A Synergistic, Cultivator Model of De Novo Gene Origination

UnJin Lee, Shawn M. Mozeika, and Li Zhao *

Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY, USA

*Corresponding author: E-mail: lzhao@rockefeller.edu.

Accepted: May 12, 2024

Abstract

The origin and fixation of evolutionarily young genes is a fundamental question in evolutionary biology. However, understanding the origins of newly evolved genes arising de novo from noncoding genomic sequences is challenging. This is partly due to the low likelihood that several neutral or nearly neutral mutations fix prior to the appearance of an important novel molecular function. This issue is particularly exacerbated in large effective population sizes where the effect of drift is small. To address this problem, we propose a regulation-focused, cultivator model for de novo gene evolution. This cultivator-focused model posits that each step in a novel variant's evolutionary trajectory is driven by well-defined, selectively advantageous functions for the cultivator genes, rather than solely by the de novo genes, emphasizing the critical role of genome organization in the evolution of new genes.

Key words: de novo genes, genome organization, lncRNAs, linkage, natural selection.

Significance

The transformation of inactive genomic sequences into fully functional de novo protein-coding genes requires the accumulation of multiple mutations unlikely to arise solely through neutral processes in populations with large effective population size. Our research introduces a novel regulation-focused model of de novo gene evolution, termed the cultivator model. This model emphasizes the existence of well-defined, selectively advantageous functions at each step of a de novo gene's evolutionary trajectory, highlighting the pivotal role of genome organization in this process.

Making a New Gene

In addition to mutation-driven variations in preexisting genes or regulatory elements, one key source of phenotypic variation is the origination of new genes. There are many sources of new gene origination, including gene duplication caused by unequal crossing-over, horizontal gene transfer, domestication of transposable elements, overprinting, and chimerism (Kaessmann 2010; Long et al. 2013; Van Oss and Carvunis 2019). In general, these sources may be classified into two broad categories: the duplication of preexisting elements or the appearance of new elements de novo—i.e. the transformation of noncoding and/or nonfunctional genomic sequences into functional units. While the majority of the abovementioned cases fall under duplication-based mechanisms, the degree to which de novo origination produces genetic novelty is a standing question in evolutionary genes. This is primarily due to the fact that it is relatively easy to identify duplication events using homology-detection software, while conversely, it is intrinsically difficult to identify de novo origination events (Levine et al. 2006; Long et al. 2013; McLysaght and Hurst 2016). Unlike duplication-based events, where the presence of homology itself can be taken as positive evidence of duplication, de novo gene

origination relies on the *absence* of orthologous genes in the homologous regions. As such, a lack of rigor in homology-detection strategies can lead to inflated estimations of the number of de novo gene origination events (Moyers and Zhang 2016; Weisman et al. 2020). Some of the first empirical studies tackling this problem characterized not only the existence of lineage-specific (Levine et al. 2006) but even subpopulation-specific de novo genes (Zhao et al. 2014; Ruiz-Orera et al. 2015; Bekpen et al. 2018; Zhang et al. 2019), suggesting that de novo genes arise naturally at some nonnegligible frequency. Further work has widely characterized the existence of de novo genes across nearly all domains of life, as reviewed in McLysaght and Hurst (2016). Despite this, a large gap remains in understanding the evolutionary mechanisms by which de novo genes arise.

Genomic "Happy Accidents"

Key differentiators between inactive sequence and protein-coding genes under the central dogma are transcription, posttranscriptional modification, and translation. As such, the initiation of each of these highly regulated molecular processes represents a unique barrier that an emerging de novo protein-coding gene must overcome. Complete evolutionary models of de novo gene origination (including both RNA and protein-coding genes) should, therefore, provide meaningful insights into how noncoding sequences gain these properties, with particular emphasis on understanding the processes by which selectable variation and function arise, and subsequently become exposed to natural selection. Ideally, a comprehensive model of eukaryotic de novo gene origination should also explicitly consider the role of genome organization and local chromatin environment, providing a well-defined selective advantage for the accumulation of multiple substitutions.

In the absence of a well-defined evolutionary role, we must instead assume that newly originated de novo variants rely on the existence of large, neutrally evolving variations that possess the potential capacity to create some undescribed RNA or protein function. These de novo variants must then continue to traverse a series of dys- or nonfunctional mutational states until a new RNA or protein function appears. Finally, the potential for new function is subsequently fulfilled upon shifted internal (e.g. mutational) or external (e.g. ecological, environmental, or sexually selective) conditions allowing for fixation. The likelihood that variants could accumulate a large number of substitutions appears negligible, leading to a significant gap in understanding how new genes could arise de novo from noncoding sequences.

One particularly interesting class of studies tackles this evolvability issue directly by presupposing some selective biological function. A large library of randomly generated synthetic polypeptides is then screened for the ability of fulfill this function (Neme et al. 2017; Babina et al. 2023; Frumkin and Laub 2023). These studies are thus able to control for putative function by screening for increased reproductive capacity (Neme et al. 2017), the deactivation of a known toxin (*MazF*/*ramF*) (Frumkin and Laub 2023), or rescue of auxotrophy (Δ*serB*) (Babina et al. 2023), thereby empirically demonstrating at least the *possibility* for the evolution of such proteins de novo. However, these studies' removal of spontaneous mutations comes at the cost of losing the interaction between mutation and selection altogether. Given the significant size of the libraries used in these studies, questions remain regarding how likely it is that natural selection alone, particularly in the context of large populations where the effect of drift is minimized, could produce such a similarly large library of random polypeptides neutrally prior to selection. Further complications arise in the interpretation of these studies as the plasmids carrying these putative random protein-coding genes also contain specific upstream and downstream sequences, ensuring proper transcription and translation of these oligonucleotides—another barrier that naturally arising de novo genes must overcome.

As such, an additional barrier occurs when considering the transition of de novo genes between the transcription and translation stages of the central dogma. While the metabolic costs of promiscuous transcription may be relatively low under natural conditions, the high cellular cost of translation (Lynch and Marinov 2015) makes it unlikely that segments of random genomic sequence alone could generate sufficient levels of protein to present a selectable difference in phenotype without such flanking sequences. Instead, many sequences will likely be subject to negative selection without a clear posttranslational selective advantage. A single neutral fixation event may not be out of the question. However, it is unlikely to produce a large number of protein-coding genes through a series of de novo genomic neutral fixation events. Rather, the context of a new variant's surrounding genomic regions may provide a key clue in understanding the fixation of de novo variants. Hence, there is a need for a model that can comprehensively account for some selective function throughout the entire process of de novo gene origination while explicitly considering positional and configurational genomic organization, including inter-promoter distance and promoter orientation.

Prior Models of De Novo Gene Evolution

The existing theoretical models of de novo gene origination can be broadly categorized into two classes: proto-gene/open reading frame (ORF)-first models (Carvunis et al. 2012; Schlötterer 2015; Vakirlis et al. 2020) and

transcription-first models (Wu and Sharp 2013; Ruiz-Orera et al. 2014; Schlötterer 2015). The proto-gene/ORF-first models begin with the observation that all genomes (and even sequences of random nucleotides) contain a variety of ORFs of various lengths. Variations in these ORFs then accumulate, allowing for further differences in frame size and translational potential—those variants that share more properties with known protein-coding genes are considered "proto-genes" under these models. Promiscuous transcription and translation of these variants, particularly under periods of high stress in single-celled organisms, subsequently expose these proto-genes to selection, allowing them to increase in population frequency during these periods (Siepel 2009).

Alternatively, the class of transcription-first models begins with the observation that a large number of noncoding genomic elements undergo widespread transcription, particularly in the case of divergent (bidirectional) transcription of eukaryotic promoters (Wu and Sharp 2013) or widespread spurious transcription in mammals (Ruiz-Orera et al. 2014; Chen et al. 2015; An et al. 2023). These transcripts then accumulate a sequence of substitutions typical of protein-coding genes, such as the loss of early termination sequences or the gain of introns and poly-A signals. As these variants appear in the population, they then undergo translation, exposing these variants to selection and thereby increasing population frequency. The protein-coding and flanking regulatory sequences of these transcripts then accumulate more mutations, allowing these nascent transcripts to eventually transform into a fully formed de novo protein-coding gene.

A Stepwise, Cultivator Model of De Novo Gene Evolution

While these models provide key insights into overcoming the barriers preventing an inactive sequence from becoming a new gene de novo, we complement these models by presenting a regulation-focused, cultivator model of de novo gene evolution. The cultivator gene is a preexisting gene that acquires new or modified regulatory elements or mechanisms in *cis* or *trans* via natural selection. These novel or modified regulatory elements or mechanisms can facilitate the expression of a cultivator gene while also producing a new de novo transcript. The crucial factor is their overall closeness, ensuring the new gene's maintenance through linkage, although physical proximity is not necessarily needed. The term cultivator gene is used because it aids in the emergence and establishment of a de novo gene.

