ADVANCED GENETICS Open Access
www.advgenet.com

REVIEW OPEN ACCESS
Editor's Pick

From Genomic Fossils to Functional Elements: The Evolving Story of Pseudogenes

Mengyao Sun[1] | Yanni Ma[1] | Jia Yu[1,2]

[1]State Key Laboratory of Common Mechanism Research for Major Diseases, Key Laboratory of RNA and Hematopoietic Regulation, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, China | [2]Institute of Blood Transfusion, Chengdu, China

Correspondence: Yanni Ma (mayanni@ibms.cams.cn) | Jia Yu (j-yu@ibms.pumc.edu.cn)

Received: 30 June 2025 | **Revised:** 7 November 2025 | **Accepted:** 7 November 2025

Keywords: evolution | gene duplication | non-coding genes | pseudogenes | retrotransposition

ABSTRACT

Pseudogenes, as important products of genomic evolution, play unique regulatory roles in species adaptation. This review systematically summarizes the major types, functions, and regulatory mechanisms of metazoans pseudogenes, with a particular focus on their formation during primate evolution and the mechanisms underlying their retention in the human genome. Previous studies suggest that the loss of function in pseudogenes releases selective pressure, allowing them to evolve neutrally; furthermore, their latent functional or adaptive potential, such as reactivation, neofunctionalization, or evolutionary advantages conferred by gene silencing, further promotes their persistence. For instance, the integration of certain pseudogenes can introduce novel regulatory functions, while pseudogenization-induced gene inactivation may also provide selective benefits. Recent technological advances, including long-read sequencing, single-cell omics, and CRISPR-based functional interrogation, have greatly expanded our understanding of pseudogenes. We propose that pseudogene-mediated regulation plays a critical role in evolutionary processes and highlight their dynamic roles in both physiological and pathological contexts. We further discuss current research progress, limitations, and future directions, offering new perspectives for understanding genomic evolution and biomedical significance of pseudogenes.

1 | Introduction

The human genome has currently annotated over 50 000 genes, yet only 30% of them are protein-coding genes [1]. A much larger proportion lacks protein-coding capacity and is collectively referred to as non-coding genes [2, 3]. Long non-coding RNAs (lncRNAs), short non-coding RNAs (including microRNAs, small nuclear RNAs, etc.), and pseudogenes constitute the major components of non-coding genes. Among them, pseudogenes account for approximately one-fourth of the entire human genome, second only to protein-coding genes in number, making them a vast and highly important family of non-coding genes [2, 3].

As early as 1977, Jacq et al. discovered the first pseudogene, a defective copy of 5S DNA, in Xenopus laevis [4], leading to the initial and fundamental definition of pseudogenes as truncated gene copies of functional genes [5]. However, for a long time afterward, despite the increasing identification of pseudogenes, their biological functions received little attention. Genomic regions annotated as pseudogenes were often excluded from functional screenings and genomic analyses [6–10]. It was only after the functional significance of non-coding RNAs (such as lncRNAs and microRNAs) gained recognition that some pseudogenes were found to be reactivated, and capable of transcription and even protein translation. The transcripts or encoded products

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2025 The Author(s). *Advanced Genetics* published by Wiley Periodicals LLC

derived from pseudogenes may participate in regulating various biological processes, functioning under both physiological and pathological conditions [11–14]. While most pseudogenes in the human genome exhibit high species specificity, with the vast majority being unique to higher primates, pseudogenes are generally considered to evolve under neutral selection, and thus far, only a few have been implicated in regulating species-specific evolutionary processes [15–17].

In this review, we systematically outline the fundamental characteristics of metazoan pseudogenes, including their types and sequence features across species, and summarize the known mechanisms of pseudogene reactivation and function. Although pseudogenes have received limited attention in evolutionary studies, emerging evidence suggests that pseudogenes that mediated functional innovations may contribute to species evolution. Pseudogenes also possess potential cis-regulatory elements and can even be functionally resurrected under specific selective pressures. Compared to previously published reviews on pseudogenes [18–20], our work incorporates recent research advances and emerging developments in the field, while also discussing and anticipating potential breakthroughs enabled by technological progress. More importantly, we provide a novel and rarely explored perspective, how pseudogenes may participate in species evolution, in what capacity, and through which mechanisms, offering a systematic synthesis of functional pseudogenes and their roles in evolutionary processes. In summary, we propose that functional pseudogenes may serve as a critical driving force in evolution, while the vast reservoir of currently silent pseudogenes with unknown functions may represent an essential genomic resource for evolutionary adaptation.

2 | Identification and Classification of Pseudogenes

Pseudogenes are characterized primarily by their sequence similarity to parental genes and the presence of disabling mutations, such as frameshifts, premature stop codons, and so on [21]. To maximize the accuracy of pseudogene annotation, current methodologies combine computational pipelines with manual curation to distinguish pseudogenes from their functional counterparts [3, 22, 23]. Sequence alignments enable the inference of pseudogene origins, facilitating their annotation and classification.

2.1 | Bioinformatic Approaches for Pseudogene Identification

Bioinformatic identification of pseudogenes is a multi-step process that relies on integrated strategies, with the core challenge being the distinction between pseudogenes and their functional parental genes. The standard pipeline begins with initial homology-based screening using tools such as BLAST or BLAT to align whole-genome sequences or assembled transcripts against reference protein-coding databases, enabling the detection of highly homologous yet potentially defective sequences as candidate pseudogenic regions. This is followed by the critical step of detecting inactivating mutations, such as premature stop codons, frameshift insertions and deletions (indels), splice site disruptions, and loss of the start codon, which serve as definitive evidence of loss of coding capacity [24]. Overall, annotated pseudogenes fall into two major categories based on the recognition features: processed pseudogenes, which specifically refer to those derived from parental gene transcripts through post-transcriptional processing, typically representing retroposed pseudogenes, and unprocessed pseudogenes, which arise without such processing and generally encompass duplicated pseudogenes and unitary pseudogenes. Different types of pseudogenes are not only generated in different ways but also differ in sequence characteristics, conservation, and even functional preferences. Specialized tools, including PseudoFinder, RetroFinder, and PPFINDER, have been developed to systematically integrate these features for genome-wide pseudogene annotation.

Evolutionary constraint analysis provides further support for pseudogene identification [21, 25]. Pseudogenes exhibit remarkable species specificity compared to protein-coding genes, primarily manifested by their distinct evolutionary patterns and genomic characteristics. Beyond differences in sequence conservation, protein-coding genes and pseudogenes are subject to fundamentally distinct selective pressures: protein-coding genes experience strong purifying selection that efficiently eliminates non-synonymous mutations disrupting protein structure or function, maintaining high sequence conservation across evolutionary timescales, while pseudogenes evolve predominantly through neutral processes, accumulating mutations at genomic background levels as “genomic fossils” that record evolutionary history without functional constraint. Thus, the calculation of nonsynonymous-to-synonymous substitution rate ratio (Ka/Ks) can serve as a proposed methodology for pseudogene identification and assessing the coding potential of putative pseudogenes, especially reliable for ancient, intact pseudogene copies that have accumulated sufficient substitutions while maintaining reading frame continuity with their progenitors.

Long-read sequencing is also extensively applied in the discovery and annotation of pseudogenes. Through high-quality long-read assembly, this technology enables the identification of previously unannotated pseudogenes that are absent from the reference genomes of humans, great apes, and other species, making it a powerful tool for novel gene discovery [26, 27]. In sum, ongoing efforts such as the GENCODE project continuously refine gold-standard pseudogene annotations mainly for human and mouse [3, 28]. Dedicated databases and resources, including Gerstein labs’ pseudogene database (http://www.pseudogene.org), psiDR (http://pseudogene.org/psidr/), and GenTree (http://gentree.ioz.ac.cn/) [3, 29, 30], provide comprehensive, curated information on pseudogenes across multiple species.

Consequently, the pseudogene repertoire is continually expanding and being refined, driven by advances in annotation software, the incorporation of evolutionary data, and the integration of long-read sequencing technologies, providing a foundation for evolutionary and functional studies.

2.2 | Classification of Pseudogenes in Metazoans

Building on relatively well-established genome annotation tools, pseudogenes are primarily categorized into three major classes

