De novo genes in the human genome: Origin, roles, and evolution a CNS perspective

Mahalakshmi B. R[1] , **Latha K**[2] , **Priya M. D**[1] , **BK Manjunatha***[3] , **Sharada Devi J. N.**[1] , and **Kiran Kumar H. B.**[4]

[1]Department of Zoology, Government Science College, Nrupathunga University, Nrupathunga Road, Bangalore, 560001, India
[2]Maharanis Science College for women, Autonomous JLB Road, Mysore, Affiliated to University of Mysore, Karnataka, India
[3]Department of Biotechnology, The Oxford College of Engineering, Bengaluru-560068, Karnataka, India
[4]Former Post-doc NCBS, Affiliated to Nrupathunga University, Bengaluru-560001, India

ABSTRACT

The human genome contains many more protein-coding regions than previously recognized, with most newly identified translated open reading frames (ORFs) being evolutionarily recent and largely restricted to humans or primates. They also do not fit the usual protein-coding annotation standards. Several of these genes are de novo, derived from non-genic ancestral sequences. De novo genes are typically characterized as those that arise from ancestrally non-genic sequences. Investigating de novo-originated genes provides numerous chances to better understand their origins and functions, regulatory mechanisms, and associated evolutionary processes. Such investigations provide unique insights into the genome's complexity and gene evolution. The origins and detection techniques of de novo genes are highlighted in this study, which gives a summary of the advancements in the field of gene evolution research. Their distinct roles in brain function, stability, and evolution are shown by the CNS roles of de novo. Their evolutionary importance and fitness roles are covered by the crucial question of de novo evolutionary roles. According to the review, a number of candidate de novo genes arise from noncoding sequences in a variety of species through a stepwise mutational process, increasing protein levels and potentially providing a source of phenotypic novelty that aids in adaptation, sex evolution, and tissue-specific traits. Human-specific characteristics are mostly shaped by de novo genes, particularly in the brain, where they contribute to higher cognitive functions, cortical expansion, and complexity. Lastly, because of their influence on the development of distinctive human characteristics, they may also contribute to a person's vulnerability to specific illnesses.

Keywords: Exon shuffling; De novo genes; Long interspersed elements(LINEs); open reading frames (ORFs).

1.0. Introduction

There are numerous striking instances of functional novelty linked to phenotypes that have changed from their ancestral states, making it difficult to determine the evolutionary origins of the derived features. The development of wing patterns in Drosophila, the formation of feathers in birds, and the origin of the vertebrate inner ear, which was derived from the modified jawbones of ancestors, are some of the most well-known instances (Weisman et al., 2021).

In the last decade, methods for generating basic gene architecture such as exon shuffling, gene duplication, and gene fusion have been identified, and our understanding of how new genes emerge in a range of animals has greatly expanded [2]. These new genes have been found to integrate into and change existing gene networks through selection and mutation, exposing new patterns and rules with consistent origination rates in a range of organisms [3]. An alternative route to duplication as the only source of novel genes has emerged over the past few decades. Genes may occasionally arise from ancestrally non-genic DNA, according to this mechanism known as de novo gene evolution. De novo gene birth is the process by which new genes arise from previously noncoding areas [4]. Previously thought to be highly unusual, de novo gene birth has since been reported in a number of diverse taxa [5, 6, 7]. Haldane and Muller's cytological observation of chromosomal duplication in the 1930s is one example of an early cytogenetics experiment, suggesting that repurposed copies of ancient genes could yield new gene functions [8]. Additionally, Hugo de Vries put forth the mutation theory of evolution, which holds that a single gene mutation can give rise to a new species [9]. Novel processes and genes have been found more quickly in the genomics age, underscoring the possible significance of de novo genes in the process of new gene origination. Many studies have described how young de novo genes that are particular to a single species can play important biological roles through species-specific molecular pathways [10][11].

The widely accepted theory that new proteins evolve from pre-existing genes has been called into question by the discovery that new genes can be created de novo from noncoding DNA regions in many animals. Consequently, de novo gene birth has garnered a lot of attention recently as a major potential source of phenotypic, structural, and genomic novelty. New genes play important roles in phenotypic and functional evolution across a range of biological processes and structures, and sexual conflict genes have demonstrable fitness effects that can impact species divergence [12].

2. Mechanisms of origin of new genes

"Exon shuffling" refers to the natural process of generating novel exon combinations by intronic recombination [13]." Based on the splice frame junctions, which distinguish three classes of introns and nine classes of exons, researchers have developed in vitro formats for exon shuffling [14]. The evolution of proteins has been guided by these forms. Alternative splicing facilitates exon shuffling, a process in which introns encourage recombination both inside and between genes. Exon shuffling is governed by splice frame principles, as demonstrated by splice frame diagrams of natural genes [15]. Numerous studies have shown that the human lineage is the source of unique genes that are essential for the development of adaptive evolutionary improvements [16][17]. Ten percent of human genes are estimated to include tandemly duplicated exons [18]. Although they are not yet listed in the genome database, 2438 unannotated exons take part in exclusive alternative splicing. Genomic sequencing analysis revealed duplicated genes in every sequenced vertebrate genome [19]. Gene copies with significantly redundant functionality are produced when the whole regulatory regions and coding sequences are reproduced in tandem, a process known as tandem, segmental, or global duplication [20]. While one copy of a pair of duplicate genes maintains the original function, the other copy may accumulate mutations to further evolve new functions, according to [21]. Domain shuffling comparison examinations of whole genomes/proteomes from humans and other eukaryotes suggest that complex domain designs encoded by human genes may be the primary source of the complexity of the human genome [22]. Combinatorial diversity and a thorough analysis of protein domains encircled by introns were the outcomes of a methodical search for evidence of shuffling events in the human genome sequence [23]. The majority of human eukaryotic genomes contain mobile elements called Alu (SINE) and LINEs, which have multiple roles in the evolution of genomes. Forty-five percent of the human genome is made up of retroposed fragments, many of which came from the primate lineage [24]. The key evolutionary success of non-LTR retrotransposons in the human lineage, which has led to ongoing activity over extended periods of time, sets them apart [25]. Although the causes of this evolutionary success are still unknown, advances in bioinformatics pipelines, model genomes, and sequencing should help shed new light on this fascinating subject. Understanding the general effects of TE on human health, genome evolution, and the distinctive characteristics of humans depends on this knowledge. It is believed that retrotransposons' notable impact on genome evolution is a consequence of their evolutionary success rather than a cause [26]. Before vertebrate diversification, when the *Homo sapiens* branch split off from the chimpanzee lineage, and sometimes in between enabled by the retroposition [27].

Whether a retrocopy-contributing gene survives or is abandoned in all or part of the succeeding lineages following a "trial period"—which can be incredibly short or persist for tens of millions of years [28]—is determined by the interaction of neutral variables and natural selection. Since point mutation evolution is a rather slow process, it is unlikely to fully explain the differences between primates and other mammals. A new, fused protein with domains from both original proteins can be produced by splicing and translating RNAs [29]. [30] Found almost 200 instances of intergenic splicing involving 421 genes in the human genome, and experimentally verified that at least half of these fusions take place in human tissues. The idea that transcription-induced chimerism can aid in the evolution of protein complexes is supported by research findings. Moreover, protein-coding genes are frequently transformed into new RNA genes by messenger RNAs derived from ancestral genes [31]. Furthermore, protein and RNA genes were created from scratch using previously nonfunctional sequences, and genomic parasites were appropriated as new genes. The majority of "new" proteins are created through tinkering, which is the modular rearrangement of preexisting domain combinations or the recruitment and adaption of smaller DNA segments from nearby genes. These rearrangements are primarily caused by fusion and terminal loss [32].

3. Origin, transcription and fixation of de novo genes

An interesting evolutionary puzzle has always been where new genes come from [33]. De novo genes are sometimes defined as genes derived from ancestral non-genic sequences. Since a gene is a DNA sequence that makes RNA, a de novo gene produces an RNA that differs from an ancestral RNA [34]. According to this description, de novo genes have two physical properties. They may be protein-coding or noncoding, to start. Secondly, they may originate from any sequence that did not previously create an orthologous RNA. There are several possible methods that produce the de novo transcripts [35]. The oldest examples are novel antisense RNA that shares sequence space with an ancestor gene and novel transcripts generated by ancestral intronic or intergenic regions [36]. In certain situations, enhancers may also exhibit promoter-like traits and generate a de novo transcript [37]. Thirdly, an ancestrally nongenic sequence can be converted into a functional de novo long noncoding RNA (lncRNA) [38]. Conversely, an ancestral, functioning noncoding gene may give birth to a derived protein-coding function, which would constitute a gain of function [39]. Research generally supports the notion that de novo gene candidates carry open reading frames (ORFs) by definition and that the acquisition or retention of ORFs is an important stage in de novo gene origination [40]. MicroRNAs and other functional short RNAs may also evolve from scratch [41]. Unlike conserved genes, de novo genes are frequently exclusive to a species or lineage, much like orphan genes. De novo genes, in contrast to most genes, often show fluctuating expression in populations, which makes sense considering how young they are? [42]. Many lineage-specific genes in genomes have been bioinformatically predicted since the advent of large-scale genome sequencing, indicating that there may be a sizable rate of de novo gene generation [43]. However, it is still unknown how many functioning de novo-originated genes there are? Also a significant percentage of these genes are probably incorrectly predicted genes. Adaptive innovation and the evolution of lineage-specific traits are significantly influenced by new genes. [44] Reported that 59 of the 60 genes that were de novo derived from the human lineage were fixed in the human population.

Compared to earlier suggested estimates, the projected de novo creation rate of 9.83–11.8 genes per million years is significantly greater. Even with this high rate, it is still less than the rate of new gene origin by gene duplication when stated in terms of each gene (0.00033–0.00039 per gene per million years).

The following are the currently proposed mechanisms of de novo origin

Epigenetic modifications

The fact that many unexpressed orphan genes are decorated with repressive histone modifications, while the absence of such modifications enhances transcription of an expressed fraction of orphans, lends credence to the theory that open chromatin promotes the synthesis of new genes [45].

Open Reading Frames/Micropeptides

A more conclusive example of the ORF emerging de novo before the promoter region is the antifreeze glycoprotein gene AFGP, which appeared de novo in Arctic codfishes [46]. Furthermore, eukaryotic genomes contain a large number of putatively non-genic ORFs long enough to encode functional peptides, which are predicted to happen frequently by accident [47].