This model consists of three distinct steps that we summarize as follows. First, an easily accessible (but likely weakly) functioning mutation (1–2 mutations) alters the *cis*-regulation of a preexisting cultivator gene, which "cultivates" the appearance and refinement of a neighboring de novo gene. Importantly, this novel or modified regulation also results in the transcription of a functionally neutral *cis*-regulating long non-coding RNA (lncRNA) transcript ("*cis*-lncRNA" or "*cis*-lncRNA transcript") from this promoter/locus (the newly evolved de novo "*cis*-lncRNA gene") (cf. Table 1, supplementary table S1, Supplementary Material online for definitions). While selection acts on the older cultivator gene, the entire locus, comprising the cultivator gene and the selectively advantageous *cis*-regulatory element (i.e. the new *cis*-lncRNA gene) rise to fixation together through linkage. Second, as the highly evolvable but novel function of the *cis*-lncRNA is likely weakly functioning, this regulatory function could potentially be further refined. This refinement could result in the fixation of a newer *trans*-lncRNA gene that also possesses a *trans*-regulatory function (via a new *trans*-lncRNA transcript). Alternatively, the *trans*-lncRNA may accumulate mutations that allow it to regulate a separate cultivator gene(s) in a selectively beneficial manner. Finally, the regulatory function of the newly evolving *cis*- or *trans*-lncRNA gene is further refined, resulting in the origination of a de novo protein-coding gene. Importantly, each step is dependent on an explicitly defined function—specifically, the purpose of each step is to regulate the function of one or more preexisting "cultivator" genes. A notable side effect of this sequential refinement of function is the potential appearance of pleiotropic (in the case of multiple cultivator genes) or redundant (in the case of both *cis* and *trans* regulation of a single cultivator gene) functionality. A list of key definitions is provided in Table 1 and Supplementary Table S1, Supplementary Material online, a summary of the model is provided in Table 2, and a comparison between existing models and the cultivator model is provided in Table 3.

Several points regarding the model need to be raised. First, while a new gene or transcript may share a divergent promoter (Wu and Sharp 2013) with a cultivator gene, this is not always the case. A new gene and a cultivator may share a novel or modified enhancer, a novel transposable element insertion that contains regulatory sequences, other regulatory elements, or even more distant regulators. Although these genes are likely linked, they do not necessarily have to be adjacent. The key point is that the new gene and the cultivator are sufficiently close that linkage influences the transmission rate to subsequent generations. Second, while *cis*-lncRNA genes may affect the expression of cultivator genes by competing for regulatory elements or other mechanisms, *trans*-lncRNA genes do not have to function directly through the cultivator genes. Third, the origination of a protein-coding de novo gene may occur concurrently with or after the emergence of the *trans*-lncRNA function. Since our model focuses on the effects of cultivator genes, we do not limit the functional or

Table 1
Key terms and definitions

Term	Description
Gene	A genomic locus with an ascribable molecular function that produces meaningful levels of transcription, regardless of whether the resulting transcript undergoes translation (e.g., "RNA genes" or "protein coding genes"). Sequences containing genes are considered "genic" sequences, while those that do not are considered to be "nongenic" even if they contain genomic sequences with ascribable molecular function (e.g. an enhancer).
Promoter*	A genomic sequence that is necessary and sufficient to produce meaningful levels of transcription.
Enhancer*	A genomic sequence that is not sufficient to produce meaningful levels of transcription alone but instead modulates and/or regulates the expression of proximal promoter elements via sequence-specific recruitment of repressive or activating transcription factors.
De novo gene	A newly evolved gene that originates from nongenic sequences.
Cultivator gene	A gene that is proximal to a newly evolved de novo gene that predates the origination event for the said de novo gene. The appearance of the de novo gene's promoter in close proximity to the cultivator gene results in the alteration of the expression of the cultivator gene. Selection for altered expression of the cultivator gene thus facilitates retention of this new de novo gene promoter and its associated transcripts via linkage.
lncRNA	A transcript that does not possess protein-coding capacity. If not specified, used herein to refer to the transcripts produced by a lncRNA gene regardless of whether or not the transcript has any ascribable genomic function (i.e. "*cis*-lncRNA*" or "*trans*-lncRNA*")
lncRNA gene	A genomic locus that produces lncRNA transcripts resulting from the activity of a promoter element. Used herein to refer to the locus and not the transcript (i.e. "*cis*-lncRNA gene*" or "*trans*-lncRNA gene*").
cis*-regulation	Alteration of transcriptional activity of affected promoter element(s) located on the same continuous DNA segment.
trans*-regulation	Alteration of transcriptional activity of affected promoter elements(s) located on any DNA segment.
Evolvability*	The relative ease or difficulty of producing a given DNA sequence or phenotypic function within a population.
Easily accessible mutation	A mutation requiring only a small number (1 to 2) of alterations, including single nucleotide changes and/or single larger insertion events. These mutations are likelier to arise in a population than mutations that require a large number of alterations.
Weakly functional	A small alteration in phenotype yielding only a small selective advantage.

*indicates extended definition in supplementary table S1, Supplementary Material online.

selective impact of the lncRNA or protein-coding genes. Thus, both noncoding and protein-coding de novo genes may be born following this model. Last, by proposing this model, we do not reject the previous models of de novo gene origination, as they all share common synergistic elements. Instead, we focus on the effect of cultivators in the evolution and fixation of new genes.

Molecular Mechanisms of the Cultivator Model

To further illustrate the model, we consider a scenario with the assumption that natural selection begins to favor the altered expression of a particular trait. For simplicity, we will assume that this trait is under monogenic control by what we will call the cultivator gene, and that there is strong negative selection against any other alteration of the wild-type phenotype of this organism. One possible solution to these new selective conditions is the appearance of a promoter in the vicinity of the cultivator gene through mutation and/or the insertion of a preexisting promoter element. This promoter may appear in one of five possible promoter syntaxes, existing locally in (i) tandem upstream or (ii) tandem downstream configurations with the cultivator gene's promoter, in a (iii) locally convergent or (iv) divergent configuration, or even (v) in a distal region of the genome (Fig. 1). The regulation of the original cultivator gene is subsequently altered because of the regulatory interference by the newly evolved promoter, allowing it to regulate the cultivator in *cis*. Note that in this context, "regulation" is defined broadly, including but not limited to the action of RNA–DNA and protein–DNA interactions but also proximity- and supercoiling-based regulatory effects. For example, in addition to altering the binding affinity of transcription factor binding sites, lncRNAs can regulate or influence neighboring genes by competing for promoters or enhancers (Lin et al. 2018) or by modulating chromatin context (Statello et al. 2021) in *cis*. Regardless, a by-product of the new promoter is the appearance of a nascent *cis*-regulatory lncRNA gene whose *transcripts* are selectively neutral. Importantly, this *cis*-lncRNA gene will rise to fixation in linkage with the cultivator gene (Fig. 2A and B).

Subsequently, the *cis*-lncRNA gene may begin to accumulate mutations that could further refine its *cis*-regulatory function through regulatory archetypes described in Wang and Chang (2011). One possible outcome of the accumulation of mutations is the appearance of a variant that will allow the *cis*-lncRNA transcript to begin to regulate the cultivator gene either directly or further upstream in *trans*,

Table 2
Summary of cultivator model

Step of model	Detail	Observation
Origination and fixation of *cis*-lncRNA	*cis*-lncRNA genes are easily evolvable	• Promoters are easily evolvable in bacteria (Yona et al. 2018) • Enhancers may evolve into promoters (but may be rare events) (Carelli et al. 2018) • Transcriptional solutions to selection are ~10^3 times more easily evolvable than protein solutions (Frumkin and Laub 2023)
	New promoters may regulate cultivator genes	• Promoter interference can alter regulation of neighboring genes (Fukaya et al. 2016; Cho et al. 2018) • Supercoiling-mediated coupling can alter regulation of neighboring genes (Johnstone and Galloway 2022) • Read-through of promoters can alter genome structure (Ibragimov et al. 2023)
	Positioning of promoter can be necessary and sufficient for lncRNA gene function	• Deletion of the lncRNA *bxd* promoter (but not gene body) yields phenotype (Petruk et al. 2006; Pease et al. 2013) • Promoter deletion is rescued by inverted promoter (Petruk et al. 2006; Pease et al. 2013)
	Many lncRNA genes may potentially function in a distance-dependent manner	• lncRNA genes appear closer to protein-coding genes than expected by chance in *D. melanogaster* (Fig. 4)
	Many lncRNA genes may potentially function only in *cis*	• CRISPR screen of lncRNA genes in *D. melanogaster* showed only a few phenotypes rescued in *trans* (Wen et al. 2016) • Needs further study
Compensatory mutations cause degeneration of *lncRNA* genes	Weak functionality (but high evolvability) may underlie rapid gain/loss patterns	• Promoter loss events are very common (Carelli et al. 2018)
Refinement into enhancer	Regulatory function of promoter resulting from easily accessible mutations may be refined by further mutation	• Promoters may evolve into enhancers (but may be rare events) (Carelli et al. 2018) • Enhancers and promoters may have common origins (Haberle and Stark 2018)
Gain of *trans*-lncRNA function	*trans*-lncRNA function is rarer than *cis*-lncRNA function	• CRISPR screen of lncRNA in *D. melanogaster* showed few phenotypes rescued in *trans* (Wen et al. 2016) • Needs further study
	Refinement of function allows for retention over evolutionary time while also yielding redundancy	• Needs further study
Gain of gene-like properties	U1-PAS axis provides plausible mechanism for gain of splicing and poly-A sites	• U1-PAS axis may elongate nascent transcripts from divergent promoters (Wu and Sharp 2013)
Gain of protein-coding function	*trans*-lncRNA genes have previously evolved to have the same protein function	• Two de novo protein-coding genes, *MYEOV* and *NCYM*, have ancestral *trans*-lncRNA function (Suenaga et al. 2014, 2020; Chen et al. 2015; Papamichos et al. 2015; Liang et al. 2020; Weisman 2022)

potentially through the appearance of substitutions that generate intronic structures and poly-A site sequences in the *cis*-lncRNA gene. A proposed molecular mechanism for producing such gene-like structures from bidirectional, divergent promoters is presented in Wu and Sharp (2013). The resulting action of this nascent *trans*-regulatory lncRNA transcript will cause a change in the selected phenotype through a more refined regulation of the cultivator gene, allowing this new variant to rise to fixation again in linkage with the cultivator gene. Not only does the *trans* function provide further refinement of the regulation particularly under heterozygous states (Fig. 2B), but such refinement could potentially allow for a new lncRNA gene to be maintained for longer periods within evolving populations. The preservation of a new variant within populations is particularly important for lncRNA given the rapid degree of turnover that they are subjected to in *Drosophila* (Nyberg and Machado 2016; Wen et al. 2016), mammals (Paralkar et al. 2014; Kern et al. 2018), and humans (Lin et al. 2023). Note that it is not strictly necessary for a *cis*-lncRNA transcript to gain a *trans* function in our model (Fig. 2B).