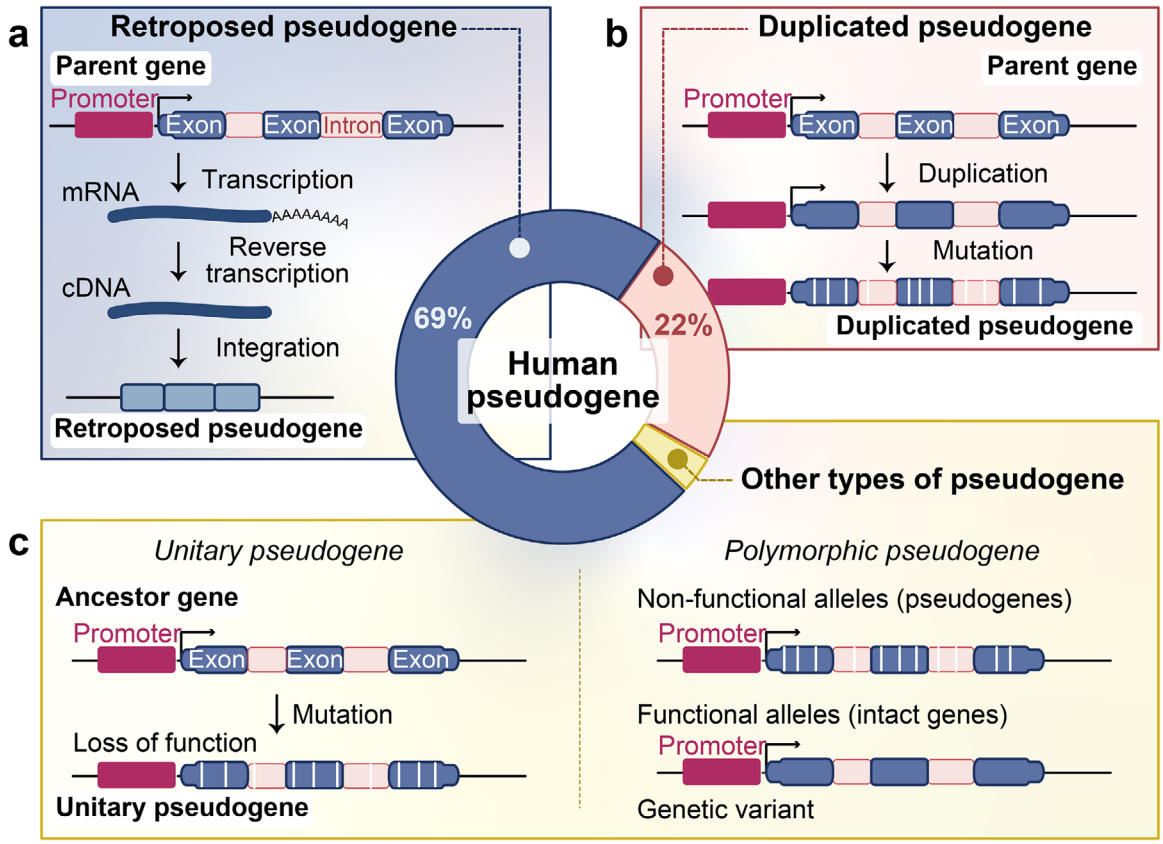

FIGURE 1 | Main types and origins of human pseudogenes. (a) Retroposed pseudogenes (~70% of total), generated via reverse transcription and genomic reintegration. (b) Duplicated pseudogenes (~20%), produced by gene duplication followed by mutation accumulation. (c) Other pseudogene types: unitary pseudogenes (ancestral gene inactivation, left) and polymorphic pseudogenes (functional alleles coexisting with non-functional variants in a population).

based on their mechanisms of formation: retroposed pseudogenes, duplicated pseudogenes, and unitary pseudogenes. Retroposed pseudogenes, also termed processed pseudogenes, are generated through retrotransposition-mediated insertion of parental gene transcripts into new genomic loci (Figure 1a) [5, 31]. Human-retroposed pseudogenes predominantly originated during the ancestral primate retrotransposition burst and currently represent the most abundant pseudogene class in the human genome, accounting for approximately 70% of all pseudogenes [32, 33]. Their formation is mechanistically linked to the transposition activity of LINE1 (L1) elements, which encode endogenous reverse transcriptases capable of co-opting the transcripts originating from distinct genes for reintegration [34]. Consequently, most retroposed pseudogenes are characterized by the absence of introns, a hallmark feature distinguishing them from their parental genes [35, 36]. Notably, approximately 8.9% of protein-coding genes are also devoid of introns and may have been generated by retrotransposons, suggesting retroposition as a significant pathway for novel gene origination [37]. Duplicated pseudogenes arise from direct genomic duplication events, retaining both intronic sequences and proximal regulatory elements from their parental genes (Figure 1b) [3]. Typically localized near their functional counterparts, duplicated pseudogenes exhibit high sequence similarity with their progenitors but are rendered nonfunctional through degenerative mutations. Derived from in situ inactivation of ancestral genes, unitary pseudogenes accumulate loss-of-function mutations while remaining genomically retained (Figure 1c) [38]. Although lacking functional parental genes in humans, orthologous functional genes often persist in other species. Additionally, there exist relatively rare pseudogene classes, such as polymorphic pseudogenes (Figure 1c), wherein a single genetic locus within a biological population concurrently harbors both functional alleles (intact genes) and non-functional alleles (pseudogenes) [3, 39]. This genetic variant may represent an intermediate evolutionary state between gene functionality and complete pseudogenization. Furthermore, two rare pseudogene types exhibit unique origins and functions. circRNA-derived pseudogenes originate from reverse-transcribed circular RNAs and display non-collinear exon arrangements relative to parental genes, potentially influencing genomic structural remodeling [40, 41]. Chimeric pseudogenes, formed by fusion of sequences from multiple genomic sources, highlight the complexity of genomic rearrangements and serve as valuable models for studying evolutionary innovation and structural diversity [42].

In sum, regardless of whether through retrotransposition, duplication, or in situ mutation and so on, pseudogenes typically lose protein-coding capacity during their generation through various molecular defects, including aberrant splicing, frameshift mutations, or premature termination codon introduction, ultimately becoming defective genes on the genome.

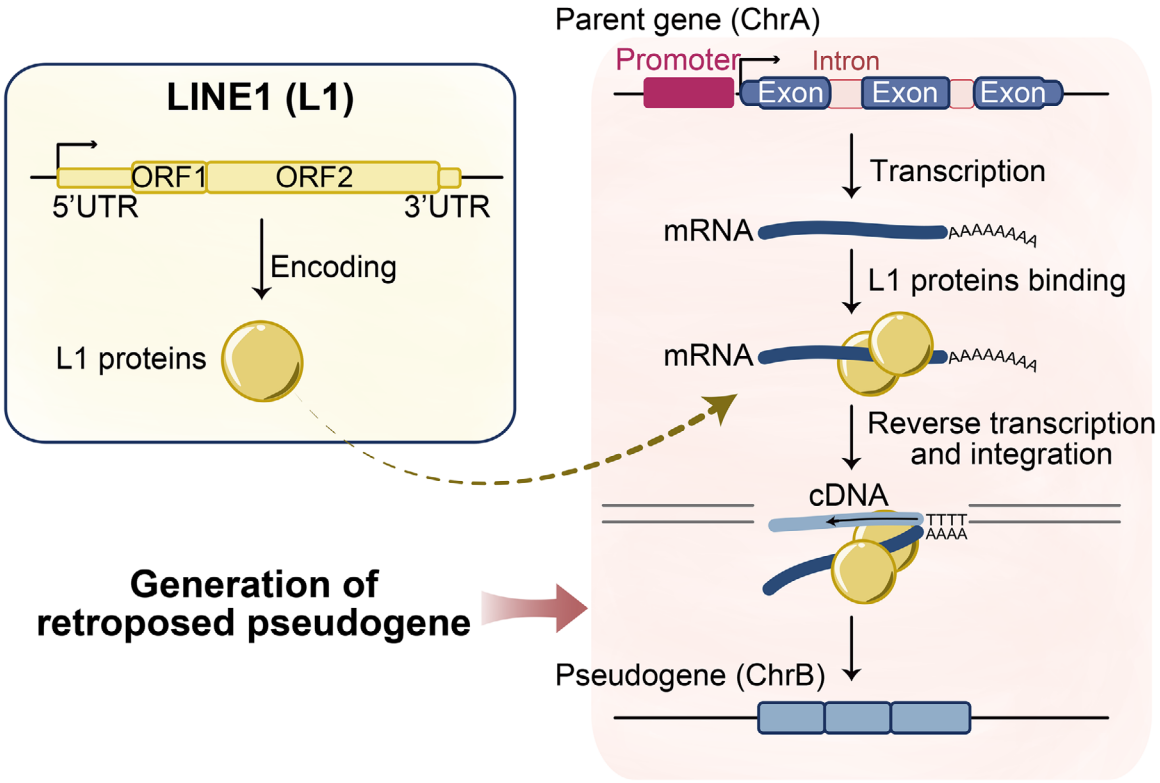

FIGURE 2 | LINE1-mediated retroposed pseudogene formation. LINE1-encoded proteins (L1 proteins), including nucleic-acid chaperone, reverse transcriptase, and endonuclease, bind gene transcripts, cleave genomic DNA, and integrate cDNA into the genome, resulting in intron-less retroposed pseudogenes typically on different chromosomes from parental genes.

2.3 | Characteristics of Major Pseudogene Types

We systematically compare the fundamental distinctions between retroposed and duplicated pseudogenes, highlighting key differential characteristics in sequence architecture and genomic distribution patterns. Retroposed pseudogenes, generating through LINE1-mediated retrotransposition (Figure 2), typically exhibit intron-less structures due to mRNA intermediate processing, truncated sequences averaging ~1.5 kb in length, frequent polyA tract remnants, and absence of parental gene regulatory elements, while duplicated pseudogenes generally maintain full-length genomic organization including intronic regions and potential retention of upstream regulatory sequences (e.g., promoters, enhancers). Regarding genomic distribution, retroposed pseudogenes demonstrate insertional bias toward regions with open-chromatin and actively expressed genes, random genomic dispersion mediated by retrotransposase activity, and transchromosomal localization with parental genes (e.g., *PTEN* on 10q23.31 versus *PTENP1* on 9p13.3) [36, 43, 44]. Whereas duplicated pseudogenes originate through non-allelic homologous recombination (NAHR) or replication errors and display paralogous chromosomal clustering (e.g., *HBB*/*HBBP1* in the β-globin locus on 11p15.4), illustrating that distinct molecular origins may fundamentally determine pseudogene architecture, genomic distribution patterns, and even regulatory potential in mammalian genomes. Moreover, retroposed pseudogenes in the human genome demonstrate pronounced primate-specificity, particularly in the New World and Old World monkey lineage [32, 33]. In contrast, duplicated pseudogenes show relatively lower species specificity, with 40% sequence conservation between human and mouse genomes (versus 10% for retroposed pseudogenes) [3].