Overprinting

The introduction of different ORF one or two frames ahead of an existing gene is referred to as overprinting. This leads to the translation of a protein that is completely different from the standard protein [48]. Moreover, it can arise downstream of the original ORF in the 3' UTR [50] or upstream of the ancestral ORF in the 5' UTR [49].

Exonization

Mutations within a gene that result in the acquisition of a new exon and possibly a new ORF are known as exonization [51]. For instance, it can happen when a splicing site is lost, turning an original intron into an exon [52].

Antisense origin

Additionally, de novo genes may arise that overlap preexisting genes on the other strand [53]. It has been revealed that transcribed antisense de novo emerging ORFs can control translation efficiency [55] or produce functional proteins [54].

From scratch in intergenic region

Future gene formation in an intergenic region requires a transcription event, an ORF, the ability to translate, some stability in the untranslated regions (UTRs), eventually introns, etc. [53].

A series of sequential steps is required for a non-genic genomic region to evolve into an expressed protein-coding gene. Initially, the region must acquire an open reading frame (ORF) capable of translation and become transcriptionally active [33]. Successful gene expression depends on the ability of the transcriptional machinery to recognize and transcribe the region. Many such transcripts are initially non-coding and lack sequence similarity or characteristic features of conventional protein-coding RNAs, leading to their classification as long non-coding RNAs (lncRNAs) [56]. Notably, a substantial fraction of lncRNAs can associate with ribosomes and localize to the cytoplasm in a manner similar to protein-coding mRNAs [57].

Comparative analyses of human, chimpanzee, and macaque tissues have shown that transcripts corresponding to 24 human de novo *genes* are present in other primates as lncRNAs but are not translated, suggesting that translation emerged specifically in the human lineage [58]. Since the first identification of human-specific de novo *genes*, tentative links to disease have been proposed. More recently, the de novo *gene* NYCM, found exclusively in human and chimpanzee genomes, has been implicated in the development of human neuroblastoma [59]. Furthermore, knockdown studies of a transcript containing a human-specific de novo ORF originating within an endogenous retrovirus demonstrated that the transcript plays an essential role in maintaining cellular pluripotency, providing the first experimental evidence supporting the functional importance of de novo *genes* in the human lineage.

[60]. Recent acquisition of downstream splice sites and upstream transcription start sites may be linked to the expression of species-specific transcripts [61]. Despite this, there are no indications of purifying selection in the sequences of the majority of these transcripts. The significance of acquiring transcriptional regulatory sequences as a stage in de novo *gene* evolution is supported by several researchers [62]. A translatable ORF may develop in non-coding transcripts once they are produced in a particular species due to the accumulation of mutations. Over time, ORFs continue to appear in non-coding transcripts, but only a small percentage of them develop the capacity to be translated [50]. Higher ORF translatability in mammals has been linked to several factors, including the translation initiation environment, codon use, and the relative position of the initiation codon [63]. Another change is the composition of scanning complexes with varied ability to unwind RNA secondary structures, which can selectively activate ORFs with worse initiation settings under certain conditions. The human putative de novo *genes* CLLU1 and DNAH10OS illustrate the importance of insertion–deletion (indel) mutations in facilitating the emergence of novel *genes* from previously non-genic sequences [64]. In addition, some de novo *genes* may arise through mechanisms in which an intact open reading frame (ORF) already exists before transcriptional activation occurs [65]. Eukaryotic genomes contain numerous intergenic and intronic ORFs [66]; although these structures occur frequently, they are typically not subject to selective pressure and are continuously gained and lost over evolutionary time. In *Drosophila melanogaster*, several de novo ORFs have been identified that existed prior to the acquisition of transcriptional activity by the host RNA transcript [67]. In closely related *Drosophila* species, orthologous ORFs may be conserved at the genomic level while the corresponding loci remain transcriptionally silent [68].

A key question concerns how translated ORFs become established as functional protein-coding *genes* and whether this transition is largely stochastic. The prevailing hypothesis proposes that occasional expression of random ORFs generates a reservoir of proto-genes from which advantageous variants are retained through evolutionary selection [69]. Proto-genes are RNA transcripts originating from non-genic regions that contain translated ORFs. This concept was first extensively studied in *Saccharomyces cerevisiae*, where approximately 1,000 proto-genes differ from those found in closely related yeast species [71]. Only a small subset of these proto-genes appears to have evolved into functional genes under selective pressure, while most show no evidence of purifying selection or stable translation. Similar evolutionary patterns have been observed in humans, mice, and *Drosophila* species [65].

Another model, known as the preadaptation or "all-or-nothing" model, suggests that de novo genes emerge only when sequences already possess properties that minimize cellular toxicity, as non-functional intermediates could be detrimental to the cell [72,73]. Under this framework, young genes often exhibit higher intrinsic structural disorder (ISD) compared to older genes over long evolutionary timescales [74]. However, correlations between evolutionary age and characteristics such as aggregation propensity or ISD are less clear over shorter evolutionary periods [75]. Gene and ORF lengths generally increase with evolutionary age, whereas newly formed genes tend to be shorter and contain fewer exons [76]. Younger genes are also typically expressed at lower levels and often show tissue-specific expression patterns [77]. In *S. cerevisiae*, ISD does not strongly correlate with de novo gene emergence, whereas thymine-rich sequences associated with transmembrane domains appear to contribute to gene formation. Notably, about 67.9% of proto-genes in humans and apes contain at least one intron [78]. An example is the MYEOV protein, which possesses a putative transmembrane domain and appears to have originated de novo in the human genome [79]. While structured proteins may integrate more effectively into cellular pathways, they may also carry increased risks of misfolding and cytotoxicity.

Interestingly, many newly evolved genes are preferentially expressed in the testes. The testis is considered a hotspot for gene innovation in animals due to strong selective pressures related to sperm competition, sexual selection, and reproductive isolation. Furthermore, the chromatin environment in spermatocytes and spermatids favors the initial transcription of emerging genes [81]. Widespread demethylation of CpG-rich promoter regions and the presence of modified histones increase transcriptional activity, leading to permissive or promiscuous transcription of many genomic regions, including nonfunctional sequences and newly emerging de novo genes [82].

The fixation of a de novo gene within a population likely differs substantially from the fixation of genes arising through partial or complete duplication of pre-existing genes [83]. In gene duplication events, the resulting copy is initially redundant and typically confers neither an immediate selective advantage nor disadvantage; although functional, it generally lacks novelty at the time of origin [84]. In contrast, studies have shown that even arbitrary sequences may exhibit selectable variation under certain conditions [85]. However, if such sequences are not expressed, neither advantageous nor deleterious open reading frames (ORFs) can be subjected to natural selection, preventing their refinement or elimination [86].

If an arbitrary ORF were suddenly expressed at high levels, it would more likely produce deleterious effects rather than confer benefits, particularly when encoding longer proteins prone to misfolding or cellular toxicity. Nevertheless, considerable interest in de novo genes stems from their potential to evolve entirely new biological functions within relatively short evolutionary timeframes [87]. Although well-characterized examples remain limited, some de novo genes have already been linked to important biological processes and disease. For instance, the human-specific de novo gene FLJ33706 shows highest expression in brain tissue, with increased expression observed in Alzheimer's disease, and a single nucleotide polymorphism within this gene has been associated with addiction disorders [88,89].

Likewise, knockdown studies have demonstrated that the human-specific gene ESRG plays a critical role in maintaining pluripotency in human naïve stem cells, providing further evidence that de novo genes can rapidly acquire essential cellular functions [90].

4. De novo Genes in Brain Development, Function, and Disease

Human de novo genes can be derived from evolutionarily neutral long non-coding RNA (lncRNA) locations, which have evolutionary significance [91]. However, it is unknown how and why this all-or-nothing flip to protein-coding activity takes place. RNA-seq expression data suggest possible functions for some of the de novo protein-coding genes [92]. Newly generated de novo genes and enhancers are much more common in the human brain, particularly in the fetal cortex, where they contribute to evolutionary novelties including increasing brain size and cognitive complexity [93]. These genes are commonly produced from non-coding DNA (e.g., lncRNAs) and exhibit distinct expression patterns that contribute to the development of human-specific brain structures such as folded cortex (gyrification). The cerebral cortex expresses more de novo genes than other examined organs [94]. The cerebral cortex, the wrinkled gray matter that covers the cerebral hemispheres, is responsible for the majority of cognitive abilities. The cerebral cortex plays a vital role in cognition, awareness, and consciousness [95]. Research has concentrated on the origins and evolution of human cognitive capacities, as well as the effects of positive natural selection on brain growth, genes, and expression changes. Studies have linked new genes created by gene duplication, new microRNAs, and novel regulatory mechanisms for old genes to various human brain development features [96].

Human-specific de novo genes, such as **SP0535**, are highly expressed in the ventricular zone of the developing fetal brain and have been shown to contribute to neuronal proliferation and cortical expansion, highlighting their potential importance in human brain development [97]. Another important aspect of brain biology involves alternative splicing of genes expressed in neural tissues. Alternative splicing of precursor mRNA is a key regulatory mechanism that increases transcriptomic and proteomic diversity while also controlling mRNA levels post-transcriptionally [98]. This process is particularly frequent in brain tissues and is essential during multiple stages of nervous system development, including cell fate specification, neuronal migration, axon guidance, and synapse formation [99].

Increasing evidence indicates that some de novo genes participate in these developmental processes. Studies have demonstrated that many human- or hominoid-specific de novo genes show enriched expression in brain tissue, suggesting their contribution to brain evolution and functional complexity [100]. In addition to de novo gene formation, genomic innovations such as gene fusion, fission, or recombination also provide rapid evolutionary novelty [101]. These rearrangements often preserve functional protein domains while placing them in new genomic contexts, sometimes conferring immediate selective advantages or disadvantages. Examples include genes such as FOXP1, NOVA1, and NOVA2, which encode neuron-specific RNA-binding proteins involved in alternative splicing and play critical roles in neurodevelopment [102].

Regulatory changes occurring before and after the onset of neurogenesis have also been proposed to explain the expansion of the primate brain.

Enlargement of the neocortex may result from increased numbers of neuroepithelial cells, while cortical layer thickening is associated with expansion of radial glial cells (RGCs) [103]. Given that many de novo genes display brain-enriched expression, especially in the human fetal brain, it is plausible that they contribute adaptively to brain development [95]. Comparative studies using human and chimpanzee cortical organoids to model early brain development revealed species-specific differences, including prolonged prometaphase metaphase duration in human neural progenitors, a reduced proportion of neurogenic basal progenitors, and delayed maturation of human neural tissues [104].