Finally, the transcribed *cis*- or *trans*-lncRNA gene may accumulate mutations or insertions that allow it to become translated, such as the key 5′-untranslated region

Table 3
Comparisons of the cultivator model and ORF-first or transcription-first models

	Cultivator model	ORF-first model	Transcription-first model
Origination of variation	Preexisting or rapidly evolvable adaptive variation is fixed in response to altered selective conditions.	Neutral variants exist prior to stress-induced transcription/ translation, favorable variants increase in population frequency during stress	Neutral and/or selection for regulatory function
Exposure to selection	Regulation of cultivator gene(s) expression exposes de novo gene variants to selection	Induction of stress exposes cryptic protein-coding variants of unknown function to selection via promiscuous transcription and translation	Reduction of deleterious effects of noncoding RNA transcription may expose variants to selection. Alternatively, the appearance of favorable protein products of unknown functions exposes variants to selection.
Gain of transcription	Appears as by-product of *cis*-regulation of cultivator gene(s), becomes fixed by linkage	Appears due to stress induction, increases in population frequency and potentially becomes fixed based on adaptive value of proto-gene product	Assumed to exist prior to selection as enhancer RNA or bidirectional promoter
Gain of posttranscriptional processing	May appear based on adaptive value of *trans*-regulation of cultivator gene(s) by de novo transcript product likely via improved regulation of same cultivator gene(s). Potentially occurs through the "U1-PAS axis"	Gain of gene-like properties like gene length and exonic structures continues to evolve based on adaptive value of proto-gene product	Reduces deleterious variation through splicing, gains easily evolvable poly-A processing through chance via the "U1-PAS axis"
Gain of translation	Appears based on the adaptive value of *trans*-regulation of cultivator gene(s) by de novo protein-coding gene product, potentially via improved regulation of same cultivator gene(s)	Gain of gene-like properties continues to evolve based on adaptive value of proto-gene product	May appear based on the adaptive value of de novo protein-coding gene product

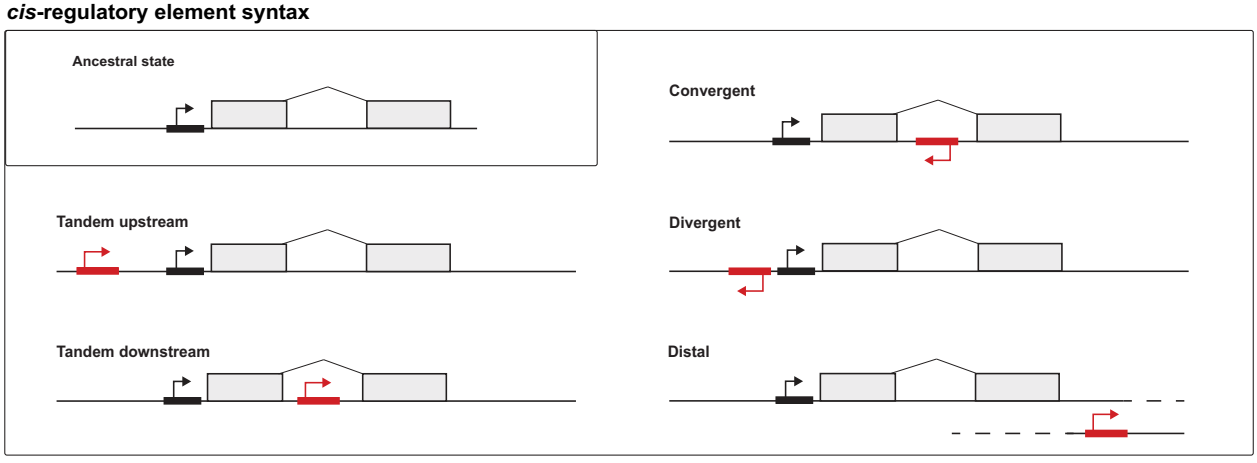

Fig. 1.—New promoter elements may evolve into one of multiple promoter syntaxes. The cultivator model begins with natural selection favoring altered expression of a preexisting cultivator gene within a given tissue (ancestral state). One highly evolvable solution for these selective conditions is the appearance of a new promoter element in close proximity to the cultivator gene. The new promoter element may appear in one of five possible promoter syntaxes: tandem upstream, tandem downstream, convergent, divergent, and distal configurations.