In conclusion, understanding the distinct characteristics of each pseudogene type, including origin, structure, and evolutionary trajectory, is essential for accurately annotating genomes and interpreting their potential biological roles. Future studies leveraging long-read sequencing and functional genomics will further refine these classifications and reveal novel subclasses.

3 | Functional Pseudogenes

Once considered nonfunctional genomic relics or molecular "fossils", pseudogenes, defective paralogs of protein-coding genes, are now gradually recognized as dynamic contributors to gene regulation. Emerging evidence demonstrates that pseudogenes can regain functional activity through diverse mechanisms [45–48]. Certain pseudogenes not only transcribe regulatory noncoding RNAs but also encode functional truncated or full-length proteins, challenging the traditional view of their functionlessness theory. And even transcriptionally silent, pseudogenes may exert regulatory functions through their DNA sequences. These functional repertoires highlight pseudogenes as versatile genomic elements capable of influencing gene expression networks, phenotypic plasticity, and evolutionary innovation. Herein, we systematically elucidate the molecular mechanisms underlying pseudogene reactivation and their diverse functional repertoires in genomic regulation.

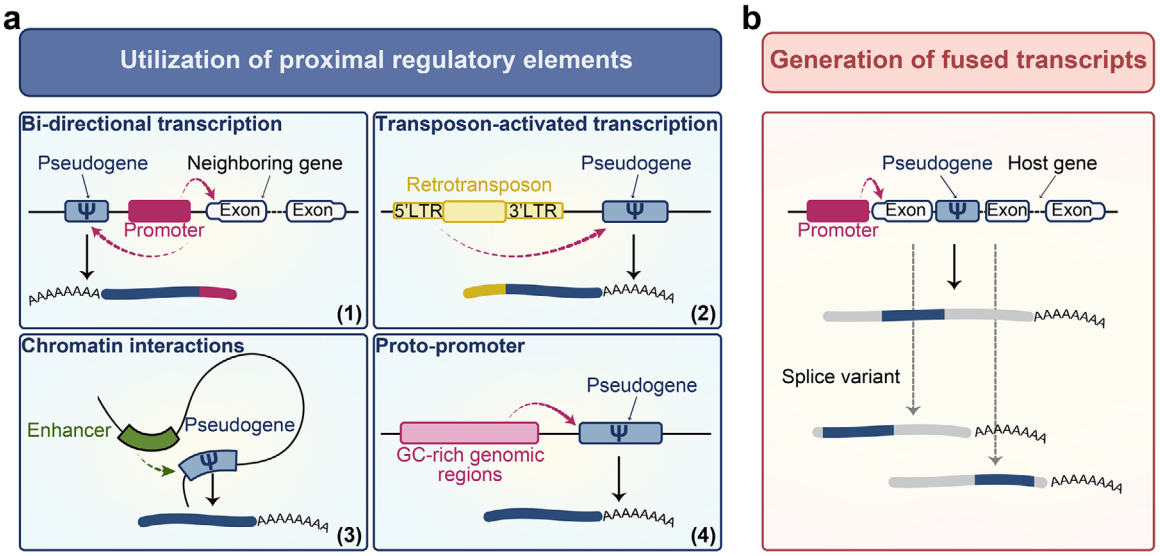

FIGURE 3 | Pathways of pseudogene reactivation. (a) Transcriptional reactivation of pseudogenes is facilitated by various pre-existing regulatory elements, including: (1) promoters from adjacent genes; (2) promoters derived from nearby transposable elements; (3) enhancers engaging in long-range chromatin interactions; and (4) GC-rich regions with intrinsic promoter potential. (b) pseudogenes derived fusion transcripts formed via co-splicing with host gene exons.

3.1 | Reactivation of Pseudogenes

Although the majority of pseudogenes in the human genome arise from retrotransposition events and therefore lack native promoter elements, their integration into the genome demonstrates significant non-randomness, strongly influenced by the epigenetic landscape of insertion sites. These sequences exhibit a pronounced preference for open chromatin regions, enriched with regulatory signatures such as DNase I hypersensitive sites, histone modifications associated with active transcription (e.g., H3K4me3, H3K27ac), and binding sites for transcription factors [36, 43, 44]. Such positioning supports their transcriptional activation through various mechanisms: certain pseudogenes integrate within intronic or exonic regions of functional genes and hijack endogenous promoters; others exploit proximal regulatory architectures, including bidirectional promoters from adjacent genes [36, 46, 49] or promoters contributed by retrotransposons like LTRs of endogenous retroviruses (Figure 3a) [36, 46, 50]. Additionally, distal enhancers can activate pseudogene expression via long-range chromatin interactions (Figure 3a) [36]. It is also noteworthy that the GC-rich genomic environments often selected by retrotransposed pseudogenes may possess intrinsic promoter-like features (Figure 3a) [51, 52]. Thus, the transcriptional competence of pseudogenes emerges from the synergistic interplay between their structural preservation and the accessibility of local and distal regulatory elements, underscoring their potential roles in gene regulatory networks.

Beyond transcriptional reactivation, pseudogenes can form fusion transcripts with adjacent or host genes via canonical or alternative splicing, repurposing their sequences into chimeric RNAs with novel functions (Figure 3b) [36, 53–56]. For instance, in prostate cancer, the pseudogene *KLKP1* splices into the *KLK4* transcript, generating an oncogenic fusion RNA [53, 57]. Alternatively, pseudogenes derived exons can introduce new protein domains, as demonstrated by a *DPH3P1*-containing *BRCA1* splice variant that produces a truncated protein with functional properties [20, 58]. These mechanisms highlight the expanding repertoire of pseudogenes driven by transcript diversification and its biological significance.

In summary, accumulating evidence demonstrates that pseudogenes are far from being genetically inert. Through either context-dependent reactivation (such as co-opting neighboring regulatory elements) or chimeric transcript formation with neighboring or host genes, some pseudogenes escape transcriptional silencing and may acquire novel biological functions.

3.2 | Major Regulatory Pathways of Pseudogenes

3.2.1 | Noncoding Mechanisms

Emerging research has established that certain pseudogenes can be transcriptionally reactivated under specific cellular contexts, producing diverse noncoding RNAs that play multifaceted regulatory roles in biological processes. These pseudogenes derived transcripts exhibit remarkable functional versatility through several well-characterized mechanisms (Figure 4a).

First, transcripts of pseudogenes may act as microRNA sponging and possess competing endogenous RNA (ceRNA) activity [59–61]. Pseudogenes derived lncRNAs frequently function as molecular decoys that sequester microRNAs, thereby modulating the expression of their parental genes, such as *PTENP1* transcripts competitively bind the *PTEN*-targeting microRNA (miR) 19/20/21 families, elevating *PTEN* mRNA levels and suppressing tumorigenesis [14], and *BRAFP1*-derived lncRNAs sponge microRNA, derepressing *BRAF* expression and activating MAPK signaling [11]. Moreover, pseudogenic lncRNAs can directly bind proteins to alter their activity or localization, including *HMGA1-p* lncRNA interacts with the RNA-stabilizing complex