Further investigations into the functional implications of newly evolved genes have examined conservation patterns and gene-expression relationships across species. Transcriptome analyses of macaque brains identified genes whose expression correlates with macaque orthologues of human de novo genes, enabling the construction of gene co-expression networks relevant to brain development [105]. Experimental systems using human embryonic stem cells and cortical organoids have also provided insight into genes regulating cortical development. These organoids exhibit brain-like organization, with PAX6-positive radial glial cells and CTIP2-positive neurons forming structures analogous to the ventricular zone and cortical plate of the developing neocortex. One hominoid-specific de novo gene, ENSG00000205704, contains functional splice and U1 recognition sites and encodes a 107-amino-acid protein localized to both the cytoplasm and nucleus. The gene is broadly expressed in human brain tissues, with expression increasing during brain and cortical organoid development [106], comparative transcriptome sequencing across multiple mammalian species—including humans, chimpanzees, macaques, and mice—combined with genomic comparisons has revealed signatures of natural selection in a subset of de novo genes with confirmed evidence of protein translation, further supporting the functional relevance of newly evolved genes in mammalian evolution [107].

The majority of these young genes, some of which generated de novo, are expressed in the neocortex, which is assumed to be responsible for many aspects of human cognition. Many of these young genes exhibit positive selection signatures, and functional annotations demonstrate that they are engaged in a wide range of molecular activities, with transcriptional regulatory genes being particularly enriched relative to other functional classes [108]. One such example is *FLJ33706*, a de novo gene discovered in GWAS and linkage analysis for nicotine addiction that is overexpressed in the brains of Alzheimer's patients [109]. In general, the fetal human brain expresses more immature, primate-specific genes than the mouse brain [110]. De novo acquired enhancers, activated by single-nucleotide mutations, are selected for cognitive qualities and remain active throughout neocortical development. [92] developed a novel sequence-specific deep learning model of embryonic neocortical enhancers using H3K27ac signatures from developing human and macaque brains. The study demonstrates that single nucleotide mutations result in a widespread de novo increase in enhancers in the progenitors and interneurons of the developing human neocortex. Human de novo genes can be produced from neutral long non-coding RNA (lncRNA) locations, which have evolutionary significance. *SP0535* is a protein-coding gene unique to humans that was created from scratch [95].

From an evolutionary perspective, these genes act as a "hopeful monster" mechanism, allowing for rapid functional adaptation and, in certain cases, contributing to the development of cognitive abilities and neurodevelopmental disorders [111]. This strategy entails rapid reconfiguration of brain circuitry to generate cognitive abilities de novo (new) [112]. According to a recent study, the embryonic neocortex contains roughly 4000 de novo enhancers that are specific to humans and not seen in macaques. These enhancers help to develop human cognitive abilities [113]. Several single-nucleotide mutations (a sort of macromutation) can create a new enhancer that activates a key transcription factor (e.g., POU3F2, SOX TFs), causing a dramatic shift in brain development [114]. These newly acquired enhancers mostly act on progenitor cells and interneurons, which may enlarge the cortical surface and increase the complexity of brain connections in a short evolutionary timeframe [115].

Beyond their evolutionary significance, de novo gene formation also has important implications for human health [116]. Newly evolved genes, particularly those unique to specific lineages, may contribute to species-specific biological traits, although many of these genes still lack detailed functional annotation. Nevertheless, growing evidence indicates that certain human-specific de novo genes are involved in disease processes, including cancer [117]. For example, NYCM, a de novo gene found only in humans and chimpanzees, has been shown to influence neuroblastoma development in experimental models, while the primate-specific long non-coding RNA PART1 has been reported to function either as a tumor suppressor or as an oncogene depending on cellular context [105].

Furthermore, due to their roles in early brain development, de novo genes have also been implicated in several neuropsychiatric disorders, including autism spectrum disorder (ASD), intellectual disability (ID), and schizophrenia. De novo mutations (DNMs) in genes such as SCN2A are frequently observed across these disorders [118], while mutations in SLC6A1 have been associated with schizophrenia [119]. Additionally, pathogenic DNMs in genes including HNRNPU, WAC, and RYR2 have been linked to intellectual disability [120]. These findings highlight the growing recognition that recently evolved genes and de novo mutations contribute to both human-specific traits and susceptibility to neurological and developmental disorders.

5. Tools and Techniques in the Detection of De Novo Genes

The initial step in identifying de novo genes or proto-genes involves selecting candidate genes from a particular species, population, or individual genome. Identification typically begins with annotated genomes, where researchers evaluate whether annotated genes lack homologs in related taxa and may therefore have originated de novo within a specific evolutionary lineage [121]. Another widely used strategy involves analyzing transcriptomic datasets, in which transcripts expressed in one or more tissues or developmental stages are examined to identify potential novel genes. However, transcriptome-based approaches usually depend on the availability of a well-annotated reference genome for accurate mapping and interpretation [122].

De novo genes may arise from various genomic regions, including intergenic spaces, intronic sequences, untranslated regions (UTRs), or through overlapping arrangements with existing genes in alternative reading frames or antisense orientations.

Depending on the hypothesized mechanism of gene emergence, certain transcripts or ORFs may be excluded during candidate selection. Tools such as BEDtools facilitate the determination of genomic overlaps between transcripts and annotated features, enabling researchers to filter and retain candidate transcripts for downstream analysis [123]. Appropriate application of such tools depends on the characteristics of the RNA-seq library used, including strand specificity.

Following genomic filtering, selected spliced transcripts are screened to identify open reading frames. Multiple computational tools are available for ORF extraction, including EMBOSS getorf, which provides information on ORF position and orientation within transcripts [124]. Additional software approaches help determine which ORFs are likely to represent true coding sequences, often using protein-based prediction methods and comparative analyses [125,126]. Homology searches are then conducted to verify novelty, typically using BLAST due to its balance of speed and accuracy, although faster alternatives such as DIAMOND are often preferred when searching large databases like NCBI nr or RefSeq [127].

A major challenge in de novo gene identification is variation in genome and transcriptome data quality among species, which can influence detection accuracy. Many candidate de novo genes are short, lineage-specific, and may lack clear evidence of protein-coding capacity. Moreover, statistical methods used to detect evolutionary conservation often require substantial sequence divergence to achieve significance, limiting their effectiveness for short sequences [128]. Currently, two principal strategies dominate systematic searches for new genes: genomic phylostratigraphy, which traces gene age based on evolutionary conservation patterns, and synteny-based methods, which examine conservation of genomic neighborhood to determine gene origin [129].

Genomic phylostratigraphy

Genomic phylostratigraphy evaluates each gene within a focal species to determine whether homologous sequences exist in ancestral lineages, typically using BLAST-based sequence alignment or related computational approaches [130]. By mapping the evolutionary origin of genes across phylogenetic levels, this method estimates the relative age of genes and identifies those potentially unique to specific taxonomic groups. However, an important limitation of phylostratigraphy arises from its reliance on sequence similarity, which makes it difficult to distinguish whether a gene truly originated de novo or represents an ancient gene that has diverged so extensively that homologous relationships are no longer detectable. To address this limitation, more sensitive similarity-search techniques can be employed. Methods such as context-specific BLAST (CS-BLAST) and Hidden Markov Model (HMM)-based searches can be used either independently or in combination to improve detection of distant homologues and reduce false identification of genes as lineage-specific [131].

Synteny-based approaches

Synteny-based approaches, although historically limited in throughput, are increasingly being applied in genome-wide surveys to identify de novo genes and represent a promising direction for improving gene age estimation methods [132]. These approaches detect potential non-genic ancestral regions of candidate de novo genes by examining conserved syntenic blocks in outgroup species, where the relative order and arrangement of genomic features remain preserved over evolutionary time [133].

Syntenic alignments typically rely on conserved sequence "anchors" to identify corresponding genomic regions across species. Protein-coding genes commonly serve as such markers, although shorter conserved elements such as k-mers or exonic sequences may also be used [134,135]. Early primate genomic studies identified approximately 270 orphan genes unique to humans, chimpanzees, and macaques, among which 15 were proposed to have originated de novo [136]. Subsequent investigations reported around 60 human-specific de novo genes supported by both transcriptional and proteomic evidence, strengthening the case for their functional expression [137]. To ensure robust identification, conservative filtering strategies are often applied, excluding candidate genes lacking syntenic sequence information in related species or those showing evidence of paralogs, distant homologues, or conserved genomic regions, which could indicate alternative evolutionary origins.

Comparative Approaches

To distinguish between de novo and ancient gene selection signatures, evolutionary traits of de novo genes should be compared to comparison sets of ancient genes with similar characteristics, such as short length and low expression [98].

Population Genetic Approaches

The identification and investigation of genetic variations responsible for de novo gene expression may enable more insightful population genetic assessments of segregating or newly fixed genes [138]. An alternative approach to studying the role of selection on de novo genes found in only one species is based on the idea that for a protein-coding de novo gene, the presence of functional constraints predicts that nonsynonymous site heterozygosity (pN) should be lower than synonymous site heterozygosity (pS) [139].

6. Evolutionary implications of de novo genes

Evolution is defined as the change in inherited traits of biological populations over repeated generations [140]. Available molecular tools and rapidly developing genome data from different creatures reveal significant variance in the number of genes among them. The evolution of living beings is the result of two processes. First, evolution is dependent on the genetic variety caused by mutations, which occur on a constant basis within populations. Second, it depends on variations in the frequency of alleles within populations throughout time [141]. Natural selection influences the fate of mutations that have an impact on the carrier's fitness. Charles Darwin's early investigations provided a scientific justification for the hypothesis of evolution by natural selection. Natural selection provides ample evidence for evolutionary change from a variety of sources [142]. The mechanism of change over time produces changes in the qualities (traits) of organisms within lineages from generation to generation, and its components include variation, inheritance, increased population growth, and differential survival and reproduction. Furthermore, Haldane and Fisher argued that in the presence of recurrent mutation, one member of a duplicate pair eventually becomes nonfunctional; that is, most duplicates should eventually die out as pseudogenes, while the other gene undergoes neofunctionalization, acquiring new function [143]. Kimura's Neutral Theory of Evolution has been fundamental to molecular evolution research, in part because it allows for robust predictions that can be evaluated against actual data [144].

According to the neutral theory, most molecular variation has no effect on fitness, and so the evolution of genetic variation is best described by stochastic processes.