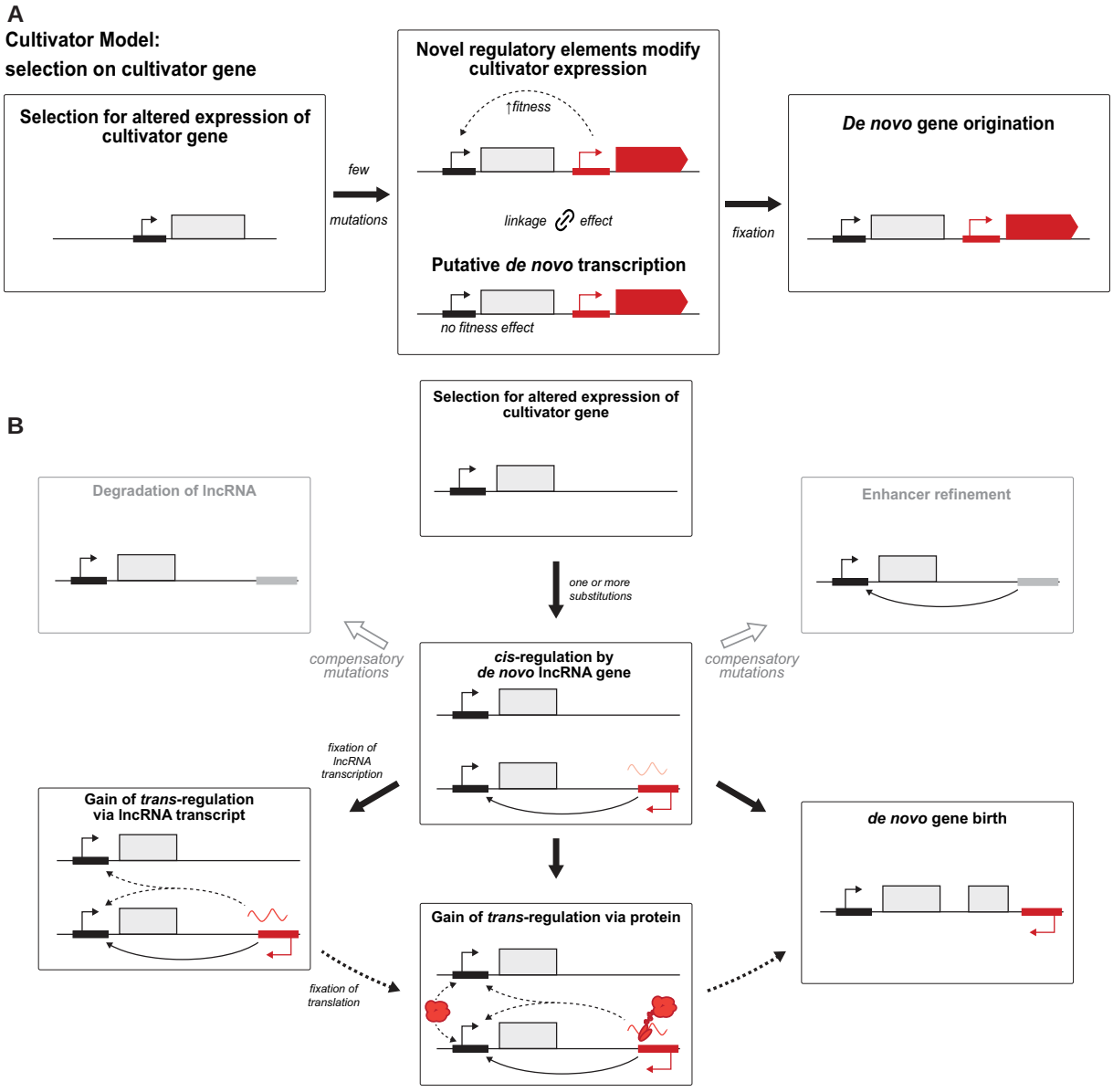

Fig. 2.—Proposed origination mechanism for de novo genes. The cultivator model posits that a new promoter element may evolve rapidly within a population in response to altered selective conditions. Without loss of generality, we assume that increased expression is favored. Additionally, only two of multiple possible promoter configurations are depicted. A) In response to selective conditions, substitutions accumulate in the existing regulatory or promoter regions of the cultivator gene. These substitutions simultaneously increase the expression while also initiating selectively neutral transcription of a new, putative de novo gene. For example, the appearance of a transcription factor binding site might increase tissue-specific expression of the cultivator gene. Simultaneously, these same changes may also initiate transcription of the putative de novo RNA gene ("*cis*-lncRNA gene"). Alternatively, the appearance of a promoter in the vicinity of the cultivator gene could alter regulation of the cultivator gene through supercoiling-mediated coupling, promoter competition, or other mechanisms. Due to the linkage between increased expression of the cultivator gene as well as the expression of the new, neutral de novo transcript ("*cis*-lncRNA transcript"), selection for increased expression drives the fixation of the entire locus, resulting in the origination of a new de novo RNA gene. B) Multiple potential outcomes await the new *cis*-lncRNA gene. One outcome is the appearance of compensatory mutations that supplant the function of the *cis*-lncRNA gene, allowing it to degenerate. Another is the refinement of the *cis*-lncRNA gene into an enhancer element acting on the cultivator gene. The last, but rarest, outcome is the gain of *trans*-acting function for the lncRNA product ("*trans*-lncRNA transcript"), either through the action of the lncRNA ("*trans*-lncRNA gene") or by the evolving protein-coding function. This allows the de novo gene product to directly or indirectly regulate either the cultivator gene or an entirely different genetic pathway in *trans*. As this results in an increase in fitness beyond the *cis*-lncRNA gene, the new changes will fix in a population. Subsequent substitutions may accumulate on the *trans-lncRNA* gene that refine and stabilize its function, potentially transforming the *trans*-lncRNA gene into a de novo protein-coding gene. Note that it is possible that a *cis*-lncRNA gene may potentially gain protein-coding function without developing *trans*-lncRNA function. In general, the de novo gene may retain/refine all prior functions if present, including as an enhancer, as a *cis*-lncRNA gene, and/or as a *trans*-lncRNA gene.

or internal ribosome entry site sequences. Should a variant of this newly originated de novo protein-coding gene further refine the regulation of the cultivator gene, either directly or indirectly, this would allow each subsequent substitution to rise to fixation. If the *cis*- or *trans*-function of the de novo protein-coding gene is maintained, this would also impart a degree of functional redundancy to the de novo gene.

In providing an evolutionary mechanism for fixation at each progressive step, the cultivator model describes a sequential, adaptive process by which a nongenic and/or nonfunctional genomic sequence may initially become transcribed, gain gene-like features such as intronic structures and poly-A sequences, and finally become translated.

Evidence Supporting a Cultivator Model of De Novo Gene Evolution

New Promoters Are Easily Evolvable

One governing principle of genomic evolution is that when natural selection is presented with multiple solutions to a given problem, it will select for the earliest evolvable solution to arise within a population, even if it is not the most efficient solution. This evolvability principle is a key principle that underlies our proposed regulation-focused cultivator model of de novo gene origination (Fig. 2). As the cultivator model proposes that the appearance of a new, proximal promoter is an easily evolvable solution for regulating nearby cultivator genes, the high evolvability of promoters is a key feature of the model.

In a screen of bacterial promoter evolvability, one or at most two activating substitutions were found to be sufficient to generate a de novo promoter from random sequences, while a similarly evolvable solution, microduplication/insertion of a preexisting promoter element, was also found to achieve the same goals (Yona et al. 2018). In total, ∼75% of their evolved promoters required only a single step to achieve promoter activity, with ∼60% of the promoters evolving via a single substitution, and an additional ∼15% evolving via duplication/insertion (Yona et al. 2018). This high evolvability was also directly coupled with inefficient function, as the median promoter activity was functionally deficient, possessing only ∼50% of the wild-type promoter activity (Yona et al. 2018). Should selective conditions for increased promoter activity persist, these de novo promoters should then continue to progressively accumulate mutations that further refine the specific activity of these promoters through a series of sequential fixation events. Throughout this process, the promoter activity levels of these intermediate fixation events should span the levels found in highly accessible de novo promoter states to the highly specific wild-type promoter activity levels that have been refined over millions of years of selection.

Appearance of a Promoter May Rapidly Regulate Neighboring Genes in *cis*

A key aspect of the cultivator model is the *cis*-regulation of the cultivator gene's expression by a proximal promoter element. One potential way this can occur is through the coupling of the modification of the transcription factor binding sites, such as those found in enhancer sequences, with the appearance of a new promoter as a by-product of this substitution (Carelli et al. 2018; Majic and Payne 2020). Alternatively, this *cis*-regulation can occur through a more direct interaction with the newly evolved promoter, either through DNA-supercoiling as demonstrated in Johnstone and Galloway (2022), or through promoter interference mediated by a preexisting enhancer as demonstrated in Fukaya et al. (2016). While the initial function of the new promoter element in our regulation-focused cultivator model is to provide regulation of the cultivator gene in *cis*, a by-product of the appearance of this promoter would be the production of a newly evolved, nascent de novo lncRNA gene (Fig. 2A). The resulting *cis*-lncRNA transcripts themselves are selectively neutral (or weakly deleterious), while the *cis*-lncRNA gene's regulation of the cultivator gene is selectively advantageous. If alteration of the expression of the cultivator gene occurs in a tissue-specific manner, the *cis*-regulatory function of the *cis*-lncRNA would also be expected to occur in a similarly specific manner. This could occur through the appearance of a promoter in an existing tissue-specific enhancer, resulting in dual enhancer–promoter function (Dao et al. 2017) (cf. *cis-lncRNAs may evolve into enhancers*) or the utilization of a tissue-specific promoter. Regardless of the mechanism, the new *cis*-lncRNA is expected to express in a tissue-specific manner, a pattern that is commonly observed in lncRNA (Cabili et al. 2011; Lianoglou et al. 2013).

Importantly, the physical organization of the genome is the key factor allowing the new promoter to regulate the cultivator gene. The central parameter of both the supercoiling-mediated coupling and the promoter competition is the physical proximity in *cis* with the cultivator gene. Regulation is thus expected to drop-off significantly when the newly evolved promoter becomes more distant. The physical linkage of the new promoter and the cultivator gene should therefore play a key role in the evolution of the nascent de novo lncRNA. This may lead to secondary effects owing to hitchhiking, potentially leaving a signature of selective sweeps. This new promoter, with its corresponding de novo lncRNA gene and local hitchhiking variants, will then rise to fixation within a population so long as the selective benefits exceed the penalty from genomic dysregulation caused by the de novo lncRNA in *trans*.

In a prior study, a minimal chemical model of DNA-supercoiling was implemented in silico to explore the distance-dependent effect of promoter syntax on

transcriptional coupling between two neighboring promoters (Johnstone and Galloway 2022). The authors found that in the case of a divergent promoter configuration—the same syntax as that proposed in Wu and Sharp (2013)—induction of one promoter allows for the dynamic coupling of transcription between its neighboring promoters via the reduction of local inter-promoter supercoiling. Similarly, the authors found that a convergent promoter configuration increased local inter-promoter supercoiling, thus generating alternating dynamic transcription. By comparing their predictions to transcription of a pair of genes in *Danio rerio* double mutants, combined either in *cis* or in *trans*, they demonstrate how their model of transcriptional coupling correctly predicts and explains patterning during zebrafish somite formation.

A similar effect of transcriptional interference was seen in a study of transcriptional bursting using an in vivo synthetic reporter assay in *Drosophila melanogaster* embryos (Fukaya et al. 2016). Not only did the authors find alternating dynamic transcriptional coupling similar to that predicted in Johnstone and Galloway (2022), but they also found that the reduction of inter-promoter and promoter–enhancer distance decreased the bursting frequency of the unaltered neighboring promoter. Interestingly, alternating transcriptional coupling was observed both in convergent and divergent promoter syntaxes in Fukaya et al. (2016), contradicting the predictions from Johnstone and Galloway (2022). A key explanatory difference between these studies may be the action of an intervening enhancer element as observed in Fukaya et al. (2016) and not in Johnstone and Galloway (2022), suggesting that it is possible that there may be a fundamental, mechanical difference between supercoiling- and enhancer-mediated transcriptional regulation. Similar transcriptional interference effects were also observed in mammals, where the *PVT1* promoter was found to compete with the *MYC* oncogene promoter for access to four different enhancer elements (Cho et al. 2018). Importantly, this competition occurred in *cis*, where the effect was only detected when the promoters were found on the same chromosome.

In addition to these more direct mechanisms, more studies suggest an even wider variety of regulatory mechanisms for new proximal promoter elements that are both dependent on and independent of promoter syntax. In another study, the lncRNA *bxd* has been shown to regulate its neighboring gene *Ubx* where disruption of the *bxd* promoter generated a *Ubx*-dependent phenotype (Petruk et al. 2006; Pease et al. 2013). Even though deletion of the *bxd* gene body did not rescue the *bxd* phenotype, reinsertion of the *bxd* promoter on the reverse strand provided a rescue of the *bxd* phenotype. This case of *bxd* illustrates an example where the mere distance-dependent presence of a promoter, not its relative orientation, was sufficient to generate phenotype. The case of the *Fub-1* lncRNA, another member of the *Ubx* locus, offers a case study where promoter syntax is key to cultivator gene regulation outside of supercoiling-mediated coupling or promoter interference (Ibragimov et al. 2023). Here, a proper read-through of the *Fub-1* promoter (*HS2*) allowed enhancers in the *bxd/pbx* domain to bypass the wild-type insulating activity of the *Fub-1* lncRNA gene. However, if the direction of the *HS2* promoter element is reversed in situ, the insulating activity of the *Fub-1* lncRNA gene is restored, preventing proper enhancer contact between the *bxd/pbx* locus and *Ubx*.

lncRNA Appear Closer to Protein-coding Genes Than Expected by Chance

A key prediction of the cultivator model is an overabundance of lncRNAs in close proximity to older cultivator genes. As the origination of these lncRNAs should be driven by short-range promoter interactions that occur on the order of a few hundred base pairs (Fukaya et al. 2016; Johnstone and Galloway 2022), rather on the order of thousands of base pairs as is typical of enhancer-driven interactions (Lee et al. 2024), these lncRNAs should also be specifically biased toward similarly shorter distances. We scanned the latest FlyBase genome annotation (dmel6r50) (Gramates et al. 2022), showing an abundance of FlyBase-annotated ncRNAs in the vicinity of known protein-coding genes (Figs. 3 and 4). Some notable examples are presented in Fig. 3, including the pair-rule gene *even-skipped*, the apoptosis factor *grim*, the Hox gene *Ultrabithorax*, and the photoreceptor development gene *seven up*. The case of *Ultrabithorax* is particularly interesting, as it is located near the lncRNA *bithoraxoid (bxd)* and *HS2*, which have been previously demonstrated as regulating *Ultrabithorax* in *cis* but not in *trans* (Petruk et al. 2006; Pease et al. 2013; Ibragimov et al. 2023).