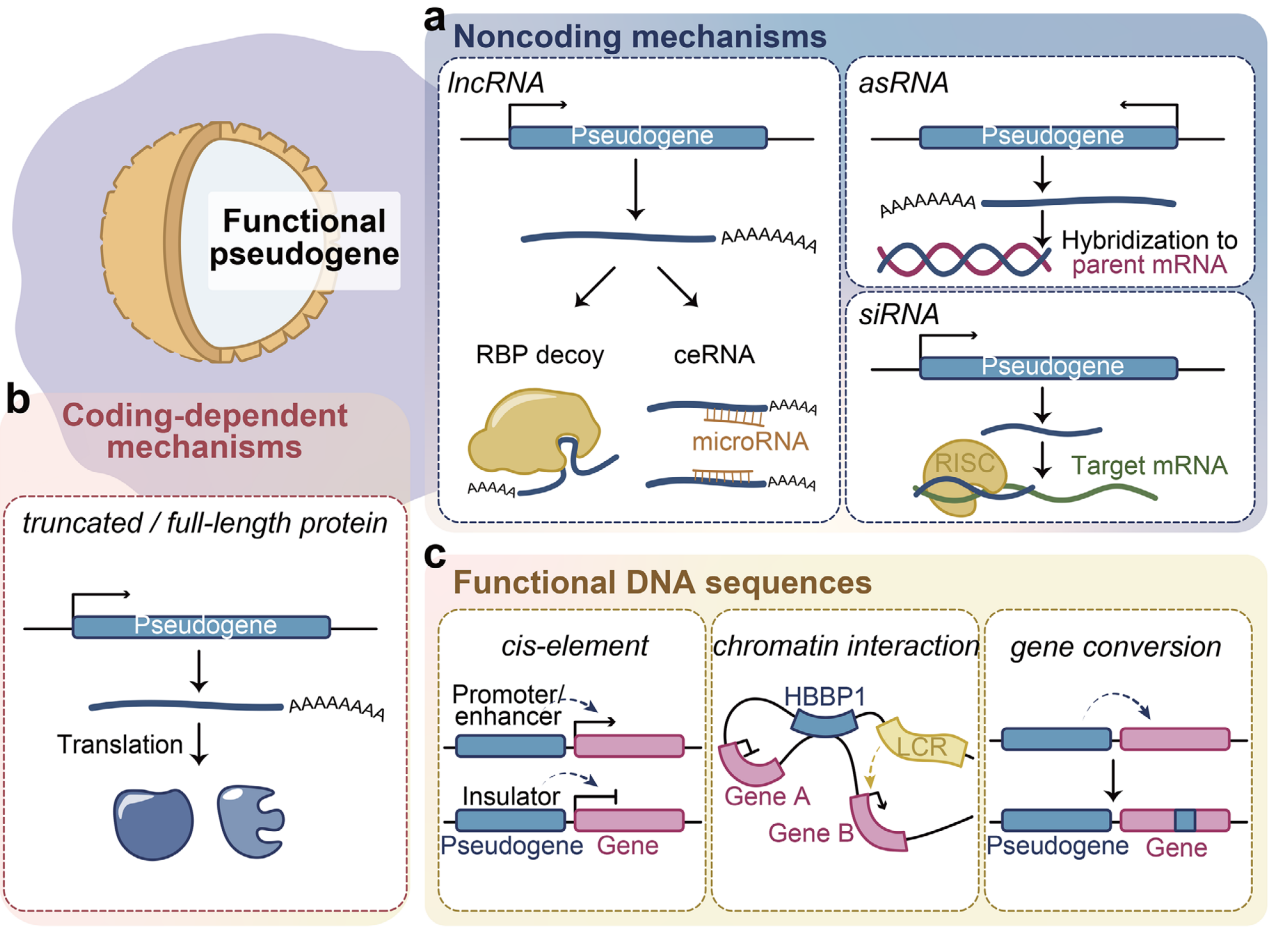

FIGURE 4 | Main functional mechanisms of pseudogenes. (a) Noncoding-dependent roles: long non-coding RNA (lncRNA), antisense RNA (asRNA), or small interfering RNA (siRNA) production. (b) Coding-dependent roles: truncated or full-length protein translation. (c) DNA sequence-mediated functions: functioning as a cis-regulatory element (left), chromatin interactions by forming loops (middle) or gene conversion (right).

αCP1, destabilizing *HMGA1* mRNA [62], *RNA5SP141* binds RIG-I during herpesvirus infection, triggering interferon production [63], and so on. Bidirectional transcription generates antisense RNAs (asRNAs) with distinct functions, and pseudogenes also can encode antisense RNAs. *PTENP1*-AS isoforms regulate *PTEN* mRNA nuclear export and stability [64], *OCT4pg5*-AS recruits EZH2 to the OCT4 promoter [65], silencing its expression, and *NOS*-AS inhibits translation of neuronal nitric oxide synthase via RNA-RNA duplex formation [66]. In addition, some pseudogenes have been found to be able to transcribe endogenous small interfering RNAs (siRNAs) and participate in the process of mRNA degradation, and pseudogenes such as *ψPPM1K*, *ATP8A2P1*, and *CNN2P1* inhibit the expression of the parent gene by transcribing siRNAs [67–70].

In summary, current research has confirmed that pseudogenes derived transcripts exert regulatory functions primarily through interactions with microRNAs, proteins, or mRNAs. Due to their sequence similarity with parental genes, pseudogene transcripts preferentially target and modulate the expression of their parental genes through bidirectional regulatory mechanisms [71]. Pseudogenes can either upregulate or downregulate parental gene expression depending on their molecular mode of action. Positive regulation generally occurs when pseudogenes function as competing endogenous RNAs that sequester microRNAs targeting the parental transcript, or when they competitively bind RNA-destabilizing proteins, thereby stabilizing parental mRNA and enhancing its expression. Conversely, negative regulation results from pseudogenes' competition for binding to RNA-stabilizing proteins or their generation of endogenous siRNAs that direct the cleavage and degradation of parental transcripts. This dual regulatory capacity enables pseudogenes to serve as precise molecular switches for fine-tuning parental gene expression levels, contributing to the maintenance of cellular homeostasis and the coordination of complex biological processes. The functional outcome, activation or repression, is determined by the specific interaction partners and molecular context of each pseudogenes parental gene pair, highlighting the sophisticated regulatory potential of these once-considered "genomic fossils".

3.2.2 | Coding-Dependent Mechanisms

While pseudogenes have been primarily characterized as sources of regulatory noncoding RNAs, accumulating evidence reveals that a biologically significant subset even retains protein-coding potential, producing functional micropeptides and full-length proteins that contribute to diverse physiological and pathological processes (Figure 4b). These protein-coding pseudogenes represent a representative exception to the traditional view of pseudogenes as nonfunctional relics [72–77].

The most typical examples demonstrate how pseudogenes derived proteins can modulate their parental gene functions through precise molecular mechanisms. The *SRGAP2C* pseudogene, for instance, encodes a truncated version of the synaptic plasticity regulator *SRGAP2* that acts as a dominant-negative inhibitor by forming nonproductive heterodimers with the full-length parental protein [15, 78]. This interaction may cause neoteny during spine maturation and increase the density of longer spines, potentially contributing to human-specific brain development patterns. Similarly, the *NOTCH2NL* gene family produces proteins that enhance cortical neurogenesis by competitively binding the NOTCH pathway inhibitor DELTA-LIKE1, thereby amplifying NOTCH signaling in neural progenitor cells [17, 79]. These cases illustrate how pseudogenes encoded proteins can evolve dosage-sensitive regulatory relationships with their parental genes. Beyond neurodevelopmental roles, several oncogenic pseudogenes derived proteins have been identified. The *POU5F1B* pseudogene encodes a protein resembling the OCT4 transcription factor that cooperates with MYC signaling in gastric cancer, with overexpression correlating with poor prognosis [80]. Similarly, *NANOGP8* produces a protein that maintains stem-like properties in cancer cells, potentially explaining its elevated expression in aggressive tumors [81]. Other examples include *GJA1P1*, which may encode a gap junction protein that suppresses tumor cell proliferation [82]. Besides all, the coding products of pseudogenes can be detected in other systems, such as *PGK2*, despite exhibiting canonical features of retrotransposition-derived sequences, also maintain an uninterrupted open reading frame and demonstrate functional testis-specific expression [12, 13]. Beyond human tissues, protein products derived from pseudogenes have also been identified in other species. For instance, the murine pseudogene *Pfh8-ps*, which retains an intact coding sequence, gives rise to the PHF8-PS protein [76]. This protein has lost the histone demethylase activity characteristic of its parental protein but localizes to the cytoplasm. Through interactions with prefolding and other co-factors, it may participate in a previously unrecognized pathway involved in protein folding.

In brief, although protein-coding pseudogenes represent a relatively small proportion of the total pseudogene repertoire, emerging evidence demonstrates their capacity to participate in critical biological processes through diverse molecular mechanisms. These functional pseudogenes encoded products often exhibit exquisite specificity in their regulatory roles, frequently targeting pathways associated with their parental genes.

3.2.3 | Functional DNA Sequences of Pseudogenes

Beyond their well-characterized roles in RNA-based regulation and potential protein-coding activities, recent studies have revealed that pseudogenes can also function at the DNA level through mechanisms such as acting as cis-regulatory elements, participating in chromatin-mediated gene regulation, and facilitating non-allelic homologous recombination (gene conversion) (Figure 4c). While less extensively studied than their transcriptional activities, these DNA-level functions represent important and underappreciated aspects of pseudogene biology.

First, certain pseudogene sequences can function as cis-regulatory elements. Features inherited from their parental genes, such as GC-rich sequences and CTCF binding sites, can exhibit promoter, enhancer, or insulator activity, enabling them to participate in the regulation of neighboring gene expression [36, 47, 68, 83, 84]. Some pseudogenes may even contribute to mediating long-range chromatin interactions, exemplified by the *HBBP1* pseudogene, which facilitates the fetal-to-adult hemoglobin switch by physically interacting with the locus control region (LCR) through chromatin looping, thereby modulating globin gene expression during erythroid development [85]. This singular reported case suggests that pseudogenes may serve as architectural elements in genome organization, potentially influencing gene expression through long-range chromatin contacts, though the generalizability of this phenomenon requires further investigation. More extensive studies are the role of pseudogenes in disease pathogenesis through gene conversion events, where non-allelic homologous recombination between pseudogenes and their functional parental genes transfers deleterious mutations [86]. This mechanism underlies numerous disorders, including hereditary pancreatitis caused by *PRSS3P2-PRSS3* conversions, congenital adrenal hyperplasia resulting from *CYP21A2P-CYP21A2* recombination, and various forms of inherited cataracts (*CRYBP1-CRYBB2* conversions) [87–89]. These pathogenic gene conversion events typically occur due to the high sequence similarity between pseudogenes and their functional counterparts, with conversion tracts ranging from 20 bp to several kilobases.

The relative paucity of research on pseudogene chromatin interactions compared to gene conversion events highlights an important gap in our understanding of pseudogene biology, suggesting that future studies employing techniques like Hi-C, ChIA-PET, and CRISPR-based chromatin manipulation could uncover additional examples of pseudogenes functioning as regulatory DNA elements [90–92]. Together, these DNA-level functions expand the functional repertoire of pseudogenes beyond their well-characterized RNA-based activities and underscore their multifaceted roles in both normal biology and disease pathogenesis.

3.3 | Dual Identities in Physiological and Pathological States

3.3.1 | Disease-Associated Pseudogenes

As discussed, pseudogenes can exert regulatory functions at the DNA, RNA, and protein levels. Consequently, the broader biological functions and implications of these regulatory activities are receiving increasing attention. The dysregulated expression of pseudogenes in various diseases has emerged as a critical area of research, with numerous studies demonstrating their functional involvement in tumor initiation and progression through diverse molecular mechanisms [71, 93–97].