De novo genes, which arise from previously noncoding DNA, are critical for evolutionary innovation because they allow for rapid adaptation to new environments, promote phenotypic novelty, and contribute to species-specific traits such as human proto-oncogenes, male-specificity in Drosophila, and antifreeze protein in Antarctic fish.

They usually start off as brief, chaotic sequences that quickly transform into organized, functional patterns that influence adaptation. However, many of these sequences have been lost, emphasizing the importance of natural selection in developing these new genes from the ground up.

De novo genes are responsible for evolutionary innovation and adaptation. Evidence supports the notion that de novo genes—genes derived from previously noncoding, intergenic DNA—provide a rapid way to gain novel protein functions and have a distinct advantage over the slower, more gradual process of gene modification. Rice studies have revealed significant patterns of rapid structural change in de novo genes over a relatively brief evolutionary span of one to two million years. [145] discusses the amazing qualities of de novo genes and encoded proteins, including high binding affinities, low molecular weights, positive net charges, diffusion, and interactions with other proteins. Similarly, [146] describes the independent emergence of antifreeze glycoprotein genes (afgps) in Arctic codfish and Antarctic notothenioids to prevent freezing. Furthermore, de novo genes evolve significantly quicker than older genes, exhibiting higher rates of adaptive modifications, accelerated sequence evolution, and frequently acquiring specialized tissue or gender-biased expression, as demonstrated in Drosophila. These immature genes typically arise from intergenic areas, particularly in the testis, and can evolve critical functions in a short evolutionary timeframe [147][148]. These genes' expression is dramatically increased and temporally enlarged in malignancies, which is related with extrachromosomal DNA amplification [149]. Novel proteins can result in unique biological properties, such as functions in development, disease, and complex characters, with examples discovered in a wide range of animals. De novo genes in plants have been linked to stress tolerance, reproductive success, and developmental regulation [150]. Many de novo genes are just ephemeral, but a tiny percentage become important, indicating how natural selection modifies these novel genetic components for specific biological tasks, many of which are [98]. In evolutionary biology, de novo genes act as selective filters by providing a continuously created, highly variable raw source of genetic material (usually in the form of short ORFs, or "proto-genes") that is subject to natural selection.

These genes have a high turnover rate (rapid gain and loss) and serve as a reservoir for adaptation and innovation, especially in harsh environmental situations. Few examples include the *BSC4* and *MDF1* genes, which are two of the most well-characterized, canonical examples of de novo evolved protein-coding genes [151]. *The QQS* orphan gene in plants modulates carbon and nitrogen allocation [152] and the *NCYM*, a Cis-antisense gene of *MYCN*, a proto-oncogene [153]. Identifying de novo gene functions and fitness effects is challenging, partly due to their overall smaller fitness effects compared to conserved genes. The birth and subsequent spread of young de novo genes may be affected by many processes that are directly or indirectly related to population genetics.

For example, organisms with a smaller effective population size or less recombination may have more noncoding DNA and more deleterious variants associated with spurious transcription, resulting in higher de novo gene birth rates than species with larger populations and more recombination [154] [155]. De novo transcripts have a high turnover rate at the population level. While there are a few examples when conserved gene loss may be adaptive, it appears that the turnover of young de novo genes represents drift loss as the selection environment changes over time [156]. An essential goal is to investigate the processes and mechanisms that underpin the rapid birth and death of functional de novo genes. The most strongly deleterious novel ORFs and transcripts are likely to be rapidly eliminated by purifying selection, making it difficult to identify such deleterious de novo ORFs in natural populations, even for model species with relatively abundant population genome and phenotypic data [157]. How these elements, as well as the distinct biological or ecological characteristics of various species, organs, and cell types, influence de novo gene origination and spread is an essential field of future research. There are indications that local adaptation may play a role in the evolution of young de novo genes [158].

7.Discussion

Although the majority of identified new genes in the human genome appear to have resulted from duplication-related mechanisms, current research suggests that genes can also develop de novo from ancestrally non-genic sequences. Thus, studying de novo-originated genes provides several chances to learn about the origin and functioning of novel genes, as well as their regulatory mechanisms and the evolutionary processes that drive them. These studies provide insight into the complexity of the human genome and gene evolution in the post-human sequencing age. With the completion of the human genome project, one of the key areas of genomics investigated is the origin of genes and their functions in Homo sapiens evolution [159]. Parallel efforts in primates and other model genomes make this process easier, and big data storage and enhanced sensitivity bioinformatics technologies help as well. Several unique problems arise in this particular section of the human genome, including when and how de novo genes develop, as well as under what conditions. The second question concerns the protein structure, location, functions, and interactions of de novo coded proteins, while the third concerns the genes' developmental and evolutionary roles. Several lines of research have revealed that the human genome has more protein-coding regions than previously thought. However, a distinctive fact is that the majority of newly discovered translated open reading frames (ORFs) are evolutionary young and limited to humans or primates, implying that the genome has evolved slowly but steadily. It is also worth noting that several of the proteins encoded by these genes do not conform to the standard definition of a protein. Several of these genes arose spontaneously from non-genic ancestral sequences. Investigating de novo-originated genes provides numerous chances to better understand their origins and functions, regulatory mechanisms, and associated evolutionary processes. These findings provide new insights into the intricacy of the human genome and gene evolution. Several mechanisms for the de novo genesis of genes have been validated experimentally and through bioinformatic pipelines. Enhancers mediated de novo transcript synthesis under specified conditions, de novo long noncoding RNA (lncRNA) origins from ancestral nongenic sequence, and the acquisition of a new exon and ORF are some

examples. Thus, researchers have sufficient opportunity to interrogate the genome data about the origins and trajectory of eukaryotes. The assumption that de novo gene candidates have ORFs, and that their acquisition or retention is critical to de novo gene origination, is now largely accepted in research. In the domain of functional short RNAs led from the De novo genes and evolution has provided unequivocal evidence for the likely roles of microRNAs in the human genome and evolution, particularly about their activities in the CNS. Small proteins or micropeptides, which are typically less than 50 or 100 amino acids, can be found throughout the enormous genomic landscape [160]. The association between micropeptides and de novo genes is only beginning to be studied, leaving many unsolved concerns. Because of their shorter length, micropeptides may play a role in de novo gene origination. Many found micropeptides are classified as noncoding regions, implying a possible transition from nonprotein-coding sequences to functional coding regions via the development of tiny genes (de novo). Micropeptides are involved in the regulation of cellular homeostasis, metabolism, and development, and they frequently function as fine-tuners for complex biological processes. Dysregulation of micropeptides has been linked to a variety of diseases, including cancer, cardiovascular disease, and neurodegeneration.

The acquisition of (rudimentary) protein domains through elongation of a de novo protein is a mechanism for a de novo gene to integrate into existing molecular machinery, which explains why neutrally developing peptides survive despite purifying selection. These alterations are chosen for a new product in order to generate unique phenotypes useful to the organism, several of which are covered in the review, de novo genes have fewer particular origins and are more lineage specific. Furthermore, de novo genes demonstrate varied expression in populations, which is understandable given their young age. These findings suggest the evolution of novelty, tissue-specificity, or improved evolution.

They also facilitate adaptation to new environments. Another part of the human genome that is being studied more vigorously around the world is the role of RNA transcripts from non-coding sequences with a translated ORF. Given the similar mechanisms of proto-genes in humans, mice, and Drosophila, proto-genes appear to have evolved from the proto-gene pool into genes, shifting from a neutral proto-gene state to a translated gene under selection, which fits the "all-or-nothing" transition origin of denovo genes. Neuroscience research has focused on the origins and evolution of human cognitive skills. These studies have related new genes created by gene duplication, new microRNAs, and novel regulatory mechanisms for old genes to specific human brain development features.

De novo genes are now linked to the formation of the human and ape brains. They are strongly expressed in the fetal neocortex, particularly in progenitor and interneuron cells, which promotes cortical growth and folding, allowing for the development of distinct human cognitive processes. Several genes have been annotated to support this finding. De novo genes are the source of evolutionary innovation and adaptability, as proven by the proteins found in many model organisms and their roles. Thus, the study of de novo genes and related corollary issues provides researchers with an exciting avenue for investigation in the eukaryotic genome, with implications for basic, evolutionary, and medicinal research. Cumulatively, de novo genes have been identified in diverse species, with many candidate de novo genes emerging from noncoding sequences through a stepwise mutational process, contributing to increased protein, with significant possible sources of phenotypic novelty contributing to adaptation, evolution of sex and tissue-specific traits. The De novo genes play significant roles in shaping human-specific traits, especially in the brain, contributing to higher cognitive functions and susceptibility to certain diseases.