A more general analysis of inter-genic distances between ncRNAs and annotated protein-coding genes reveals that ncRNAs are found to be significantly closer to annotated neighboring genes than would be expected by chance (Fig. 4, $P < 2 \times 10^{-16}$). More specifically, we first calculated the distribution of inter-genic distances between existing ncRNAs and protein-coding genes. We then took each ncRNA annotation and moved it to another location on the same chromosome *in silico*. The new position was chosen at random with a uniform probability between the first and last annotated gene locations on each chromosome. The distance between these shuffled ncRNA positions and the real positions of protein-coding genes was calculated. This procedure was repeated 1,000 times to generate the distribution of inter-genic distances that would be expected by chance. It is important to note that this simple procedure does not account for GC bias or open chromatin, which would help to increase the evolvability of promoters in the vicinity of existing genes (Majic and Payne 2020).

While this result is not definitive and does not necessarily exclude other mechanisms, one possible interpretation is that proximity-based regulation of neighboring cultivator

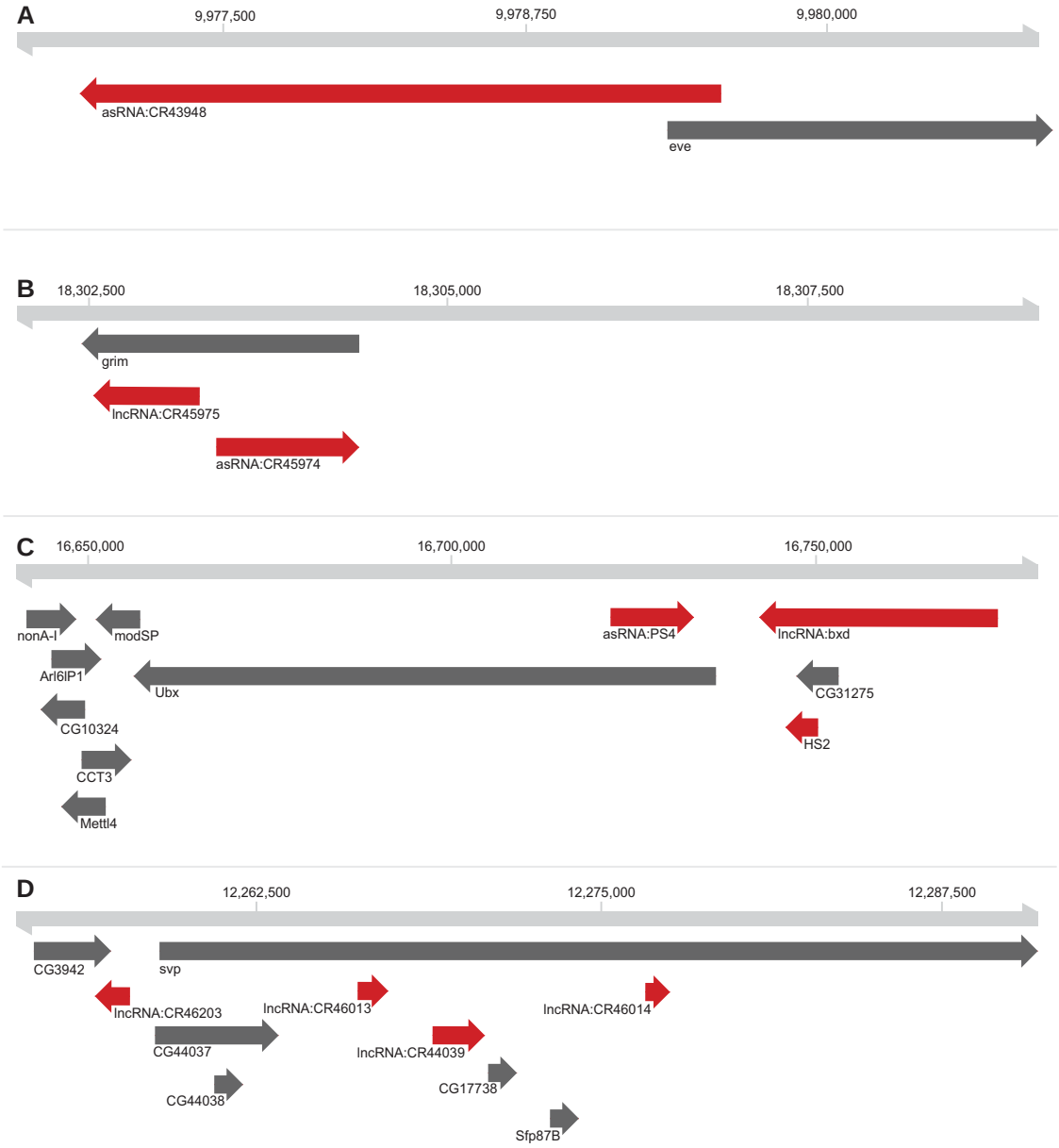

Fig. 3.—Functionally important genes often have annotated lncRNA and de novo genes near their transcription start sites. A) *even-skipped* is a pair-rule gene located on chromosome 2R that plays a key role in body plan specification and contains an uncharacterized lncRNA, CR43948, within its intron. B) *grim* is a crucial regulator of cellular apoptosis located on chromosome 3L, which contains two uncharacterized lncRNA, *CR45975*, and *CR45974*, near its transcription start site. C) *Ultrabithorax* is a homeobox gene located in the canonical *Hox* gene cluster on chromosome 3R with notable lncRNA near its transcriptional start site, including *PS4*, *bithoraxoid* (*bxd*), and the *Fab-1* locus (*HS2* promoter). Notably, *bxd* and *HS2* have been shown to regulate *Ultrabithorax* in *cis* but not in *trans*. D) *seven up* is a key regulator of photoreceptor development located on chromosome 3R with a large number of uncharacterized lncRNA located near its transcriptional start site. Notably, *CG44037* and *CG44038* have been identified as de novo protein-coding genes.

genes may play a role in lncRNA origination. Notably, these calculated inter-genic distances do not account for 3D chromosomal configurations and thus represent an upper limit of true physical distance. Further exploration of *cis*-acting lncRNAs, including their evolution and potential relation to the origination of protein-coding genes, is reviewed in Gil and Ulitsky (2020).

cis-lncRNAs May Have Multiple Evolutionary Fates

There are multiple possible fates of the nascent de novo *cis*-lncRNA gene subsequent to fixation (Fig. 2B). The first possible fate is the trivial and likeliest case: appearance of compensatory mutations elsewhere in the genome resulting in a degeneration and loss of the nascent de novo *cis*-lncRNA gene. This would likely occur through the

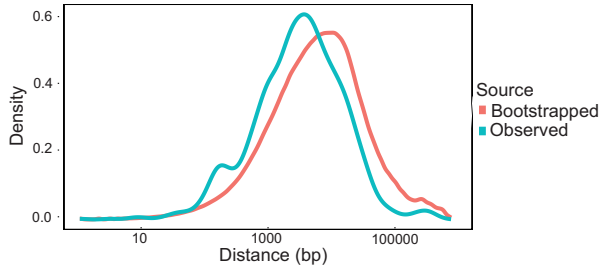

Fig. 4.—Annotated ncRNA are closer to genic promoters than expected by chance. The distances between annotated ncRNA and their nearest neighboring genes were calculated using the latest FlyBase *D. melanogaster* annotations (dmel6r50) using chromosomes 2L, 2R, 3L, 3R, 4, X, and Y. The expectation of the same distribution of distances was bootstrapped 1,000 times by placing simulated ncRNA with uniform probability across each chromosome between the first and last annotated genes per chromosome, with the total number of ncRNA per chromosome being preserved in each bootstrap. The distribution of promoter distances was significantly lower for the observed genomic ncRNA than for the bootstrapped distribution ($P < 2 \times 10^{-16}$, Student's *t*-test). These results suggest that the nonrandom distribution of genomic ncRNA is driven by interactions with neighboring genic elements, which is potentially the result of supercoiling- or enhancer-mediated promoter regulation.

accumulation of inactivating mutations that would propagate neutrally through a population. The next possible fate is further refinement of the *cis*-lncRNA's function, potentially transforming into an enhancer element (cf. Carelli et al. 2018). If this occurs, the new enhancer variant would subsequently fix within a population, resulting in decreased transcriptional activity of the de novo *cis*-lncRNA gene. Another outcome, albeit unlikely, would be the evolution of selectively advantageous regulatory activity of the de novo *cis*-lncRNA transcript in *trans*, as potentially demonstrated in Wen et al. (2016). Importantly, the fixation of this new de novo *trans*-lncRNA gene is contingent on the production of some selectively advantageous phenotype downstream of the de novo *trans*-lncRNA transcript. This could occur through a selectively advantageous consequence of the regulation of either the original cultivator gene or a different cultivator gene entirely. The regulatory activity of the nascent *trans*-lncRNA gene thus provides the necessary selective pressure for further accumulation and fixation of mutations that produce increased transcriptional activation. A final possible outcome would be the appearance of protein-coding capacity directly from the *cis*-lncRNA gene.