Among cancer-associated pseudogenes, *PTENP1* represents a paradigm-shifting discovery that fundamentally altered our

understanding of pseudogene biology, serving as the representative well-characterized example of a tumor-suppressive pseudogene functioning through microRNA decoy mechanisms. The *PTEN* tumor suppressor gene has at least 17 pseudogenes in the human genome, with *PTENP1* being an extensively studied retroposed pseudogene sharing high 3'UTR sequence homology with its parental gene. *PTENP1* exerts its tumor-suppressive function primarily by competitively binding a panel of oncogenic microRNAs, including miR-17, miR-19, miR-21, miR-26, miR-214, and so on, that normally target *PTEN* mRNA for degradation, thereby acting as a molecular sponge that protects *PTEN* transcripts from microRNA-mediated suppression [14]. Beyond this canonical ceRNA mechanism, *PTENP1* displays remarkable regulatory complexity through the production of two distinct antisense RNA isoforms (α and β) that differentially modulate *PTEN* expression: the α isoform mediates trans-regulation by recruiting DNA methyltransferases to the *PTEN* promoter to epigenetically modulate its transcription, while the β isoform forms RNA-RNA duplexes with *PTEN* mRNA to regulate its stability, nuclear-cytoplasmic transport, and microRNA accessibility [64]. The clinical relevance of *PTENP1* is underscored by its frequent downregulation in gastric cancer, squamous cell carcinoma, and other malignancies, where reduced *PTENP1* expression correlates with decreased *PTEN* levels and worse clinical outcomes [98–102]. Parallel to *PTENP1*, other pseudogenes like *KRASP1* and *BRAFP1* have been implicated in cancer pathogenesis through similar ceRNA mechanisms, with *KRASP1* sponging miR-143 and let-7 family to protect *KRAS* mRNA [14] and *BRAFP1* enhancing oncogenic signaling by sequestering microRNAs that target *BRAF* [11, 103]. The pervasive dysregulation of pseudogenes across cancer types—with some functioning as tumor suppressors (e.g., *PTENP1*) and others as oncogenic drivers (e.g., *KRASP1*, *BRAFP1*)—highlights their importance as a novel class of cancer genes. These findings have catalyzed a new era of pseudogene research in oncology, suggesting their potential as diagnostic biomarkers, therapeutic targets, and mediators of treatment resistance. Future studies integrating single-cell analyses, CRISPR screening, and clinical correlative studies will be essential to fully exploit the translational potential of cancer-associated pseudogenes.

Beyond transcriptional dysregulation, pseudogene copy number variations (CNVs) have emerged as significant contributors to human disease pathogenesis, with the *NOTCH2NL* locus serving as a paradigmatic example of how pseudogene structural variations can exert bidirectional effects on neurodevelopment [79, 104]. The *NOTCH2NL* gene family, as we mentioned above, derived from partial duplications of the *NOTCH2* receptor gene, exhibits dosage-sensitive effects on cortical development, where increased copy numbers are associated with autism spectrum disorder and macrocephaly, while deletions correlate with microcephaly and schizophrenia. This dual pathology stems from *NOTCH2NL*'s role in modulating NOTCH signaling dynamics-copy number gains amplify cortical progenitor expansion through enhanced *NOTCH* activation, whereas losses impair neurogenesis due to insufficient NOTCH pathway stimulation [79]. More broadly, the *NOTCH2NL* paradigm underscores the importance of systematically evaluating pseudogene CNVs across the genome, particularly for human-specific duplicates that may underlie both our unique cortical expansion and associated neuropsychiatric vulnerabilities.

The association between pseudogenes and disease is not limited to humans but also extends to susceptibility traits in other species. A notable example is found in canids, where numerous retrocopies (a class of pseudogenes) have been identified [26]. Among these, a retrocopy of fibroblast growth factor 4 (*FGF4*) shows a strong association with a chondrodysplasia phenotype (characterized by short limbs) and is a major genetic factor underlying both this morphology and a high prevalence of intervertebral disc disease (IVDD) in certain dog breeds [105, 106]. This case illustrates how pseudogenes can contribute to disease susceptibility across species.

3.3.2 | Physiological Roles of Pseudogenes

In addition to being associated with disease progression, current research has also established that pseudogenes play critical regulatory roles during normal development, with several well-characterized examples demonstrating their involvement in stem cell maintenance, development, cellular immunity, and so on.

Our previous research has uncovered a pseudogenes expression network during human early embryonic development, where we systematically identify 140 functionally active pseudogenes that exhibit stage-specific expression patterns across critical developmental windows, and these embryonic-expressed pseudogenes display critical regulation functions in both naive (pre-implantation) and primed (post-implantation) pluripotency states [107]. Meanwhile, in addition to the above-mentioned DNA sequences of *HBBP1* involved in the formation of chromatin interactions, we also found that the transcription product of *HBBP1* can participate in the process of erythropoiesis by competitively binding to hnRNPA1 with *TAL1* mRNA, suggesting the critical role of *HBBP1* in the normal development of the human blood system [16]. Pseudogenes may also participate in immune responses, as evidenced by the identification of 54 pseudogenes derived from lncRNAs within the TNFα-induced inflammatory signaling pathway in mice, which demonstrate precisely regulated expression patterns [108].

In summary, although research on pseudogenes in normal physiological processes remains relatively limited compared to their pathological roles, their crucial regulatory functions cannot be overlooked. The dual functionality of pseudogenes, operating in both pathological and physiological contexts, renders certain pseudogenes indispensable within the human genome. Moreover, in recent years, high-throughput sequencing technologies have revealed the phenomenon of "pervasive transcription" in mammalian genomes, yielding vast quantities of non-coding RNAs with unknown functions, often referred to as "dark matter RNA" [109, 110]. This discovery has sparked considerable debate within the scientific community regarding whether these widespread transcriptional outputs represent regulatory molecules with specific biological roles or merely reflect non-functional "transcriptional noise". The transcripts derived from pseudogenes, which are the focus of this review, lie at the heart of this controversy. Although pseudogenes originate from inactivated gene copies and are often regarded as evolutionary by-products, accumulating evidence demonstrates that some pseudogene transcripts play crucial regulatory roles through mechanisms such as ceRNA activity. Therefore, elucidating the

functions of pseudogenes not only advances our understanding of pseudogenes themselves but also provides important insights into the biological significance of pervasive transcription.

4 | Roles of Pseudogenes in Evolution

As duplicated sequences are highly similar in sequence to their parental genes, pseudogenes can be functionally reactivated under specific conditions to exert regulatory roles at multiple molecular levels. While parental genes represent the predominant targets of pseudogenes mediated regulation, emerging evidence reveals their involvement in broader signaling networks as well. Due to their diverse mechanisms of action and wide range of molecular targets, pseudogenes demonstrate the capacity to perform critical biological functions in both physiological and pathological contexts. Gene family expansion through sequence duplication represents a well-established mechanism for functional innovation [111]. Similarly, as products of sequence duplication, the high degree of species specificity observed in human pseudogene sequences suggests that human- or primate-specific pseudogenes in particular may play pivotal roles in species-specific regulatory processes. Growing research supports this evolutionary perspective, highlighting the regulatory potential of pseudogenes in shaping species-specific adaptations. Current findings position pseudogenes as important genomic elements that contribute to innovation through their unique regulatory capabilities.

4.1 | Functional Pseudogenes During Evolution

4.1.1 | Functionalization Caused by Activated Pseudogenes

While pseudogenes are commonly thought to be retained in the genome due to their selective neutrality, growing evidence reveals that these "genomic fossils" can also acquire critical biological functions during species evolution. This occurs primarily through two evolutionary paths: the resurrection of the original function via compensatory mutations or the neofunctionalization through gain-of-function mutations (Figure 5a).

The pseudogenes we will discuss, *NOTCH2NL*, *HBBP1*, and *IRF5P1*, exemplify species-specific regulatory innovations, each revealing distinct paths to functional evolution. The human-specific *NOTCH2NL* duplicates, originating from partial duplication of the *NOTCH2* gene, exhibit dosage-sensitive regulation of cortical neurogenesis by modulating *NOTCH* signaling dynamics, a mechanism fundamentally linked to human brain expansion [79, 17]. Notably, *NOTCH2NL* copy number variations demonstrate pleiotropic effects on human neurodevelopmental disorders. Similarly, the hemoglobin cluster pseudogene *HBBP1* has evolved human-specific functionality in erythropoiesis through a unique RNA-protein interaction network wherein its transcript stabilizes *TAL1* mRNA by competitively binding hnRNPA1, thereby creating a human-specific regulatory circuit. Intriguingly, in some species, the *HBBP1* locus even retains protein-coding capacity, suggesting that its regulatory function was remodeled during human-specific adaptation [16]. Furthermore, the higher primate-specific pseudogene *IRF5P1* is a chimeric element composed of sequences derived from multiple origins, likely originating approximately 60 million years ago through an ancient retroviral integration event [42]. Located within the *TRIQK* gene, it is transcribed into antisense RNA, which may bind to *IRF5* mRNA and post-transcriptionally regulate its expression. This mechanism suggests a significant evolutionary role for pseudogenes in shaping immune system adaptations in primates.

