 93 (11):1889–96. https://doi.org/10.1016/j.biochi.2011.07.001 PMID: 21771634.

113. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320(5881):1344–9. https://doi.org/10.1126/science.1158441 PMID: 18451266; PubMed Central PMCID: PMCPMC2951732.

114. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. PLoS Biol. 2011; 9(7):e1000625; discussion e1102. https://doi.org/10.1371/journal.pbio.1000625 PMID: 21765801; PubMed Central PMCID: PMCPMC3134446.

115. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014; 8 (5):1365–79. https://doi.org/10.1016/j.celrep.2014.07.045 PMID: 25159147; PubMed Central PMCID: PMCPMC4216110.

116. Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Canas JL, Messeguer X, Alba MM. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat Ecol Evol. 2018; 2(5):890–6. https://doi.org/10.1038/s41559-018-0506-6 PMID: 29556078.

117. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014; 3:e03523. https://doi.org/10.7554/eLife.03523 PMID: 25233276; PubMed Central PMCID: PMCPMC4359382.

118. Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol. 2011; 3:1245–52. https://doi.org/10.1093/gbe/evr099 PMID: 21948395; PubMed Central PMCID: PMCPMC3209793.

119. Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LE, Loudet O, et al. Extensive natural epigenetic variation at a de novo originated gene. PLoS Genet. 2013; 9(4):e1003437. https://doi.org/10.1371/journal.pgen.1003437 PMID: 23593031; PubMed Central PMCID: PMCPMC3623765.

120. Kimmins S, Sassone-Corsi P. Chromatin remodelling and epigenetic features of germ cells. Nature. 2005; 434(7033):583–9. Epub 2005/04/01. https://doi.org/10.1038/nature03368 PMID: 15800613.

121. Keefe AD, Szostak JW. Functional proteins from a random-sequence library. Nature. 2001; 410 (6829):715–8. https://doi.org/10.1038/35070613 PMID: 11287961; PubMed Central PMCID: PMCPMC4476321.

122. Tretyachenko V, Vymetal J, Bednarova L, Kopecky V Jr., Hofbauerova K, Jindrova H, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci Rep. 2017; 7(1):15449. Epub 2017/11/15. https://doi.org/10.1038/s41598-017-15635-8 PMID: 29133927; PubMed Central PMCID: PMCPMC5684393.

123. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015; 16(1):18–29. Epub 2014/12/23. https://doi.org/10.1038/nrm3920 PMID: 25531225; PubMed Central PMCID: PMCPMC4405151.

124. Neme R, Amador C, Yildirim B, McConnell E, Tautz D. Random sequences are an abundant source of bioactive RNAs or peptides. Nat Ecol Evol. 2017; 1(6):0217. Epub 2017/06/06. https://doi.org/10.1038/s41559-017-0127 PMID: 28580432; PubMed Central PMCID: PMCPMC5447804.

125. Abrusan G. Integration of new genes into cellular networks, and their structural maturation. Genetics. 2013; 195(4):1407–17. https://doi.org/10.1534/genetics.113.152256 PMID: 24056411; PubMed Central PMCID: PMCPMC3832282.

126. Rajon E, Masel J. Evolution of molecular error rates and the consequences for evolvability. Proc Natl Acad Sci U S A. 2011; 108(3):1082–7. https://doi.org/10.1073/pnas.1012918108 PMID: 21199946; PubMed Central PMCID: PMCPMC3024668.

127. Willis S, Masel J. Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes. Genetics. 2018; 210(1):303–13. Epub 2018/07/22. https://doi.org/10.1534/genetics.118.301249 PMID: 30026186; PubMed Central PMCID: PMCPMC6116962.

128. Foy SG, Wilson BA, Bertram J, Cordes MHJ, Masel J. A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution. Genetics. 2019. Epub 2019/01/30. https://doi.org/10.1534/genetics.118.301719 PMID: 30692195.

129. Bornberg-Bauer E, Schmitz J, Heberlein M. Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult'. Biochem Soc Trans. 2015; 43(5):867–73. https://doi.org/10.1042/BST20150089 PMID: 26517896.

130. Kleppe AS, Bornberg-Bauer E. Robustness by intrinsically disordered C-termini and translational read-through. Nucleic Acids Res. 2018; 46(19):10184–94. Epub 2018/09/25. https://doi.org/10.1093/nar/gky778 PMID: 30247639.

131. Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. Origins and structural properties of novel and de novo protein domains during insect evolution. FEBS J. 2018; 285(14):2605–25. Epub 2018/05/29. https://doi.org/10.1111/febs.14504 PMID: 29802682.

132. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nat Rev Genet. 2013; 14 (9):645–60. https://doi.org/10.1038/nrg3521 PMID: 23949544; PubMed Central PMCID: PMCPMC4236023.

133. Suenaga Y, Islam SM, Alagu J, Kaneko Y, Kato M, Tanaka Y, et al. NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3beta resulting in the stabilization of MYCN in human neuroblastomas. PLoS Genet. 2014; 10(1):e1003996. https://doi.org/10.1371/journal.pgen.1003996 PMID: 24391509; PubMed Central PMCID: PMCPMC3879166.

134. Lin B, White JT, Ferguson C, Bumgarner R, Friedman C, Trask B, et al. PART-1: a novel human prostate-specific, androgen-regulated gene that maps to chromosome 5q12. Cancer Res. 2000; 60 (4):858–63. PMID: 10706094.

135. Kang M, Ren M, Li Y, Fu Y, Deng M, Li C. Exosome-mediated transfer of lncRNA PART1 induces gefitinib resistance in esophageal squamous cell carcinoma via functioning as a competing endogenous RNA. J Exp Clin Cancer Res. 2018; 37(1):171. https://doi.org/10.1186/s13046-018-0845-9 PMID: 30049286; PubMed Central PMCID: PMCPMC6063009.

136. Samusik N, Krukovskaya L, Meln I, Shilov E, Kozlov AP. PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. PLoS One. 2013; 8 (2):e56162. https://doi.org/10.1371/journal.pone.0056162 PMID: 23418531; PubMed Central PMCID: PMCPMC3572036.

137. Guerzoni D, McLysaght A. De Novo Genes Arise at a Slow but Steady Rate along the Primate Lineage and Have Been Subject to Incomplete Lineage Sorting. Genome Biol Evol. 2016; 8(4):1222–32. https://doi.org/10.1093/gbe/evw074 PMID: 27056411; PubMed Central PMCID: PMCPMC4860702.

138. Pekarsky Y, Rynditch A, Wieser R, Fonatsch C, Gardiner K. Activation of a novel gene in 3q21 and identification of intergenic fusion transcripts with ecotropic viral insertion site I in leukemia. Cancer Res. 1997; 57(18):3914–9. PMID: 9307271.

139. Papamichos SI, Margaritis D, Kotsianidis I. Adaptive Evolution Coupled with Retrotransposon Exaptation Allowed for the Generation of a Human-Protein-Specific Coding Gene That Promotes Cancer Cell Proliferation and Metastasis in Both Haematological Malignancies and Solid Tumours: The Extraordinary Case of MYEOV Gene. Scientifica (Cairo). 2015; 2015:984706. https://doi.org/10.1155/2015/984706 PMID: 26568894; PubMed Central PMCID: PMCPMC4629056.

140. Kozlov AP. Expression of evolutionarily novel genes in tumors. Infect Agent Cancer. 2016; 11:34. https://doi.org/10.1186/s13027-016-0077-6 PMID: 27437030; PubMed Central PMCID: PMCPMC4949931.

141. Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, et al. A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol. 2010; 6(3):e1000734. https://doi.org/10.1371/journal.pcbi.1000734 PMID: 20376170; PubMed Central PMCID: PMCPMC2845654.

142. Zhang YE, Landback P, Vibranovski MD, Long M. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 2011; 9(10):e1001179. https://doi.org/10.1371/journal.pbio.1001179 PMID: 22028629; PubMed Central PMCID: PMCPMC3196496.

143. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014; 516(7531):405–9. https://doi.org/10.1038/nature13804 PMID: 25317556.

144. Dolstra H, Fredrix H, Maas F, Coulie PG, Brasseur F, Mensink E, et al. A human minor histocompatibility antigen specific for B cell acute lymphoblastic leukemia. J Exp Med. 1999; 189(2):301–8. PMID: 9892612; PubMed Central PMCID: PMCPMC2192993.

145. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009; 37(Database issue):D211–5. Epub 2008/10/23. https://doi.org/10.1093/nar/gkn785 PMID: 18940856; PubMed Central PMCID: PMCPMC2686546.

146. Murphy DN, McLysaght A. De novo origin of protein-coding genes in murine rodents. PLoS One. 2012; 7(11):e48650. https://doi.org/10.1371/journal.pone.0048650 PMID: 23185269; PubMed Central PMCID: PMCPMC3504067.

147. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein diversity by de novo origination in Oryza. Nat Ecol Evol. 2019; 3(4):679–90. Epub 2019/03/13. https://doi.org/10.1038/s41559-019-0822-5 PMID: 30858588.