cis-lncRNAs May Refine Their Function After Initial Appearance

The molecular mechanisms we propose for the initial appearance of the *cis*-lncRNA gene are crude. However, a review of known lncRNAs (Wang and Chang 2011) has identified four general archetypes of lncRNA functions: signal, decoy, guide, and scaffold. Of these, the signal and guide archetypes may provide further refinement of the *cis*-lncRNA function beyond the initial appearance of a new promoter element. In the case of the signal archetype, the nascent de novo *cis*-lncRNA gene may evolve sequences that allow the transcript to demarcate particular spatiotemporal patterns upstream of preexisting regulatory machinery in *cis*. Alternatively in the case of the guide archetype, the nascent *cis*-lncRNA gene may accumulate mutations that allow the transcript to physically recruit regulatory factors to specific genomic locations also in *cis*. While it may be tempting to hypothesize that these functional archetypes themselves may serve as the evolutionary "motive" for the origination of new lncRNA, the sequence-dependent nature of these archetypal functions would likely require the appearance of multiple independent mutations. As this large number of mutations may be difficult to achieve in large populations, the rapid appearance of a promoter element that serves as a crude initial regulatory mechanism, and thus the appearance of the *cis*-lncRNA transcript, is likely to precede the appearance of these more refined, highly regulated functions. This observation is consistent with the previously discussed study of the *bxd* lncRNA, where in situ inversion of the promoter element demonstrated that the *cis*-regulatory function of *bxd* on *Ubx* is not dependent on the sequence of the *bxd* transcript, but instead on the relative positioning of the *bxd* promoter element (Pease et al. 2013).

cis-lncRNAs May Evolve Into Enhancers

A central element of the cultivator model is the adaptive progression from a nascent transcript to a de novo protein-coding gene over a series of successive steps (Fig. 2). While we provide an evolutionary mechanism connecting the first appearance of a new promoter element to the translation of a new polypeptide, there are also multiple alternate but terminal pathways that do not result in the production of a de novo protein-coding gene. One possible outcome is the evolution of a *cis*-lncRNA gene into a stable enhancer element. Eleven known examples of lncRNA with homologous enhancer sequences were previously found in Engreitz et al. (2016). Additionally, a more comprehensive study used histone modification data to identify the directionality of these changes (Carelli et al. 2018). In accordance with the trivial case of the cultivator model, the authors found widespread promoter loss events. The authors also detected a significantly smaller fraction of enhancer-to-promoter and promoter-to-enhancer interconversion events (~1,300 promoter loss, 27 enhancer-to-promoter, 1 promoter-to-enhancer events in primates.) The authors note that their detected events are a conservative estimate of the true number of these interconversion events, as they excluded cases in which the same sequence displayed both enhancer and promoter activity. The authors further

mention that bias toward enhancer-to-promoter events (and not vice versa) is likely real, as their study design would be more sensitive to detecting promoter-to-enhancer events. This suggests that lncRNA loss is orders of magnitude likelier to be the eventual fate of new de novo lncRNA gene/promoters than enhancer refinement. An alternative possibility is that refinement of the regulatory function of a nascent promoter occurs on significantly faster time scales than speciation. This latter hypothesis would be in line with a model of enhancer evolution where promoters and enhancers share a common ancestral promoter-like state (Haberle and Stark 2018). Regardless, these enhancer-to-promoter and promoter-to-enhancer observations are in line with the cultivator model, providing an insight into the origination of new promoter elements, including bidirectional promoters.

cis-lncRNA Genes May Evolve into *trans*-lncRNA Genes

Under the cultivator model, the fixation of the de novo *cis*-lncRNA gene is dependent on a balance of selective forces. It is possible that the nascent *cis*-lncRNA gene, whose function is to regulate a cultivator gene in *cis*, may produce deleterious downstream effects owing to some downstream interactions of the newly evolved transcript. However, so long as the beneficial selective effects outweigh these deleterious effects, the new lncRNA should become fixed within the population. These residual deleterious effects then produce further selective pressure to alleviate downstream effects. If these downstream effects are sequence-dependent, a simple evolutionary solution would be the appearance of a new snRNP site that would splice out the deleterious sequences, allowing for the further evolution of intronic structures. This development of complex structures would subsequently increase the search space for potential sequence-dependent *trans*-regulatory functions and transcript-stabilizing poly-A site sequences (cf. *U1- poly-A signal (PAS) axis for lncRNA elongation is compatible with the cultivator model*) (Wu and Sharp 2013).

One prediction of the cultivator model would be the existence of a larger number of *cis*-lncRNA genes than *trans*-lncRNA genes, as specific alterations following the fixation of new *cis*-lncRNA sequences are required for the appearance of *trans*-regulatory function. In a systematic study of lncRNAs in *D. melanogaster* (Wen et al. 2016), the authors performed a large-scale CRISPR knockout screen of 108 lncRNAs expressed in the *D. melanogaster* testis. Of these 108 lncRNAs, 19 lncRNAs were found to have gross loss-of-function phenotypes during spermatid individualization, and of these 19 lncRNAs, six were found to produce a phenotype that was rescuable in *trans*. It is neither clear whether the authors attempted *trans* rescue of all 19 of these lncRNAs nor was it clear whether the rescue of these phenotypes in *trans* was complete or partial. Further intrinsic difficulties in the interpretation of *cis*- versus *trans*-regulatory functions of lncRNAs are discussed and reviewed in Gil and Ulitsky (2020), highlighting the need for more comprehensive genome-wide studies on lncRNA function. In the absence of larger-scale studies, these results may suggest that it is unlikely that the majority of lncRNAs act on *Drosophila* spermatogenesis in *trans*, instead acting in *cis*. In our model, a de novo *trans*-lncRNA could either regulate the original cultivator gene potentially far upstream of it or a different cultivator gene entirely. The de novo *trans*-lncRNA gene would be expected to fix in a given population so long as the benefits outweigh the costs, either through a further refinement of the regulatory action of the *trans*-lncRNA transcript on the cultivator gene or with a reduction of the penalties caused by genomic dysregulation downstream of the *trans*-lncRNA transcript, both of which could potentially occur in parallel.

U1-PAS Axis for lncRNA Elongation Is Compatible With Cultivator Model

The authors of the transcription-first model presented in Wu and Sharp (2013) argue that transcription of divergent promoters in mammals and yeast (one of five possible promoter syntaxes, Fig. 1A) may provide a basis from which de novo protein-coding genes could arise. The central observation of their model is that the binding of U1 snRNPs suppresses the binding of poly-A signal (PAS) machinery and subsequent transcript cleavage, which they refer to as the "U1-PAS axis." They propose that the appearance of a U1 snRNP splicing site in the antisense strand of a bidirectional promoter would subsequently suppress the function of a PAS site. While it is not explicitly stated, this would allow the repressed PAS site to accumulate neutral mutations while simultaneously elongating the nascent antisense transcript. Similarly, transcription of the antisense strand would cause R-loop formation, which is known to increase the mutation rate toward G–T rich substitutions, increasing the likelihood of the appearance of a G–T rich U1 snRNP motif while causing A-rich PAS motifs to be eliminated. The collective effect of these phenomena would constitute, as the authors suggest, a positive feedback loop that continues to elongate nascent transcripts from divergent promoters over evolutionarily relevant timescales, thus potentially providing the genetic diversity from which de novo genes may emerge.

While this model provides mechanistic insights into how lncRNA genes may begin to gain gene-like properties, their transcription-first model lacks detail on a mechanism by which new variants would fix or accumulate mutations, instead tacitly relying on high polymorphism rates within a population to allow for the accumulation of multiple sequential mutations. This allows for a synergy between their model and the cultivator gene model, which provides an

evolutionary mechanism explaining the preservation and sequential fixation of such mutations. In fact, the application of the transcription-first model of bidirectional promoters would provide a mechanism that would allow for the acceleration of the process described in the cultivator model, where selectively advantageous regulation would elevate the G–T rich mutation rate of the nascent *cis*-lncRNA genes. This would subsequently allow for longer transcribing variants to appear more rapidly in the population than would be expected by chance alone.

Some De Novo Proteins May Evolve From *trans-lncRNA* Genes

The last step of our cultivator model is the appearance of substitutions that further refine the regulatory activity of the de novo *trans*-lncRNA transcripts through the birth of a de novo protein-coding gene (Fig. 2B). Crucially, the appearance and fixation of a de novo protein-coding gene is strictly dependent on a selective advantage of the new variant above and beyond the function already provided by either the *cis*-lncRNA and/or *trans*-lncRNA genes. As before, the regulatory action of this de novo polypeptide on some cultivator gene drives the fixation of this new variant, which is activated through the appearance of 5′-UTR, internal ribosomal entry site, and other sequences required for proper translation of the *trans*-lncRNA transcript. While this could occur through the gradual accumulation of single nucleotide changes, there may not be enough selective pressure to allow such multiple mutations. In the absence of studies demonstrating that these changes are easily evolvable, as in the case of new promoters, a rapidly evolvable mechanism might be the duplication of preexisting material into the 5′ region of the *trans*-lncRNA gene. Alternatively, the promiscuous transcription and translation of de novo variants, particularly under periods of stress as described in the proto-gene model (Carvunis et al. 2012; Vakirlis et al. 2020), may expose new de novo protein-coding variants to natural selection.