These examples illustrate the diverse mechanistic paths by which pseudogenes acquire new functions. As introduced in Section 3, under certain conditions, pseudogenes can be reactivated to produce functional RNAs or even encode proteins. The transcriptional or translational products of pseudogenes may either exert functions analogous to those of their parental genes, such as *NOTCH2NL* and *IRF5P1*, or acquire novel biological roles, as exemplified by *HBBP1*. Regardless of the mechanism of their functional "resurrection", these revitalized pseudogenes may contribute to adaptive evolutionary processes, thereby becoming selectively retained in the genome. In sum, pseudogenes can acquire lineage-restricted regulatory roles through structural innovation, develop entirely new interaction networks, and become genetically embedded in species-defining biological processes. The biomedical significance is also profound, as these species-specific pseudogene functions simultaneously represent both evolutionary adaptations and disease association. Future research should explore whether there are more pseudogenes contributing to human-specific or primate-specific traits, whether such species-specific pseudogene networks follow predictable evolutionary patterns or emerge stochastically, and how their manipulation might address human-specific disorders.

4.1.2 | Evolutionary Advantages Resulted From Pseudogenization

Pseudogenes can still provide an evolutionary advantage to an organism even if it has lost its original coding function (Figure 5b). A prime example is *Xist*, essential for X-chromosome dosage compensation in placental mammals, which evolved through the pseudogenization of the ancestral *Lnx3* protein-coding gene [112]. Intriguingly, its transition to a noncoding regulator appears to have preceded the acquisition of X-inactivation functions, exemplifying how pseudogenes can serve as genomic substrates for the emergence of new regulatory elements. Furthermore, the *MYH16* pseudogenization event represents a pivotal transition in human evolution, as reported by Stedman et al. [113]. The inactivation of this masticatory muscle myosin gene triggered a cascade of morphological changes, such as the decreased mechanical constraints on cranial growth resulting from reduced temporal muscle mass, thus enabling brain expansion and facilitating cognitive advancements characteristic of the Homo genus. Moreover, the pseudogenization of *ZP3R* in primates, a key gene mediating sperm-egg adhesion, is likely linked to a relaxation of sexual selection pressure [114, 115]. During the evolution of humans and higher primates, mutations in the *GULOP* gene led to the loss of the ability to synthesize vitamin C endogenously [116–118]. Similarly, the pseudogenization of the urate oxidase (*UOX*) gene likely contributed to altered purine metabolism [119, 120]. These functional losses may represent metabolic adaptations that confer evolutionary advantages by conserving energy and optimizing

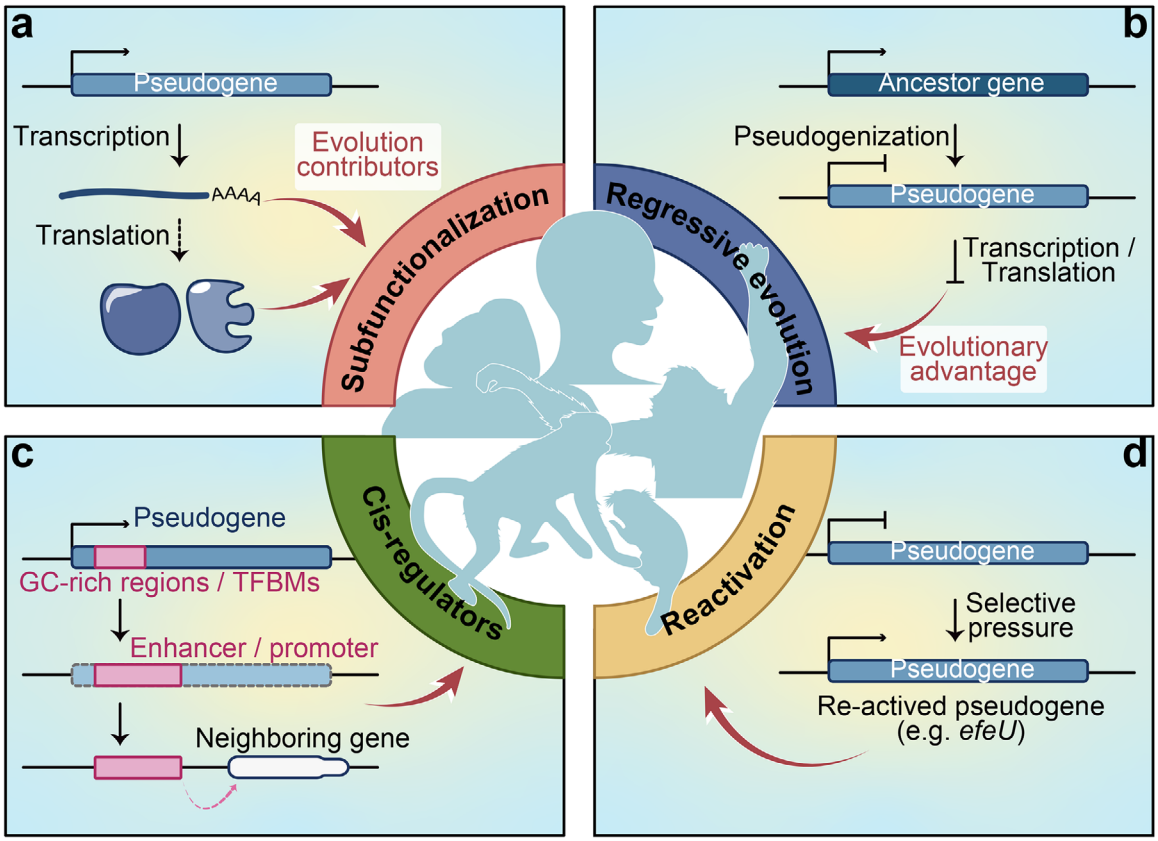

FIGURE 5 | Pseudogene contributions to species evolution. (a) Novel functions from transcribed or translated pseudogenes. (b) Adaptive advantages via functional gene pseudogenization. (c) Evolution into cis-regulatory elements such as enhancers or promoters, via GC-rich sequences or transcription factor binding motifs (TFBMs) on pseudogenes. (d) Reactivation of pseudogenes under certain selective pressures as a genomic reservoir.

resource allocation. Furthermore, in felids, the pseudogenization of the sweet taste receptor gene *Tas1r2* is likely an adaptive trait associated with their obligate carnivorous diet [121, 122].

Beyond mammals, classic examples of "loss-of-function" mutations and pseudogenization driving adaptive evolution are also prevalent in other metazoans. For instance, the ZRS (ZPA regulatory sequence), a critical enhancer for limb development in tetrapods, has undergone severe degeneration or even complete loss in the snake genome [123, 124]. This loss led to the termination of limb development programs in snakes, which may have facilitated adaptation for burrowing or slithering locomotion. Similarly, in fish species inhabiting dark cave environments, such as Astyanax mexicanus, visual degeneration is commonly observed [125, 126]. This trait is strongly associated with inactivating mutations in multiple eye development genes. The loss of eyesight provides significant adaptive evolutionary advantages for survival in nutrient-poor, lightless environments.

Throughout evolution, certain genes have lost their protein-coding capacity, while others have become transcriptionally silent. Paradoxically, such losses may themselves represent evolutionary gains. It is possible that sporadic gene inactivation events conferred adaptive advantages, such as increased brain size or more efficient metabolic pathways, thereby promoting their retention in the genome. These "regressive evolution" pathways, where loss-of-function mutations drive adaptive complexity, reveal pseudogenes as dual-natured genomic elements that, while molecularly "broken" relative to ancestors, pseudogenes may still gain an evolutionary advantage for innovation. The above paradigms underscore how pseudogenization contributes to species-defining adaptations, and it's possible that having more pseudogenes that lose their function is instead an evolutionarily better choice.

4.2 | Evolutionary Potential of Pseudogenes as Cis-Regulators

Regardless of whether pseudogenes can resurgence, their reintegration itself represents an expansion of the genomic repertoire [127–129]. Pseudogene integration increases sequence complexity within the genome, in which certain pseudogenes may serve as reservoirs of potential cis-regulatory elements (Figure 5c) [36, 47, 68, 83, 84]. For instance, retroposed pseudogenes derived from protein-coding genes may carry GC-rich regions and transcription factor binding motifs (TFBMs), providing raw material for the evolution of promoters or enhancers. The presence of TFBMs in retroposed pseudogenes enhances their potential to evolve into enhancer-like regions capable of modulating neighboring gene expression, and these motifs may degenerate over evolutionary time in the absence of selective pressure. Moreover, some pseudogenes can inherit evolutionarily competent promoter activity from their parental genes, as exemplified by the interaction between *PPP1R26P1* and the *RB1* locus [130]. The retrotransposition event generating *PPP1R26P1* occurred prior to

the divergence of New and Old World monkeys, resulting in its integration within intron 2 of *RB1*. The *PPP1R26P1* sequence derived from exon 4 of its parental gene contains a differentially methylated CpG island that evolved promoter activity, which can serve as an alternative promoter contributing to generating novel splice isoforms for the host gene. In addition to evolving promoter activity to facilitate gene expression, pseudogenes may also function as cis-regulatory elements in gene silencing processes. For instance, small RNAs derived from pseudogenes can recruit SETDB1 (a histone H3K9 methyltransferase), leading to trimethylation of histone H3 at lysine 9 (H3K9me3), which induces localized chromatin repression and transcriptional silencing [68].

Functioning as cis-regulatory elements may be a key reason why some pseudogenes, despite losing protein-coding capacity, have not been eliminated during evolution. Studies have shown that for genes with lost coding potential (GLCPs), although their ability to produce proteins is abolished, enhancer elements embedded within their genomic regions can retain critical biological functions [131]. Natural selection acts upon these enhancer sequences to maintain their regulatory roles, thereby preserving the entire locus. We propose that this evolutionary mechanism also applies to pseudogenes. As a class of GLCPs, pseudogenes may retain or even evolve regulatory elements such as promoters, enhancers, or insulators. The regulatory signals in which they participate may confer adaptive advantages during evolution, facilitating their retention in the genome.