Tables

| Sl.no	Organisms	Denovo genes	Functions
1	*D.melanogaster*	5	X-linked,testis based expression
2	*D.Yokuba and D.erecta*	7+3	X-linked
3	*D.melanogaster subgroup*	1	Testis based expression
4	*D.willistoni*	15	Testis based expression
5	*D.melanogaster*	208(fixed) proto-genes	Testis based expression, male biased expression
	P. xylostella		Lushu, testis expression
6	Primates(*H.sapiens, P.troglodytes, M. mullata*)	15	*PART1*, Prostrate carcinogenesis
7	hominids	24	RNA expression, cerebellum expression
8	*H.sapiens*	3	*CLLU1*, upregulated in chronic lymphoid leukemia
9	*H.sapiens*	1	*FLJ33706*, neuronal expression
10	*M.musculs*	1	*Poldi*, testis expression
11	*M.musculs*	1	○ *Pldi* (Placenta-derived transcript male fertility and regulates sperm motility
12	*R.norvegicus*	6	*Poldi*, testis expression
13	*Oryza*	1	*OsDR10*, defence gene
14	*A.thaliana*	1	*QQS*. Starch biosynthesis
15	*A.thaliana*	1	Sun Wu-Kong (*SWK*), seed germination in osmotic stress
16	*Rose*	1	*SCREP, rose scent*
17	*P.Vivax*	13	Introns in coding sequence
18	*T.trichiura*	3	Mitochondrial *cox1*, *rrnL*, and *nd1* genes
18	*S.Cerevisae*	1	*BSC4*, lethal, DNA repair
19	*S.Cerevisae*	1	*MDF1*, promotes vegetative growth
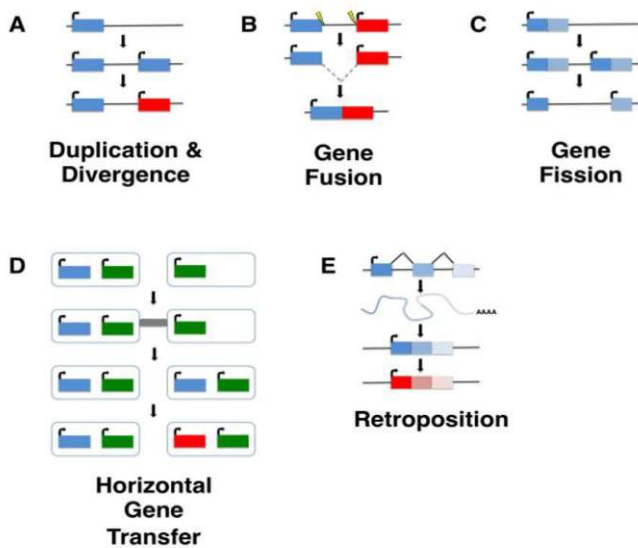
Figure-1. Illustration of the common mechanisms involved in Gene evolution in the human genome(Oss and Carvunis.Plos genetics 2019)
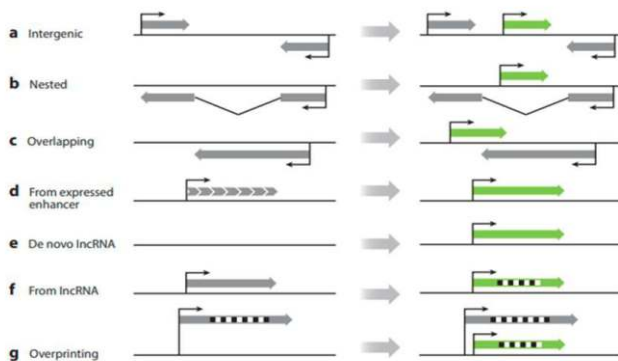

Figure-2. Locations de novo transcript and mechanisms of de novo gene origination.(Zhao, Zao et al., Ann.Rev.Gent 2024)
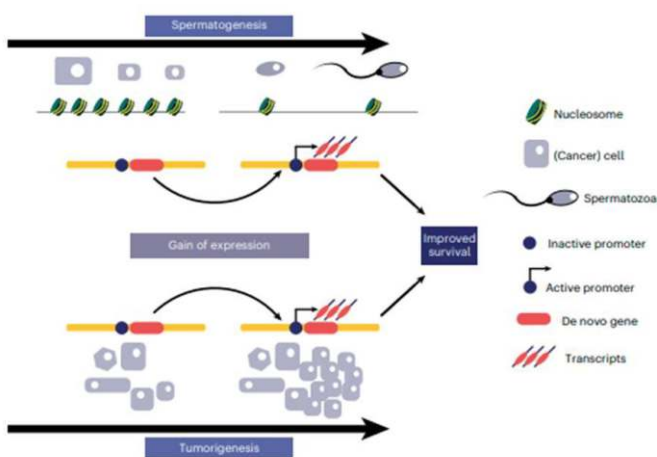

Figure-3.Illustartion of tissue specific advantage and gain of function attributes of the de novo genes.(Broeils et al., Nature ecology and evolution 2023.)

References

1. Weisman, C.M. (2022). The Origins and Functions of De Novo Genes: Against All Odds? J Mol Evol, 90(3-4):244-257. doi: 10.1007/s00239-022-10055-3.
2. Sowjanya, B. A., Narayana, B. D., & Shreyas, S. (2019). Origin and evolution of new genes: A review. Journal of Pharmacognosy and Phytochemistry, 8(2): 2394–2405.
3. Long, M., VanKuren, N.W., Chen, S., & Vibranovski, M.D. (2013). New gene evolution: little did we know. Annu Rev Genet, 47:307-33. doi: 10.1146/annurev-genet-111212-133301.
4. Van Oss, S. B., & Carvunis, A. R. (2019). De novo gene birth. PLoS Genetics, 15(5): e1008160. https://doi.org/10.1371/journal.pgen.1008160.
5. Baalsrud, H.T., Tørresen, O.K., Solbakken, M.H., Salzburger, W., Hanel, R., Jakobsen, K.S., & Jentoft, S. (2018). De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. Mol Biol Evol, 35(3):593-606. doi: 10.1093/molbev/msx311.
6. Weisman, C. M. (2022). The origins and functions of de novo genes: Against all odds? Journal of Molecular Evolution, 90: 244–257. https://doi.org/10.1007/s00239-022-10055-3.
7. Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K., & Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife, 8:e44392. https://doi.org/10.7554/eLife.44392.
8. Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Research, 20(10):1313-1326. doi: 10.1101/gr.101386.109.
9. Nei, M., & Nozawa, M. (2011). Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. Genome Biol Evol, 3:812-29. doi: 10.1093/gbe/evr028.
10. Bornberg-Bauer, E., & Eicholt, L.A. (2026). Emergence and evolution of protein-coding de novo genes. Nat Rev Genet. https://doi.org/10.1038/s41576-025-00929-9.
11. Cherezov, R. O., Vorontsova, J. E., Kuvaeva, E. E., Akishina, A. A., Zavoloka, E. L., & Simonova, O. B. (2025). The lawc gene emerged de novo from conserved genomic elements and acquired a broad expression pattern in Drosophila. Journal of Genetics and Genomics, 52(7): 901–914. https://doi.org/10.1016/j.jgg.2024.12.014.
12. Grandchamp, A., Aubel, M., Eicholt, L.A., Roginski, P., Luria, V., Karger, A., & Dohmen, E. (2025). De Novo Gene Emergence: Summary, Classification, and Challenges of Current Methods. Genome Biol Evol, 17(11):evaf197. https://doi.org/10.1093/gbe/evaf197.
13. Singh, P.K., Singh, P., Singh, R.P., & Singh, R.L. (2021). Chapter 2 - From gene to genomics: tools for improvement of animals. Editor(s): Mondal, S., & Singh, R.L. Advances in Animal Genomics. Academic Press, 13-32. https://doi.org/10.1016/B978-0-12-820595-2.00002-3.
14. Al-Balool, H.H., Weber, D., Liu, Y., Wade, M., Guleria, K., Nam, P.L., Clayton, J., Rowe, W., Coxhead, J., Irving, J., Elliott, D.J., Hall, A.G., Santibanez-Koref, M., & Jackson, M.S. (2011). Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant. Genome Res, 21(11):1788-99. doi: 10.1101/gr.116442.110.
15. Patthy, L. (2021). Exon Shuffling Played a Decisive Role in the Evolution of the Genetic Toolkit for the Multicellular Body Plan of Metazoa. Genes (Basel), 12(3):382. doi: 10.3390/genes12030382.
16. Llamas, B., Willerslev, E., & Orlando, L. (2017). Human evolution: A tale from ancient genomes. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1713): 20150484. https://doi.org/10.1098/rstb.2015.0484.
17. Pollen, A.A., Kilik, U., Lowe, C.B., & Camp, J.G. (2023). Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. Nat Rev Genet, 24(10):687-711. doi: 10.1038/s41576-022-00568-4.
18. Ivanov, T.M., & Pervouchine, D.D. (2022). Tandem Exon Duplications Expanding the Alternative Splicing Repertoire. Acta Naturae, 14(1):73-81. doi: 10.32607/actanaturae.11583.
19. Verma, P., Thakur, D., & Pandit, S. B. (2024). Exon nomenclature and classification of transcripts database (ENACTdb): A resource for analyzing alternative splicing mediated proteome diversity. Bioinformatics Advances, 4(1): vbae157. https://doi.org/10.1093/bioadv/vbae157.

20. Kono, T.J.Y., Brohammer, A.B., McGaugh, S.E., & Hirsch, C.N. (2018). Tandem Duplicate Genes in Maize Are Abundant and Date to Two Distinct Periods of Time. G3 (Bethesda), 8(9):3049-3058. doi: 10.1534/g3.118.200580.

21. Birchler, J.A., & Yang, H. (2022). The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell, 34(7):2466-2474. doi: 10.1093/plcell/koac076.

22. Alvarez-Ponce, D., & Krishnamurthy, S. (2025). Organismal complexity strongly correlates with the number of protein families and domains. Proceedings of the National Academy of Sciences of the United States of America, 122: e2404332122. https://doi.org/10.1073/pnas.2404332122.

23. Koeppel, J., et al. (2025). Randomizing the human genome by engineering recombination between repeat elements. Science, 387: eado3979. https://doi.org/10.1126/science.ado3979.

24. Navarro, F.C., & Galante, P.A. (2015). A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biol Evol, 7(8):2265-75. doi: 10.1093/gbe/evv142.

25. Platt, R.N., Vandewege, M.W., & Ray, D.A. (2018). Mammalian transposable elements and their impacts on genome evolution. Chromosome Res, 26: 25–43. https://doi.org/10.1007/s10577-017-9570-z.

26. Ferrari, R., Grandi, N., Tramontano, E., & Dieci, G. (2021). Retrotransposons as Drivers of Mammalian Brain Evolution. Life, 11(5):376. https://doi.org/10.3390/life11050376.

27. Casola, C., & Betrán, E. (2017). The genomic impact of gene retrocopies: What have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biology and Evolution, 9(6): 1351–1373. https://doi.org/10.1093/gbe/evx081.

28. Zhang, W., & Tautz, D. (2022). Tracing the Origin and Evolutionary Fate of Recent Gene Retrocopies in Natural Populations of the House Mouse. Mol Biol Evol, 39(2):msab360. doi: 10.1093/molbev/msab360.

29. Tao, Y., Zhang, Q., Wang, H. *et al.* (2024). Alternative splicing and related RNA binding proteins in human health and disease. Sig Transduct Target Ther, 9: 26. https://doi.org/10.1038/s41392-024-01734-2.

30. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R. Transcription-mediated gene fusion in the human genome. Genome Res. 2006 Jan;16(1):30-6. doi: 10.1101/gr.4137606. Epub 2005 Dec 12.Sampath, K., & Ephrussi, A. (2016). CncRNAs: RNAs with both coding and non-coding roles in development. Development, 143(8), 1234–1241. https://doi.org/10.1242/dev.133298.

31. Tang, H., Peng, Q., Oyang, L., Tan, S., Jiang, X., Ren, Z., Xu, X., Shen, M., Li, H., Peng, M., Xia, L., Yang, W., Li, S., Wang, J., Han, Y., Wu, N., Tang, Y., Lin, J., Liao, Q., & Zhou, Y. (2025). Fusion genes in cancers: Biogenesis, functions, and therapeutic implications. Genes Dis, 12(5):101536. doi: 10.1016/j.gendis.2025.101536.

32. McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation 1. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1678), 20140332. https://doi.org/10.1098/rstb.2014.0332.

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

34. Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (2022). A simple guide to de novo transcriptome assembly and annotation. Briefings in Bioinformatics, 23(2): bbab563. https://doi.org/10.1093/bib/bbab563.

35. Bartonicek, N., Clark, M.B., Quek, X.C., Torpy, J.R., Pritchard, A.L., Maag, J.L.V., Gloss, B.S., Crawford, J., Taft, R.J., Hayward, N.K., Montgomery, G.W., Mattick, J.S., Mercer, T.R., & Dinger, M.E. (2017). Intergenic disease-associated regions are abundant in novel transcripts. Genome Biol, 18(1):241. doi: 10.1186/s13059-017-1363-3.

36. Majic, P., & Payne, J.L. (2020). Enhancers Facilitate the Birth of De Novo Genes and Gene Integration into Regulatory Networks. Mol Biol Evol. 2020 Apr 1;37(4):1165-1178. doi: 10.1093/molbev/msz300.

37. Chillón, I., & Marcia, M. (2020). The molecular structure of long non-coding RNAs: emerging patterns and functional implications. Critical Reviews in Biochemistry and Molecular Biology, 55(6): 662–690. https://doi.org/10.1080/10409238.2020.1828259.

38. Schmitz, J.F., & Bornberg-Bauer, E.(2017) Fact or fiction: updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA. F1000Res, 6:57. doi: 10.12688/f1000research.10079.1.

39. Bornberg-Bauer, E., & Eicholt, L.A. (2026). Emergence and evolution of protein-coding de novo genes. Nat Rev Genet. https://doi.org/10.1038/s41576-025-00929-9.

40. Lu, S. (2019). De novo origination of MIRNAs through generation of short inverted repeats in target genes. RNA Biol, 16(6):846-859. doi: 10.1080/15476286.2019.1593744.

41. Acuna-Hidalgo, R., Veltman, J.A. & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome Biol, 17:241. https://doi.org/10.1186/s13059-016-1110-1.

42. Lu, Y., Li, M., Gao, Z., Ma, H., Chong, Y., Hong, J., Wu, J., Wu, D., Xi, D., & Deng, W. (2025). Advances in Whole Genome Sequencing: Methods, Tools, and Applications in Population Genomics. Int J Mol Sci, 26(1):372. doi: 10.3390/ijms26010372.

43. Wu, D.D., Irwin, D.M., & Zhang, Y.P. (2011). De novo origin of human protein-coding genes. PLoS Genet, 7(11):e1002379. doi: 10.1371/journal.pgen.1002379.

44. Werner, M.S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C., & Sommer, R.J. (2018). Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Res, 28(11):1675–1687. doi:10.1101/gr.234872.118.

45. Zhuang, X., Yang, C., Murphy, K.R., & Cheng, C.C. (2019). Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1817138116.

46. Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., et al. (2012). Proto-genes and de novo gene birth. Nature, 487(7407):370–374. doi:10.1038/nature11184.

47. Carter, J.J., et al. (2013). Identification of an overprinting gene in merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. Proc Natl Acad Sci U S A., 110:12744–12749. https://doi.org/10.1073/pnas.1303526110.

48. Renz, P.F., Valdivia-Francia, F., & Sendoel, A. (2020). Some like it translated: small ORFs in the 5'utr. Exp Cell Res, 396:112229. https://doi.org/10.1016/j.yexcr.2020.112229.

49. Wu, Q., Wright, M., Gogol, M.M., Bradford, W.D., Zhang, N., & Bazzini, A.A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. EMBO J, 39(17):e104763. doi: 10.15252/embj.2020104763.

50. Schmitz, J., & Brosius, J. (2011). Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie, 93:1928–1934. https://doi.org/10.1016/j.biochi.2011.07.014.

51. Koralewski, T.E., & Krutovsky, K.V. (2011). Evolution of exon-intron structure and alternative splicing. PLoS One, 6:e18055. https://doi.org/10.1371/journal.pone.0018055.

52. Iyengar, B.R., & Bornberg-Bauer, E. (2023). Neutral models of de novo gene emergence suggest that gene evolution has a preferred trajectory. Mol Biol Evol, 40:msad079. https://doi.org/10.1093/molbev/msad079.

53. Thomas, K.E., Gagniuc, P.A., & Gagniuc, E. (2023). Moonlighting genes harbor antisense ORFs that encode potential membrane proteins. Sci Rep, 13:12591. https://doi.org/10.1038/s41598-023-39869-x.

54. Liang, X-H., Shen, W., Sun, H., Migawa, M.T., Vickers, T.A., & Crooke, S.T. (2016). Translation efficiency of mRNAs is increased by antisense oligonucleotides targeting upstream open reading frames. Nat Biotechnol, 34:875–880. https://doi.org/10.1038/nbt.3589.

55. Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.L., Chen, R., Dean, C., Dinger, M.E., Fitzgerald, K.A., Gingeras, T.R., Guttman, M., Hirose, T., Huarte, M., et al. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol, 24(6):430-447. doi: 10.1038/s41580-022-00566-8.

56. Thomas, K.E., Gagniuc, P.A., & Gagniuc, E. (2023). Moonlighting genes harbor antisense ORFs that encode potential membrane proteins. Sci Rep, 13:12591. https://doi.org/10.1038/s41598-023-39869-x.

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

58. Carlevaro-Fita, J., Rahim, A., Guigó, R., Vardy, L.A., & Johnson, R. (2016). Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. RNA, 22(6):867-82. doi: 10.1261/rna.053561.115.

59. Delihas, N. (2020). Formation of human long intergenic non-coding RNA genes, pseudogenes, and protein genes: Ancestral sequences are key players. PLoS ONE, 15(3): e0230236. https://doi.org/10.1371/journal.pone.0230236.

60. Suenaga, Y., Islam, S.M., Alagu, J., Kaneko, Y., Kato, M., Tanaka, Y., Kawana, H., Hossain, S., Matsumoto, D., Yamamoto, M., Shoji, W., Itami, M., Shibata, T., Nakamura, Y., Ohira, M., Haraguchi, S., Takatori, A., & Nakagawara, A. (2014). 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, 10(1):e1003996. doi: 10.1371/journal.pgen.1003996.

61. Izsvák, Z., Ma, J., Singh, M., & Hurst, L.D. (2025). Co-option of an endogenous retrovirus (LTR7-HERVH) in early human embryogenesis: becoming useful and going unnoticed. Mob DNA, 16(1):27. doi: 10.1186/s13100-025-00361-0.

62. Leader, Y., Lev Maor, G., Sorek, M., Shayevitch, R., Hussein, M., Hameiri, O., Tammer, L., Zonszain, J., Keydar, I., Hollander, D., Meshorer, E., & Ast, G. (2021). The upstream 5' splice site remains associated to the transcription machinery during intron synthesis. Nat Commun, 12(1):4545. doi: 10.1038/s41467-021-24774-6.

63. Rebeiz, M., Patel, N.H., & Hinman, V.F. (2015). Unraveling the Tangled Skein: The Evolution of Transcriptional Regulatory Networks in Development. Annu Rev Genomics Hum Genet, 16:103-31. doi: 10.1146/annurev-genom-091212-153423.

64. Xu, C., & Zhang, J. (2020). Mammalian Alternative Translation Initiation Is Mostly Nonadaptive. Mol Biol Evol, 37(7):2015-2028. doi: 10.1093/molbev/msaa063.

65. Schmitz, J., & Brosius, J. (2011). Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie, 93:1928–1934. https://doi.org/10.1016/j.biochi.2011.07.014.

66. Schrider, D. R., & Kern, A. D. (2015). Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. Genome Biology and Evolution, 7(12), 3511–3528. https://doi.org/10.1093/gbe/evv228.

67. Grandchamp, A., Berk, K., Dohmen, E., & Bornberg-Bauer, E. (2022). New genomic signals underlying the emergence of human proto-genes. Genes, 13: 284. https://doi.org/10.3390/genes13020284.

68. Huang, Y., Chen, S.-Y., & Deng, F. (2016). Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. Computational and Structural Biotechnology Journal, 14: 298–303. https://doi.org/10.1016/j.csbj.2016.07.002.

69. Zhao, L., Saelao, P., Jones, C.D., & Begun, D.J. (2014). Origin and spread of de novo genes in Drosophila melanogaster populations. Science, 343(6172):769-72. doi: 10.1126/science.1248286.

70. Zheng, E.B., & Zhao, L. (2022). Protein evidence of unannotated ORFs in *Drosophila* reveals diversity in the evolution and properties of young proteins. Elife, 11:e78772. doi: 10.7554/eLife.78772.

71. Papadopoulos, C., Callebaut, I., Gelly, J.C., Hatin, I., Namy, O., Renard, M., Lespinet, O., & Lopes, A. (2021). Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. Genome Res, 31(12):2303-2315. doi: 10.1101/gr.275638.121.

72. Weisman, C. M., & Eddy, S. R. (2017). Gene evolution: Getting something from nothing. Current Biology, 27(13):R661–R663. https://doi.org/10.1016/j.cub.2017.05.056.

73. Seçkin, E., Colinet, D., Sarti, E., & Danchin, E. G. J. (2025). Orphan and de novo genes in fungi and animals: Identification, origins, and functions. Genome Biology and Evolution, 17(12): evaf220. https://doi.org/10.1093/gbe/evaf220.

74. Wilson, B.A., Foy, S.G., Neme, R., & Masel, J. (2017). Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of *De Novo* Gene Birth. Nat Ecol Ev, l.1(6):0146-146. doi: 10.1038/s41559-017-0146.

75. Schmitz, J., & Brosius, J. (2011). Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie, 93:1928–1934. https://doi.org/10.1016/j.biochi.2011.07.014.

76. Heames, B., Buchel, F., Aubel, M. *et al.* (2023). Experimental characterization of de novo proteins and their unevolved random-sequence counterparts. Nat Ecol Evol, 7: 570–580. https://doi.org/10.1038/s41559-023-02010-2.

77. Nielly-Thibault, L., & Landry, C.R. (2019). Differences Between the Raw Material and the Products of *de Novo* Gene Birth Can Result from Mutational Biases. Genetics, 212(4):1353-1366. doi: 10.1534/genetics.119.302187.