While the transformation of a lncRNA gene into a de novo protein-coding gene may seem exceedingly unlikely, examination of *seven up*, a key gene involved in the specification of rhabdomere identity in *D. melanogaster*, demonstrates the possibility of this outcome (Fig. 3). The *seven up* locus contains three annotated lncRNA genes and four protein-coding genes within its first intron along with a divergent lncRNA gene just upstream of the *seven up* promoter. Interestingly, the two intronic protein-coding genes closest to the *seven up* transcription start site, CG44037 and CG44038, were both identified as de novo protein-coding genes that evolved between the *D. willistoni/S. lebanonensis* and the *D. ananassae/D. kikkawai* divergences, respectively (Peng and Zhao 2024). While these origination events have not been definitively characterized as the result of a *trans*-lncRNA gene evolving into a de novo protein-coding gene through the action of a cultivator gene, there are two known examples of such occurrences occurring outside of *Drosophila*: (i) in the case of the *MYEOV* protein-coding gene evolving from the *MYEOV trans*-lncRNA gene, both of which regulate *HES1* in *trans* (Chen et al. 2015; Papamichos et al. 2015; Liang et al. 2020) and (ii) in the case of the *NCYM* protein-coding gene evolving from the *MYCNOS trans*-lncRNA gene, both of which regulate *MCYN* (Suenaga et al. 2014), as reviewed in Suenaga et al. (2020) and Weisman (2022). Notably, while *MYEOV* protein and its transcript are located on a different chromosome than *HES1*, *NCYM* and the *MYCNOS* lncRNA genes are contained within the first intron of *MYCN*. One interesting consequence of this mechanism is the appearance and retention of redundant functions. While it may be tempting to speculate that such redundancy allows for robust function under uncertain circumstances such as inactivating mutations, the cultivator model suggests that such redundancy may instead be the result of sequential refinement of functions.

Concluding Observations

The Cultivator Model Provides a Stepwise Evolutionary Mechanism for Fixation

In contrast to prior models of de novo gene origination, the cultivator model of de novo gene origination may be summarized as an iterative process of refinement of an explicitly defined function. In this case, the explicit function of the nascent de novo gene is to regulate one or more preexisting cultivator gene(s). The initial steps consist of a crude but rapid solution to the original selective demand that is facilitated by the physical organization of the genome, with subsequent steps involving further refinement of this function. While there are multiple potential outcomes of this process, there remains an unlikely but viable evolutionary pathway allowing for the fixation of each subsequent putative de novo variant, starting with the initial activation of a transcriptionally inactive region of the genome, to the appearance of a fully functional de novo protein-coding gene. Crucially, the cultivator model generates key predictions about how the transcriptional characteristics of neighboring genes would be altered upon the origination of a new de novo gene (Supplementary Table S2, Supplementary Material online). When comparing the expressional divergence of such key genes between in-group and out-group species, the cultivator model predicts that the appearance of new de novo variants should have a strong regulatory influence on future neighboring genes in a distance-dependent manner. Furthermore, the cultivator model integrates observations from studies into both ORF- and transcription-first models into a cohesive and synergistic evolutionary model of both de novo RNA and protein-coding gene origination.

Regulatory Changes Are Highly Evolvable

One key aspect of the cultivator model of de novo gene evolution is its strong focus on regulatory rather than protein function—this is mirrored by insights gained from the *MazF/ramF* study in relation to the high relative evolvability of regulatory versus protein function. In their study, they found that of their ~10^8 strains producing random proteins, ~2,000 of these strains were able to fulfill their selection criteria through upstream inactivation of the toxin *MazF*'s promoter. Alternatively, a "direct" amino acid solution, *ramF*, was found in only a single strain, suggesting that the evolution of regulatory solutions may be roughly three orders of magnitude more evolvable than direct amino acid solutions. Similarly, the rescue of *ΔserB* auxotrophs occurred through upstream regulatory changes that derepressed the expression of *serB* enzyme through protein–RNA interactions. While not a major focus of either study, the mechanism underlying the observed phenotypes in both studies suggests that characterization of the regulatory functioning of nascent de novo lncRNA genes may be a crucial step for understanding the origination and evolution of de novo protein-coding genes.

Genome Organization May Play a Key Role in De Novo Gene Origination

In a prescient commentary (Siepel 2009), Siepel pointed out that genomic context could play a key role in the origination process of both de novo and duplicate gene evolution. Interestingly, this work suggested that nucleotide content could also play a strong role in ORF-expansion, similar in nature to the "U1-PAS" axis. Beyond this idea, the commentary also suggested that the appearance of a new gene may be coupled with alteration in the regulatory environment of neighboring genes. As a result of a beneficial change in regulation, the newly evolved de novo variant could thus rise to fixation through a linkage-based effect. Similarly, while lacking a well-articulated selective function, a more mechanistic model of promoter origination suggested a common, unified origination mechanism for enhancer and messenger RNAs (Haberle and Stark 2018). While these ideas were not further explored beyond their mention in a handful of sentences, these commentaries hint at a deeper interplay between genomic evolution and new gene evolution, suggesting how transcription, regulation, and genomic context might be an underexplored factor in new gene evolution.

In comparison to prior models of de novo gene origination, the proto-gene model is generally agnostic to genomic organization. Alternatively, the transcription-first models consider only the context of immediate neighboring genes in the case of divergent promoters, which is only one of multiple possible promoter syntaxes. Similarly, linear genomic proximity between the new promoter element and the cultivator gene is a key driving factor in the cultivator model, particularly in the case of supercoiling-mediated mechanisms. However, as the cultivator model considers the entirety of the *cis*-regulatory environment of important cultivator genes, the 3D organization of the genome should also be expected to play a strong role in the origination of de novo genes. As described earlier, the modification of preexisting transcription factor binding sites could cause the simultaneous appearance of a new promoter element. In this case, the interaction between the preexisting *cis*-regulatory element and the cultivator gene may be quite distant in linear chromosomal distance, but instead made proximal through the action of chromatin looping. Similarly, as the promoter interference mechanism operates in a distance-dependent manner, it is likely that the key distances involved would be the inter-promoter distance in 3D space, rather than the linear genomic distance along the chromosome. As a result, the cultivator model highlights the role that larger-scale genome organization could potentially play in new gene evolution.

Future Directions

While we present the regulation-focused cultivator model of de novo gene origination as a major mechanism shaping the appearance of new genes, we wish to stress that we are not using our model to claim that ORF- or transcription-first models are invalid. In fact, these models are highly compatible and synergistic with various elements of each model, providing key mechanistic insights into de novo gene origination. Although a unique feature of our model is that the birth and appearance of a de novo RNA-based gene is coupled simultaneously with the appearance of *cis*-regulatory function, the cultivator model could, in a very broad sense, be loosely classified under the general umbrella of transcription-first models. Conversely, our model underscores the potential importance of genome organization, cultivator genes, and linkage. Additionally, the exposure of amino acid products of untranslated ORFs to natural selection remains insufficiently defined not only in transcription-first models but in our models as well. Strong evidence supporting the stress-based induction of promiscuous translation as described in the proto-gene model provides key mechanistic insights into how this could occur. As such, we emphasize that these models are not competing models, but instead are complementary in nature.

Given the concrete predictions of the cultivator model, future experiments and analysis should help illuminate different patterns of natural selection. To date, comprehensive and systematic exploration of the function of noncoding RNA genes using CRISPR/Cas9 has been relatively limited. With continuing growth in the number of genetically tractable model systems, comparative analyses should allow for testing associations between complex regulation and the origination rates of *trans*-regulatory function in noncoding RNA and protein-coding genes. For example,

after knocking out lncRNA genes, it would be useful to understand the changes in the transcription and translation levels, and patterns for nearby genes, assuming one of them might be a cultivator gene. Further research into the function of the vast number of uncharacterized protein-coding genes should also help test associations between the appearance of noncoding RNAs, external environmental changes, and differing gene ontologies.

In general, models of de novo gene origination provide key insights into natural selection by distilling different aspects of nature's deep complexity into a series of simplified ideas. Some models may be more applicable than others in different scenarios, such as in the case of single-cell versus multicellular organisms, but it is unlikely that any single one of these models provides a clear explanation for the origination of all de novo genes. The models through which we understand natural selection should also continue to grow as we extrapolate increasingly sophisticated principles with the further exploration and characterization of different examples of de novo genes. We expect that further exploration into newer conceptual paradigms and the unveiling of as-yet-undiscovered biological mechanisms will continue to highlight the deeply multifaceted nature of our understanding of gene birth and evolution.

Supplementary Material

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

Acknowledgments

We would like to thank Dr Esther Betran and the reviewers for providing very helpful commentary and ideas that helped to refine the cultivator model and this paper.

Funding

This work was supported by National Institutes of Health (NIH) MIRA R35GM133780, the Robertson Foundation, and an Allen Distinguished Investigator Award from Paul G. Allen Family Foundation to L.Z. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Data Availability

The data that support the findings of this study are available in FlyBase at flybase.org (Gramates et al. 2022). No new data were generated in this study.

Literature Cited

An NA, Zhang J, Mo F, Luan X, Tian L, Shen QS, Li X, Li C, Zhou F, Zhang B, et al. De novo genes with an lncRNA origin encode unique human brain developmental functionality. Nat Ecol Evol. 2023:7(2): 264–278. https://doi.org/10.1038/s41559-022-01925-6.