4.3 | Reactivation of Pseudogenes Under Selective Pressure

The above findings demonstrate that pseudogenes can serve distinct yet equally important evolutionary functions by directly participating in species adaptation as key evolutionary drivers. Some pseudogenes, even though they may not be currently indispensable for the survival of organisms, may provide an evolutionary basis for species innovation (Figure 5c). The phenomena we currently observe represent products of billions of years of evolution, yet we must recognize that evolution remains an ongoing process, so the current lack of detectable function does not preclude a sequence's potential to acquire regulatory roles under selective pressures.

A notable example in primate evolution involves the loss and functional recovery of the *IRGM* gene, a key member of the Immunity-Related GTPase (IRG) family that defends against intracellular pathogens like Mycobacterium tuberculosis and Toxoplasma [132, 133]. After anthropoids diverged from prosimians around 50 million years ago, most IRG clusters were deleted, though *IRGM* was retained. It was later inactivated when an Alu retrotransposon disrupted its open reading frame, turning it into a pseudogene. However, a critical mutation in the common ancestor of humans and great apes restored the coding sequence by converting a stop codon into an amino acid codon, allowing functional protein production to resume. This cycle of loss and revival illustrates how evolutionary pressures, such as host-pathogen conflicts, can drive functional innovation even after gene inactivation, highlighting the dynamic nature of genomic evolution in response to changing selective demands.

However, the regulatory systems governing gene expression in eukaryotes have evolved to be highly complex, making it experimentally challenging to verify whether certain pseudogenes could be reactivated under specific selective pressures to resume transcription or even protein production. In contrast, the comparatively simpler gene regulatory architecture of prokaryotes may provide a critical model system for investigating the molecular mechanisms and evolutionary potential of pseudogene resurrection. A compelling example emerges from bacterial systems, that under iron-limiting conditions, *Escherichia coli* (*E. coli*) can reactivate the pseudogene *efeU* through adaptive mutations that restore its promoter activity and repair frameshift errors, reconstituting an intact open reading frame [134]. Then the resuscitated *efeU* protein forms a positive feedback loop with the entCEBA operon to enhance iron uptake efficiency. This case fundamentally demonstrates how specific environmental pressures can drive functional pseudogene resurrection through targeted genomic editing. Although the E. coli genome is less complex than the human genome, it is worth exploring whether pseudogenes in the human genome can also serve as an evolutionary "pre-adaptive" reservoir.

In summary, we propose several principal mechanisms through which pseudogenes may contribute to species evolution. A well-documented pathway involves the reactivation of pseudogenes, which can restore, either partially or fully, the function of their parental genes or even acquire independent regulatory roles. The transcriptional or translational products of such reactivated pseudogenes may participate in gene regulatory networks, thereby conferring evolutionary advantages. Conversely, the pseudogenization of functional genes leads to the loss of their protein-coding or transcriptional capacity. This loss may itself represent an adaptive outcome, wherein the elimination of non-essential genetic elements optimizes energy allocation and refines survival strategies, ultimately promoting the retention of such pseudogenes in the genome. Furthermore, we highlight additional evolutionary potentials of pseudogenes, including their roles as cis-regulatory elements modulating the expression of other genes, as well as their capacity for functional resurrection under specific selective pressures. These observations underscore the dynamic and plastic nature of genomic evolution, illustrating how pseudogenes serve as versatile substrates for evolutionary innovation.

5 | Technological Advances Driving the Mining of Functional Pseudogenes

As previously discussed, pseudogenes have been found to exert regulatory functions through diverse mechanisms, and their relationship with evolutionary processes is increasingly being explored. Technological advances have been indispensable in uncovering pseudogene functionality, such as the development of more accurate algorithms and the application of long-read sequencing in pseudogene annotation. Furthermore, the rapid advancement and biological application of deep learning, particularly in genomics, transcriptomics, and evolutionary studies, are poised to revolutionize pseudogene research by improving detection, functional classification, and evolutionary annotation across species [135–139]. Here, we focus on and discuss several transformative technological developments that have recently

contributed most significantly to the field, including the use of long-read sequencing for accurate quantification of pseudogenes, the integration of single-cell omics in pseudogene studies, the potential of large-scale comparative genomics, and the facilitation of functional validation through gene editing technologies.

5.1 | Application of Long-Read Sequencing Technologies in Pseudogene Quantification

Accurately quantificating pseudogenes derived transcripts is crucial for elucidating their biological functions. Although parallel RNA sequencing (RNA-seq) can identify lncRNA transcripts, the short-read fragments obtained may be insufficient to distinguish sequence differences between pseudogenes and their parental genes [140–142]. Specialized pipelines have been developed to focus only on reads from divergent regions of pseudogenes [143], but for evolutionarily younger pseudogenes, RNA-seq-captured sequences may not yet exhibit sufficient divergence from their parental genes. A notable example involves *GBA1*, a gene whose mutations cause Gaucher disease and represent the most significant genetic risk factor for Parkinson's disease. Its highly homologous pseudogene, *GBAP1*, poses major mapping challenges: over 50% of short RNA-seq reads cannot be uniquely aligned to either *GBA1* or *GBAP1*, confounding accurate expression quantification and significantly complicating transcriptomic studies at this locus [144]. The emergence of long-read sequencing technologies fundamentally addresses this limitation [145–149]. By spanning multiple divergent sites within a single read, long-read platforms enable unambiguous assignment and precise quantification of transcripts originating from *GBA1* versus *GBAP1*.

At a broader scale, PacBio long-read cDNA sequencing applied to normal human tissues and cancer cell lines has led to the discovery of hundreds of novel transcribed pseudogenes, some of which even retain protein-coding potential [50, 150, 151]. Integrating long-read sequencing technologies to identify dynamically expressed pseudogenes throughout mammalian development, and to further investigate their functional and evolutionary characteristics, represents a major application of these advances in pseudogene research. Thus, despite possessing the most complete eukaryotic transcriptome annotation available, human reference genomes continue to reveal numerous previously unannotated genes through long-read sequencing, underscoring both the incompleteness of current annotations and the transformative potential of emerging technologies.

In sum, long-read sequencing technologies (e.g., PacBio HiFi and Oxford Nanopore) overcome key limitations of short-read approaches in pseudogene research by providing complete, unambiguous sequence contexts. They enable accurate detection of pseudogene-specific features and resolve complex genomic regions, including segmental duplications and structural variations. These advances support precise classification of pseudogenes and reveal evolutionary dynamics in repetitive gene families. Additionally, native epigenetic profiling (e.g., 5mC detection via Nanopore) links structural variation to regulatory mechanisms, offering multidimensional insights into pseudogene silencing and functional evolution. Long-read sequencing thus represents a transformative tool for decoding pseudogene biology and genome evolution. Certainly, it is undeniable that long-read sequencing currently suffers from limited sequencing depth, which may hinder the detection of pseudogenes with low expression abundance. Therefore, when performing quantitative analysis of pseudogenes in research, a combined approach utilizing both long-read and short-read sequencing can be adopted. This strategy involves applying pseudogene-specific discrimination methods during short-read data analysis, while simultaneously leveraging long-read sequencing to reconfirm pseudogene fragments that exhibit high sequence similarity to their parental genes.

5.2 | Single-Cell Technologies Revealing Pseudogene Expression and Function

Long-read sequencing enhances the accuracy of gene detection and quantification, while single-cell sequencing technologies enable the resolution of cell type-specific expression patterns across diverse tissues, developmental stages, and disease conditions at unprecedented resolution, effectively overcoming the limitations of cellular heterogeneity inherent in conventional bulk sequencing approaches. These advances have also significantly propelled research in the field of pseudogenes.

For instance, through the application of single-cell sequencing, the pseudogene *mOct4pg9* was identified as being specifically highly expressed in mouse neural stem/progenitor cells [152]. Its transcribed non-coding RNA was shown to play a critical role in impairing neural stem cell (NSC) function and disrupting brain development following fetal ethanol exposure. Another example involves *CLEC9A*, a gene encoding a key receptor protein essential for cross-presentation in mammalian immunity. In Sus scrofa, this gene has been pseudogenized due to mutation. Single-cell sequencing revealed specific expression of this pseudogene primarily in dendritic cells, suggesting a potential association with dendritic cell maturation and functional status [153]. Furthermore, the expanding repertoire of surface markers in single-cell studies now includes certain pseudogenes, indicating that not only can their expression be detected via single-cell technologies, but some pseudogenes derived transcripts may also serve as markers for specific cell types [154].

In summary, single-cell sequencing technologies enable the identification of pseudogenes with cell-specific expression patterns and facilitate the analysis of their cellular and spatial expression heterogeneity. These capabilities provide critical insights into the functional roles and regulatory mechanisms of pseudogenes, thereby advancing our understanding of their contributions to development, homeostasis, and disease.

5.3 | Charting the Pseudogene Landscape via Large-Scale Comparative Genomics Studies

Fueled by advances in the depth and breadth of sequencing technologies, large-scale comparative genomics studies are flourishing. These efforts, leveraged by consortium projects like ENCODE, GTEx, TCGA, and gnomAD, provide public data

resources that enable the systematic, system-level investigation of pseudogene biology.