78. Foy, S.G., Wilson, B.A., Bertram, J., Cordes, M.H.J., & Masel, J. (2019). A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution. Genetics, 211(4):1345-1355. doi: 10.1534/genetics.118.301719.

79. Lebherz, M., Iyengar, B., & Bornberg-Bauer, E. (2024). De novo ORFs are more likely to shrink than to elongate during neutral evolution. bioRxiv. https://doi.org/10.1101/2024.02.12.579890.

80. Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., Mazin, P.V., Xenarios, I., et al. (2019). Gene expression across mammalian organ development. Nature, 571(7766):505-509. doi: 10.1038/s41586-019-1338-5.

81. Bozorgmehr, J. H. (2024). Two "de novo" genes in Saccharomyces cerevisiae may be highly diverged duplicates. ResearchGate. https://doi.org/10.13140/RG.2.2.24875.82721.

82. Papamichos, S.I., Margaritis, D., & Kotsianidis, I. (2015). 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), 2015:984706. doi: 10.1155/2015/984706.

83. Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Research, 20(10):1313-1326. doi: 10.1101/gr.101386.109.

84. Yuen, B.T., Bush, K.M., Barrilleaux, B.L., Cotterman, R., & Knoepfler, P.S. Histone H3.3 regulates dynamic chromatin states during spermatogenesis. Development, 141(18):3483-94. doi: 10.1242/dev.106450.

85. Donkin, I., & Barrès, R. (2018). Sperm epigenetics and influence of environmental factors. Molecular Metabolism, 14: 1–11. https://doi.org/10.1016/j.molmet.2018.02.006.

86. Lange, A., Patel, P.H., Heames, B. et al. (2021). Structural and functional characterization of a putative de novo gene in *Drosophila*. Nat Commun, 12:1667. https://doi.org/10.1038/s41467-021-21667-6.

87. Hunter, P. (2022). Understanding redundancy and resilience: Redundancy in life is provided by distributing functions across networks rather than back-up systems: Redundancy in life is provided by distributing functions across networks rather than back-up systems. EMBO Rep, 23(3): e54742. doi: 10.15252/embr.202254742.

88. Good, B.H., Walczak, A.M., Neher, R.A., & Desai, M.M. (2014). Genetic Diversity in the Interference Selection Limit. PLoS Genet, 10(3): e1004222. https://doi.org/10.1371/journal.pgen.1004222.

89. Patraquim, P., Mumtaz, M.A.S., Pueyo, J.I. et al. (2020). Developmental regulation of canonical and small ORF translation from mRNAs. Genome Biol, 21:128. https://doi.org/10.1186/s13059-020-02011-5.

90. Wu, D.-D., & Zhang, Y.-P. (2013). Evolution and function of de novo originated genes. Molecular Phylogenetics and Evolution, 67(2), 541–545. https://doi.org/10.1016/j.ympev.2013.02.013.

91. Li, C.Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P.W., Lu, S.J., Li, X.M., Yu, Q., Zheng, X., Du, Q, Uhl, G.R., Liu, Q.R., & Wei, L. (2010). A human-specific de novo protein-coding gene associated with human brain functions. PLoS Computational Biology, 6(3):e1000734. https://doi.org/10.1371/journal.pcbi.1000734.

92. Li, S., Liu, H., Liu, W., Shi, N., Zhao, M., Wanggou, S., Luo, W., Wang, L., Zhu, B., Zuo, X., Xie, W., Zhao, C., Zhou, Y., Luo, L., Gao, X., Jiang, X., & Ren, C. (2023). *ESRG* is critical to maintain the cell survival and self-renewal/pluripotency of hPSCs by collaborating with MCM2 to suppress p53 pathway. Int J Biol Sci, 19(3):916-935. doi: 10.7150/ijbs.79095.

93. Shiraishi, T., & Matsumoto, A. (2025). From non-coding to coding: The importance of long non-coding RNA translation in de novo gene birth. Biochimica et Biophysica Acta (BBA) – General Subjects, 1869(2). https://doi.org/10.1016/j.bbagen.2024.130747.

94. Costa, S.S.F., Rosikiewicz, M., Roux, J., et al. (2022). Robust inference of expression state in bulk and single-cell RNA-Seq using curated intergenic regions. bioRxiv; doi: 10.1101/2022.03.31.486555.

95. Li, S., Hannenhalli, S., & Ovcharenko, I. (2023). De novo human brain enhancers created by single-nucleotide mutations. Sci Adv, 9(7):eadd2911. doi: 10.1126/sciadv.add2911.

96. Zimmer-Bensch, G. (2019). Emerging Roles of Long Non-Coding RNAs as Drivers of Brain Evolution. Cells, 8(11):1399. doi: 10.3390/cells8111399.

97. Moini, J., Gutierrez, A., & Avgeropoulos, N. (2023). Function and dysfunction of the cerebral lobes. In J. Moini, A. Gutierrez, & N. Avgeropoulos (Eds.), Clinical neuroepidemiology of acute and chronic disorders (pp. 133–152). Academic Press. https://doi.org/10.1016/B978-0-323-95901-8.00012-2.

98. Qi, J., Mo, F., An, N.A., Mi, T., Wang, J., Qi, J.T., Li, X., Zhang, B., Xia, L., Lu, Y., Sun, G., Wang, X., Li, C.Y., & Hu, B. (2023). A Human-Specific De Novo Gene Promotes Cortical Expansion and Folding. Adv Sci (Weinh), 10(7):e2204140. doi: 10.1002/advs.202204140.

99. Tao, Y., Zhang, Q., Wang, H. *et al.* (2024). Alternative splicing and related RNA binding proteins in human health and disease. Sig Transduct Target Ther, 9: 26. https://doi.org/10.1038/s41392-024-01734-2.

100. Bae BI, Jayaraman D, Walsh CA. Genetic changes shaping the human brain. Dev Cell. 2015 Feb 23;32(4):423-34. doi: 10.1016/j.devcel.2015.01.035.

101. Li, C., et al. (2025). Alternative splicing categorizes organ development by stage and reveals unique human splicing variants linked to neuromuscular disorders. Journal of Biological Chemistry, 301(6), 108542.

102. Casola, C., Luria, V., Vakirlis, N., & Zhao, L. (2025). De novo genes: Current status and future goals. Genome Biology and Evolution, 17(12): evaf230. https://doi.org/10.1093/gbe/evaf230.

103. Zhou, Y., Zhang, C., Zhang, L. *et al.* (2022). Gene fusion as an important mechanism to generate new genes in the genus *Oryza*. Genome Biol, 23:130. https://doi.org/10.1186/s13059-022-02696-w.

104. Piton, A. (2025). NOVA1/2 genes and alternative splicing in neurodevelopment. Current Opinion in Genetics & Development, 93. https://doi.org/10.1016/j.gde.2025.102373.

105. Benito-Kwiecinski, S., Giandomenico, S.L., Sutcliffe, M., Riis, E.S., Freire-Pritchett, P., Kelava, I., Wunderlich, S., Martin, U., Wray, G.A., McDole, K., & Lancaster, M.A. (2021). An early cell shape transition drives evolutionary expansion of the human forebrain. Cell, 184(8):2084-2102.e19. doi: 10.1016/j.cell.2021.02.050.

106. Mora-Bermúdez, F., Badsha, F., Kanton, S., Camp, J.G., Vernot, B., Köhler, K., Voigt, B., Okita, K., Maricic, T., He, Z., Lachmann, R., Pääbo, S., Treutlein, B., & Huttner, W.B. (2016). Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife, 5:e18683. doi: 10.7554/eLife.18683.

107. An, N.A., Zhang, J., Mo, F., Luan, X., Tian, L., Shen, Q.S., Li, X., Li, C., Zhou, F., Zhang, B., Ji, M., Qi, J., Zhou, W.Z., Ding, W., Chen, J.Y., Yu, J., Zhang, L., Shu, S., Hu, B., & Li, C.Y. (2023). De novo genes with an lncRNA origin encode unique human brain developmental functionality. Nat Ecol Evol, 7(2):264-278. doi: 10.1038/s41559-022-01925-6.

108. https://www.genecards.org/

109. Li, S., Hannenhalli, S., & Ovcharenko, I. (2023). De novo human brain enhancers created by single-nucleotide mutations. Sci Adv, 9(7):eadd2911. doi: 10.1126/sciadv.add2911.

110. Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., & Albà, M.M. (2015). Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet, 11(12):e1005721. doi: 10.1371/journal.pgen.1005721.

111. González-Buenfil, R., Vieyra-Sánchez, S., Quinto-Cortés, C.D., Oppenheimer, S.J., Pomat, W., Laman, M., Cervantes-Hernández, M.C., Barberena-Jonas, C., Auckland, K., Allen, A., Allen, S., Phipps, M.E., Huerta-Sanchez, E., Ioannidis, A.G., Mentzer, A.J., & Moreno-Estrada, A. (2024). Genetic Signatures of Positive Selection in Human Populations Adapted to High Altitude in Papua New Guinea. Genome Biol Evol, 16(8): evae161. doi: 10.1093/gbe/evae161.

112. Li, C.Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P.W., Lu, S.J., Li, X.M., Yu, Q., Zheng, X., Du, Q, Uhl, G.R., Liu, Q.R., & Wei, L. (2010). A human-specific de novo protein-coding gene associated with human brain functions. PLoS Computational Biology, 6(3):e1000734. https://doi.org/10.1371/journal.pcbi.1000734.

113. Konopka, G. (2012). The fetal brain provides a raison d'être for the evolution of new human genes. Brain, Behavior and Evolution, 79(4): 213–214. https://doi.org/10.1159/000336723.

114. Dittrich-Reed, D.R., & Fitzpatrick, B.M. (2013). Transgressive Hybrids as Hopeful Monsters. Evol Biol, 40(2):310-315. doi: 10.1007/s11692-012-9209-0.

115. Iwama, S., Zhang, Y., & Ushiba, J. (2022). *De Novo* Brain-Computer Interfacing Deforms Manifold of Populational Neural Activity Patterns in Human Cerebral Cortex. eNeuro, 9(6):0145-22. doi: 10.1523/ENEURO.0145-22.2022.

116. Prodromidou, K., & Matsas, R. (2022). Evolving features of human cortical development and the emerging roles of non-coding RNAs in neural progenitor cell diversity and function. Cell. Mol. Life Sci, 79:56. https://doi.org/10.1007/s00018-021-04063-7.