Babina AM, Surkov S, Ye W, Jerlström-Hultqvist J, Larsson M, Holmqvist E, Jemth P, Andersson DI, Knopp M. Rescue of Escherichia coli auxotrophy by de novo small proteins. Elife. 2023:12:e78299. https://doi.org/10.7554/eLife.78299.

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. https://doi.org/10.1186/s12862-018-1232-z.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011:25(18):1915–1927. https://doi.org/10.1101/gad.17446611.

Carelli FN, Liechti A, Halbert J, Warnefors M, Kaessmann H. Repurposing of promoters and enhancers during mammalian evolution. Nat Commun. 2018:9(1):4066. https://doi.org/10.1038/s41467-018-06544-z.

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

Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, Li Y, Liu C-J, Luan X, Ding W, Li S, et al. Emergence, retention and selection: a trilogy of origination for functional De Novo proteins from ancestral LncRNAs in primates. PLoS Genet. 2015:11(7):e1005391. https://doi.org/10.1371/journal.pgen.1005391.

Cho SW, Xu J, Sun R, Mumbach MR, Carter AC, Chen YG, Yost KE, Kim J, He J, Nevins SA, et al. Promoter of lncRNA gene *PVT1* is a tumor-suppressor DNA boundary element. Cell. 2018:173(6): 1398–1412.e22. https://doi.org/10.1016/j.cell.2018.03.068.

Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, Charbonnier G, Griffon A, Vanhille L, Stephen T, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Genet. 2017:49(7): 1073–1081. https://doi.org/10.1038/ng.3884.

Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature. 2016:539(7629):452–455. https://doi.org/10.1038/nature20149.

Frumkin I, Laub MT. Selection of a de novo gene that can promote survival of Escherichia coli by modulating protein homeostasis pathways. Nat Ecol Evol. 2023:7(12):2067–2079. https://doi.org/10.1038/s41559-023-02224-4.

Fukaya T, Lim B, Levine M. Enhancer control of transcriptional bursting. Cell. 2016:166(2):358–368. https://doi.org/10.1016/j.cell.2016.05.025.

Gil N, Ulitsky I. Regulation of gene expression by cis-acting long noncoding RNAs. Nat Rev Genet. 2020:21(2):102–117. https://doi.org/10.1038/s41576-019-0184-5.

Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, Goodman JL, Goutte-Gattat D, Jenkins VK, Kaufman T, et al. FlyBase: a guided tour of highlighted features. Genetics. 2022:220(4):iyac035. https://doi.org/10.1093/genetics/iyac035.

Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018:19(10): 621–637. https://doi.org/10.1038/s41580-018-0028-8.

Ibragimov A, Bing XY, Shidlovskii YV, Levine M, Georgiev P, Schedl P. lncRNA read-through regulates the BX-C insulator *Fub-1*. eLife. 2023:12:e84711. https://doi.org/10.7554/eLife.84711.

Johnstone CP, Galloway KE. Supercoiling-mediated feedback rapidly couples and tunes transcription. Cell Rep. 2022:41(3):111492. https://doi.org/10.1016/j.celrep.2022.111492.

Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010:20(10):1313–1326. https://doi.org/10.1101/gr.101386.109.

Kern C, Wang Y, Chitwood J, Korf I, Delany M, Cheng H, Medrano JF, Van Eenennaam AL, Ernst C, Ross P, et al. Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. BMC Genomics. 2018:19(1):684. https://doi.org/10.1186/s12864-018-5037-7.

Lee U, Arsala D, Xia S, Li C, Ali M, Svetec N, Langer CB, Sobreira DR, Eres I, Sosa D, et al. The Role of the 3-Dimensional Genome in New Gene Evolution. bioRxiv. 2022.11.30.518413. https://doi.org/10.1101/2022.11.30.518413., 06 January 2024, preprint: not peer reviewed

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 USA. 2006:103(26):9935–9939. https://doi.org/10.1073/pnas.0509809103.

Liang E, Lu Y, Shi Y, Zhou Q, Zhi F. MYEOV increases HES1 expression and promotes pancreatic cancer progression by enhancing SOX9 transactivity. Oncogene. 2020:39(41):6437–6450. https://doi.org/10.1038/s41388-020-01443-4.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 2013:27(21):2380–2396. https://doi.org/10.1101/gad.229328.113.

Lin D, Hiron TK, O'Callaghan CA. Intragenic transcriptional interference regulates the human immune ligand MICA. EMBO J. 2018:37(10):e97138. https://doi.org/10.15252/embj.201797138.

Lin J, Wen Y, Tang J, Zhang X, Zhang H, Zhu H. Human-specific lncRNAs contributed critically to human evolution by distinctly regulating gene expression. bioRxiv 2023.05.31.543169. https://doi.org/10.1101/2023.05.31.543169., 02 September 2023, preprint: not peer reviewed

Long M, VanKuren NW, Chen S, Vibranovski MD. New gene evolution: little did we know. Annu Rev Genet. 2013:47(1):307–333. https://doi.org/10.1146/annurev-genet-111212-133301.

Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015:112(51):15690–15695. https://doi.org/10.1073/pnas.1514974112.

Majic P, Payne JL. Enhancers facilitate the birth of De Novo genes and gene integration into regulatory networks. Mol Biol Evol. 2020:37(4):1165–1178. https://doi.org/10.1093/molbev/msz300.

McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016:17(9):567–578. https://doi.org/10.1038/nrg.2016.78.

Moyers BA, Zhang J. Evaluating phylostratigraphic evidence for widespread De Novo gene birth in genome evolution. Mol Biol Evol. 2016:33(5):1245–1256. https://doi.org/10.1093/molbev/msw008.

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. https://doi.org/10.1038/s41559-017-0127.

Nyberg KG, Machado CA. Comparative expression dynamics of intergenic long noncoding RNAs in the genus Drosophila. Genome Biol Evol. 2016:8(6):1839–1858. https://doi.org/10.1093/gbe/evw116.

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.

Paralkar VR, Mishra T, Luan J, Yao Y, Kossenkov AV, Anderson SM, Dunagin M, Pimkin M, Gore M, Sun D, et al. Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. Blood. 2014:123(12):1927–1937. https://doi.org/10.1182/blood-2013-12-544494.

Pease B, Borges AC, Bender W. Noncoding RNAs of the Ultrabithorax domain of the Drosophila bithorax complex. Genetics. 2013:195(4):1253–1264. https://doi.org/10.1534/genetics.113.155036.

Peng J, Zhao L. The origin and structural evolution of de novo genes in Drosophila. Nat Commun. 2024:15(1):810. https://doi.org/10.1038/s41467-024-45028-1.

Petruk S, Riley KM, Hodgson J, Schweisguth F, Hirose S, Jaynes JB, Brock HW, Mazo A. Transcription of bxd noncoding RNAs promoted by trithorax represses ubx in cis by transcriptional interference. Cell. 2006:127(6):1209–1221. https://doi.org/10.1016/j.cell.2006.10.039.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. Origins of De Novo genes in human and chimpanzee. PLoS Genet. 2015:11(12):e1005721. https://doi.org/10.1371/journal.pgen.1005721.

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.

Schlötterer C. Genes from scratch – the evolutionary fate of de novo genes. Trends Genet. 2015:31(4):215–219. https://doi.org/10.1016/j.tig.2015.02.007.

Siepel A. Darwinian alchemy: human genes from noncoding DNA. Genome Res. 2009:19(10):1693–1695. https://doi.org/10.1101/gr.098376.109.

Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021:22(2):96–118. https://doi.org/10.1038/s41580-020-00315-9.

Suenaga Y, Islam SMR, Alagu J, Kaneko Y, Kato M, Tanaka Y, Kawana H, Hossain S, Matsumoto D, Yamamoto M, et al. NCYM, a cis-antisense gene of MYCN, encodes a De Novo evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas. PLoS Genet. 2014:10(1):e1003996. https://doi.org/10.1371/journal.pgen.1003996.

Suenaga Y, Nakatani K, Nakagawara A. De novo evolved gene product NCYM in the pathogenesis and clinical outcome of human neuroblastomas and other cancers. Jpn J Clin Oncol. 2020:50(8):839–846. https://doi.org/10.1093/jjco/hyaa097.

Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW, Hines CP, Iannotta J, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat Commun. 2020:11(1):781. https://doi.org/10.1038/s41467-020-14500-z.

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

Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011:43(6):904–914. https://doi.org/10.1016/j.molcel.2011.08.018.

Weisman CM. The origins and functions of De Novo genes: against all odds? J Mol Evol. 2022:90(3-4):244–257. https://doi.org/10.1007/s00239-022-10055-3.

Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020:18(11):e3000862. https://doi.org/10.1371/journal.pbio.3000862.

Wen K, Yang L, Xiong T, Di C, Ma D, Wu M, Xue Z, Zhang X, Long L, Zhang W, et al. Critical roles of long noncoding RNAs in Drosophila spermatogenesis. Genome Res. 2016:26(9):1233–1244. https://doi.org/10.1101/gr.199547.115.

Wu X, Sharp PA. Divergent transcription: a driving force for new gene origination? Cell. 2013:155(5):990–996. https://doi.org/10.1016/j.cell.2013.10.048.

Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. Nat Commun. 2018:9(1):1530. https://doi.org/10.1038/s41467-018-04026-w.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. Rapid evolution of protein diversity by de novo origination in Oryza. Nat Ecol Evol. 2019:3(4):679–690. https://doi.org/10.1038/s41559-019-0822-5.

Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de Novo genes in Drosophila melanogaster populations. Science. 2014:343(6172):769–772. doi: 10.1126/science.1248286.

Associate editor: Esther Betran