117. Stevanovic, M., Drakulic, D., Lazic, A., Ninkovic, D.S., Schwirtlich, M., & Mojsin, M. (2021). SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. Front Mol Neurosci, 14:654031. doi: 10.3389/fnmol.2021.654031.

118. Ball, G., Oldham, S., Kyriakopoulou, V. et al. (2024). Molecular signatures of cortical expansion in the human foetal brain. Nat Commun, 15:9685. https://doi.org/10.1038/s41467-024-54034-2.

119. Samusik, N., Krukovskaya, L., Meln, I., Shilov, E., & Kozlov, A.P. (2013). PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. PLoS One, 8(2):e56162. doi: 10.1371/journal.pone.0056162.

120. Li, J, Cai, T, Jiang, Y., Chen, H., He, X., Chen, C., Li, X., Shao, Q., Ran, X., Li, Z., Xia, K., Liu, C., Sun, Z.S., & Wu, J. (2016). Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. Mol Psychiatry, 21(2):290-7. doi: 10.1038/mp.2015.40.

121. Rees, E., Han, J., Morgan, J., Carrera, N., Escott-Price, V., Pocklington, A.J., Duffield, M., Hall, L.S., Legge, S.E., Pardiñas, A.F., Richards, A.L., Roth, J., Lezheiko, T., et al. (2020). De novo mutations identified by exome sequencing implicate rare missense variants in SLC6A1 in schizophrenia. Nat Neurosci, 23(2):179-184. doi: 10.1038/s41593-019-0565-2.

122. Hamdan, F.F., Srour, M., Capo-Chichi, J.M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D., Diallo, O., Henrion, E., et al. (2014). De novo mutations in moderate or severe intellectual disability. PLoS Genet, 10(10):e1004772. doi: 10.1371/journal.pgen.1004772.

123. Roginski, P, Grandchamp, A., Quignot, C., & Lopes, A. (2024). De novo emerged gene search in eukaryotes with DENSE. bioRxiv. https://doi.org/10.1101/2024.01.30.578014.

124. Rana, K. M. S., Sonoda, T., Sato, Y., Kondo, Y., Ohtsuka, S., Kotani, T., Ueno, D., & Tasumi, S. (2025). De novo transcriptomic analysis to identify candidate genes potentially related to host recognition during infective stage of Caligus fugu (Crustacea: Copepoda). Comparative Biochemistry and Physiology Part D: Genomics and Proteomics, 54: 101433. https://doi.org/10.1016/j.cbd.2025.101433.

125. Quinlan, A.R., & Hall, I.M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.

126. Singh, U., & Wurtele, E.S. (2021). orfipy: a fast and flexible tool for extracting ORFs. Bioinformatics, 37(18):3019-3020. doi: 10.1093/bioinformatics/btab090.

127. Vitting-Seerup, K., Porse, B.T., Sandelin, A., & Waage, J. (2017). splicer: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC Bioinformatics, 15:1–7. https://doi.org/10.1186/1471-2105-15-81.

128. Varabyou, A., Erdogdu, B., Salzberg, S.L., & Pertea, M. (2023). Investigating open reading frames in known and novel transcripts using ORFanage. Nat Comput Sci., 3:700–708. https://doi.org/10.1038/s43588-023-00496-1.

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

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

131. Oss, S.B.V., & Carvunis, A.R. (2019). De novo gene birth. PLoS Genet, 15(5):e1008160. doi: 10.1371/journal.pgen.1008160.

132. Saripella, G.V., Sonnhammer, E.L., & Forslund, K. (2016). Benchmarking the next generation of homology inference tools. Bioinformatics, 32(17):2636–2641. doi:10.1093/bioinformatics/btw305.

133. Lu, T.C., Leu, J.Y., & Lin, W.C. (2017). A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. Mol Biol Evol, 34(11):2823–2838. doi:10.1093/molbev/msx210.

134. Seçkin, E., Colinet, D., Sarti, E., & Danchin, E. G. J. (2025). Orphan and de novo genes in fungi and animals: Identification, origins, and functions. Genome Biology and Evolution, 17(12): evaf220. https://doi.org/10.1093/gbe/evaf220.

135. Ghiurcuta, C.G., & Moret, B.M. (2014). Evaluating synteny for improved comparative studies. Bioinformatics, 30(12): i9–18. doi:10.1093/bioinformatics/btu259.

136. Gehrmann, T., & Reinders, M.J. (2015). Proteny: discovering and visualizing statistically significant syntenic clusters at the proteome level. Bioinformatics, 31(21):3437–3444. doi:10.1093/bioinformatics/btv389.

137. Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., et al. (2009). Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol, 26(3):603–612. doi:10.1093/molbev/msn281.

138. Wu, D.D., Irwin, D.M., & Zhang, Y.P. (2011). De novo origin of human protein-coding genes. PLoS Genet, 7(11):e1002379. doi: 10.1371/journal.pgen.1002379.

139. Cridland, J. M., Majane, A. C., Zhao, L., & Begun, D. J. (2022). Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster. Genetics, 220(1):iyab207. https://doi.org/10.1093/genetics/iyab207.

140. Montañés, J. C., Huertas, M., Messeguer, X., & Albà, M. M. (2023). Evolutionary trajectories of new duplicated and putative de novo genes. Molecular Biology and Evolution, 40(5): msad098. https://doi.org/10.1093/molbev/msad098.

141. Forbes, A. A., & Krimmel, B. A. (2010). Evolution is change in the inherited traits of a population through successive generations. Nature Education Knowledge, 3(10): 6.

142. Ashraf, M.A., & Sarfraz, M. (2015). Biology and evolution of life science. Saudi J Biol Sci, 23(1):S1-5. doi: 10.1016/j.sjbs.2015.11.012.

143. Avise, J.C., & Ayala, F.J. (2009). In the Light of Evolution: Volume III: Two Centuries of Darwin. National Academy of Sciences (US); editors.Washington (DC): National Academies Press (US).

144. Andersson, D.I., Jerlström-Hultqvist, J., & Näsvall, J. (2015). Evolution of new functions de novo and from preexisting genes. Cold Spring Harb Perspect Biol, 7(6):a017996. doi: 10.1101/cshperspect.a017996.

145. Ohta, T. (2013). Neutral Theory, Editor(s): Maloy, S., Hughes, K. Brenner's Encyclopedia of Genetics (Second Edition), Academic Press, 67-68. https://doi.org/10.1016/B978-0-12-374984-0.01039-1.

146. Chen, J., Li, Q., Xia, S., Arsala, D., Sosa, D., Wang, D., & Long, M. (2024). The rapid evolution of de novo proteins in structure and complex. Genome Biology and Evolution, 16(6): evae107. https://doi.org/10.1093/gbe/evae107.

147. Baalsrud, H.T., Tørresen, O.K., Solbakken, M.H., Salzburger, W., Hanel, R., Jakobsen, K.S., & Jentoft, S. (2018). De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. Mol Biol Evol, 35(3):593-606. doi: 10.1093/molbev/msx311.

148. Rivard, E.L., Ludwig, A.G., Patel, P.H., Grandchamp, A., Arnold, S.E., Berger, A., Scott, E.M., Kelly, B.J., Mascha, G.C., Bornberg-Bauer, E., & Findlay, G.D. (2021). A putative de novo evolved gene required for spermatid chromatin condensation in Drosophila melanogaster. PLoS Genet, 17(9):e1009787. doi: 10.1371/journal.pgen.1009787.

149. Glaser-Schmitt, A., & Parsch, J. (2025). Conservation and divergence of sex-biased gene expression across 50 million years of Drosophila evolution. bioRxiv. https://doi.org/10.1101/2025.10.10.681669.

150. Xiao, C., Liu, X., Liu, P., Xu, X., Yao, C., Li, C., Xiao, Q., Guo, T., Zhang, L., Qian, Y., Wang, C., Dong, Y., Wang, Y., Peng, Z., Han, C., Cheng, Q., An, N.A., & Li, C.Y. (2025). Oncogenic roles of young human de novo genes and their potential as neoantigens in cancer immunotherapy. Cell Genom, 5(9):100928. doi: 10.1016/j.xgen.2025.100928.

151. Zhao, L., Svetec, N., & Begun, D.J. (2024). De Novo Genes. Annual Review of Genetics, 58:211–232. https://doi.org/10.1146/annurev-genet-111523-102413.

152. Seçkin, E., Colinet, D., Bailly-Bechet, M., Seassau, A., Bottini, S., Sarti, E., & Danchin, E. G. J. (2025). Identification, evolutionary history and characteristics of orphan genes in root-knot nematodes. bioRxiv. https://doi.org/10.64898/2025.12.19.695360.

153. Li, L., & Wurtele, E.S. (2015). The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. Plant Biotechnol J, 13(2):177-87. doi: 10.1111/pbi.12238.

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

155. Latrille, T., & Lartillot, N. (2021). Quantifying the impact of changes in effective population size and expression level on the rate of coding sequence evolution. Theoretical Population Biology, 142, 57–66. https://doi.org/10.1016/j.tpb.2021.09.005.

156. Marino, A., Debaecker, G., Fiston Lavier, A.-S., Haudry, A., & Nabholz, B. (2024). Effective population size does not explain long-term variation in genome size and transposable element content in animals. bioRxiv. https://doi.org/10.1101/2024.02.26.582137.

157. Glaser-Schmitt, A., Lebherz, M., Saydam, E., Bornberg-Bauer, E. & Parsch, J. (2025). Expression of *De Novo* Open Reading Frames in Natural Populations of *Drosophila melanogaster*. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 344: 415-427. https://doi.org/10.1002/jez.b.23297.

158. Chen, Y., Feng, X., Reid, K., Zhang, C., Löytynoja, A., & Merilä, J. (2025). Dynamics of Deleterious Mutations and Purifying Selection in Small Population Isolates. Mol Biol Evol, 42(7): msaf110. doi: 10.1093/molbev/msaf110.

159. Cridland, J. M., Majane, A. C., Zhao, L., & Begun, D. J. (2022). Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster. Genetics, 220(1):iyab207. https://doi.org/10.1093/genetics/iyab207.

160. Stuglik Rlund. Human Genetics & Embryology. Opinion Volume 15:04, 2024

161. Mohsen JJ, Martel AA, Slavoff SA. Microproteins-Discovery, structure, and function. Proteomics. 2023 Dec;23(23-24):e2100211. doi: 10.1002/pmic.202100211. Epub 2023 Aug 21.