For example, epigenomic analyses from ENCODE have revealed that certain pseudogene loci exhibit activating histone marks, such as H3K4me3 and H3K27ac, indicative of active enhancer or promoter activity [109]. Their chromatin accessibility profiles and transcription factor binding patterns further suggest that some pseudogenes may function as independent regulatory elements. Leveraging multi-tissue RNA-seq data from the GTEx project, it is now feasible to construct a comprehensive expression atlas of pseudogenes, identify those with tissue-specific expression, and infer biological functions through co-expression network analyses that reveal coordinated activities between pseudogenes and protein-coding genes [155]. Cancer genomics initiatives like TCGA have further propelled pseudogene research into the clinical arena [156]. Pan-cancer analyses comparing pseudogene expression in tumors versus normal tissues have identified cancer-specific pseudogenes, many of which show correlations with cancer subtypes, tumor stage, and patient prognosis. Population genomics resources such as gnomAD enable the assessment of evolutionary constraints on pseudogene regions [157]. Although pseudogenes generally experience weaker purifying selection than coding genes, some exhibit significant mutational constraint across human populations, suggesting potential functional importance.

In summary, large-scale comparative studies leveraging these rich datasets are making substantial contributions to uncovering the functional landscape of pseudogenes and elucidating the principles governing their biological and clinical relevance.

5.4 | Functional Validation of Pseudogenes via Gene Editing

Advances in omics technologies are providing an unprecedented, comprehensive view of pseudogenes, and the application of gene editing tools is enabling their precise manipulation.

When performing genetic manipulations to investigate pseudogene function, due to sequence similarity between pseudogenes and their parental genes, precisely manipulating pseudogene expression without off-target effects on parental genes remains challenging. Traditional RNA interference (e.g., small interfering RNA) employs short fragments that may not reliably target pseudogenes specific divergent regions, leading to high off-target risks. The emergence of CRISPR technology provides a solution for DNA-level pseudogene targeting, enabling knockout via pseudogenes specific flanking sequences or expression modulation (e.g., CRISPR interference, CRISPR activation) by targeting promoter regions [158–161]. Employing CRISPR-based knockout or knockdown strategies, such as targeting *NANOGP8* for deletion or *PTENP1* for transcriptional repression, enables functional investigation of pseudogenes while minimizing off-target effects on their parental genes due to improved guide RNA specificity [162, 163]. Furthermore, large-scale CRISPR screens have been increasingly applied to systematically identify functional pseudogenes across the genome. A recent study has successfully used CRISPR libraries to knock down a subset of pseudogenes, demonstrating the feasibility of CRISPR-based pseudogene interference [164].

When applying CRISPR technology, it is essential to design gRNAs specifically targeting sequence divergence sites between the pseudogene and its parental gene, so as to minimize off-target effects and unintended disruption of the functional parental gene. Suitable target sites may include exon-intron junctions, which are absent in the parental gene due to the intron-less nature of retroposed pseudogenes or external sequences such as upstream regulatory regions unique to the pseudogene locus. However, for pseudogenes integrated into host genes or sharing cis-regulatory elements with nearby genes, DNA-level manipulations may disrupt the integrity or expression of neighboring genes. Thus, when performing CRISPR, it is essential to exclude interference with adjacent gene functions. For CRISPR library screens targeting functional pseudogenes, optimal candidates may be either host-free, integrate into intronic regions, or sufficiently distant from neighboring genes to minimize unintended effects. Another important experimental consideration is that when using Cas9 to target the parental gene, the high sequence similarity between pseudogenes and their parental counterparts may lead to gene conversion events during DNA repair, where the broken parental allele is repaired using the pseudogene sequence as a template [165]. Conversely, a similar mechanism may occur during pseudogene knockout, where the parental gene sequence is used for repair, potentially restoring functionality. To enhance the efficacy of pseudogene knockout, one may introduce an exogenously designed homology-directed repair (HDR) template to introduce out-of-frame insertions or deletions. This strategy can induce premature termination codons and trigger nonsense-mediated mRNA decay (NMD), ultimately achieving stable knockout.

In summary, although several technical challenges warrant careful attention in CRISPR-based editing of pseudogenes, the technology remains highly advantageous due to its precision and ability to minimize confounding effects from homologous parental genes.

6 | Conclusions and Perspective

An increasing number of studies have revealed the regulatory functions that pseudogenes may exert at various levels. As products of species evolution, the functional roles of pseudogenes are also likely to participate in evolutionary regulation. The growing understanding of pseudogene functions has led us to question whether the traditional binary classification of functional genes is still suitable for distinguishing so-called pseudogenes from true genes. For example, the functional pseudogene *NOTCH2NL*, as mentioned above, despite its typical sequence truncation and pseudogenic origin, has been redefined as a protein-coding gene due to its known function. Where, then, lies the boundary between pseudogenes and true genes? Even if they currently lack function, pseudogenes may be reactivated under special conditions—so does that mean they can then be called protein-coding genes? We argue that the functional manifestation of pseudogenes is context-dependent, making them more representative of a potential state rather than a simple gene-type classification. Because labeling as pseudogenes, they are often excluded from genomic analyses and functional screenings,

leading to the potential oversight of many functional individuals. As proposed by Seth W. Cheetham et al. [18], retroposed pseudogenes, which are generated through retrotransposition, can be collectively termed "retrocopy", while duplicated pseudogenes, arising from DNA segment duplication, may be referred to as "gene copy" or "paralogue". Current gene-type classifications may no longer serve as criteria for assessing functional potential; instead, auxiliary genotyping based on the mechanism of origin may provide more insight into their sequence composition, relative conservation, and even functional modes [5].

The functional roles of pseudogenes and their potential resurrection under selective pressures have reshaped our understanding of the mechanisms driving their persistence and functionalization. The evolutionary trajectories of pseudogenes are influenced by a dynamic interplay between neutral evolution and positive selection, which may operate on different timescales. Initially, pseudogenes accumulate mutations neutrally in the absence of functional constraints. However, under shifting evolutionary pressures, such as pathogen invasion or metabolic changes, a once-neutral pseudogene may be co-opted for a novel function, leading to the onset of positive selection that refines its sequence for this new role. This transition from a neutral "genomic fossil" to a functional element underscores the adaptive potential of pseudogenes. Beyond sequence-level evolution, structural variation (SV), including segmental duplications and copy number variations (CNVs), significantly accelerates pseudogene birth and death rates, which exerts a profound and dramatic impact on the abundance of pseudogenes [166]. SV generates redundant gene copies that serve as substrates for evolutionary innovation, while deletions within CNVs can eliminate pseudogenes containing segments, increasing their genomic turnover. Divergence in SV patterns among species partly explains the observed variation in pseudogene repertoires across taxa. The computational simulation framework established in previous work also provides a highly flexible and powerful tool for investigating the origin and evolution of gene families, reflecting the processes of gene "birth" and "death", thereby enabling more nuanced studies of pseudogene sequence evolution [167]. Taken together, pseudogenes serve as crucial sources of genomic plasticity. They provide a reservoir of low-constraint sequences that facilitate rapid functional exploration, complementing the more rigid optimization of protein-coding genes and offering dynamic modules for evolutionary innovation.

As previously introduced, existing studies have revealed that individual pseudogenes may perform important regulatory functions at various molecular levels and under different physiological states. Species-specific pseudogene functions may even contribute to evolutionary processes. Therefore, investigating whether more pseudogenes participate in biological activities and potentially contribute to species evolution, in other words, systematically studying and elucidating the functional roles of the pseudogene population, is of critical importance. For example, profiling the expression patterns of pseudogenes across different developmental stages and physiological states to identify differentially expressed pseudogenes; or comparing orthologous developmental processes (e.g., neurodevelopment, embryogenesis) across species to explore whether pseudogenes may be associated with species-specific phenotypic differences. Beyond functional exploration, the molecular mechanisms underlying pseudogene regulation remain largely unclear. While some functions at the DNA, RNA, and protein levels are known, the vast number of pseudogenes still lack well-defined mechanisms. Could they function at other regulatory levels? These are compelling questions for further investigation. Meanwhile, it should be noted that this review primarily focuses on pseudogenes derived from protein-coding genes, whereas the study of pseudogenes originating from non-coding RNA genes (e.g., miRNA-derived pseudogenes, lncRNA-derived pseudogenes) presents more distinct and considerable challenges. Unlike protein-coding pseudogenes, which can often be readily identified as loss-of-function through frameshift mutations or premature stop codons, confirming the loss of function in a non-coding RNA pseudogene typically requires extensive experimental validation. This often includes functional characterization of homologous loci in other species to demonstrate the loss of specific regulatory activities, among other approaches. The current lack of such data in the literature has resulted in the underrepresentation of non-coding pseudogenes in evolutionary studies. The identification of pseudogenes derived from different types of parent genes remains a significant challenge in the field.

Thus, pseudogenes are, in fact, crucial sequence resources for species evolution. Although current research still faces challenges, we believe that, with advancing technologies and deepening understanding, the important roles of pseudogenes will be further unveiled.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFA0805703), the National Natural Science Foundation of China (82330007, 92268205, and 32470847), CAMS Innovation Fund for Medical Sciences (2021-I2M-1-019 and 2024-I2M-ZH-021).

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. P. Amaral, S. Carbonell-Sala, F. M. De La Vega, et al., "The Status of the Human Gene Catalogue," *Nature* 622 (2023): 41–47.

2. Q. Li, Z. Li, C. Feng, S. Jiang, Z. Zhang, and L. Ma, "Multi-Omics Annotation of Human Long Non-Coding RNAs," *Biochemical Society Transactions* 48 (2020): 1545–1556.

3. B. Pei, C. Sisu, A. Frankish, et al., "The GENCODE Pseudogene Resource Genome Biology," *Genome Biology* 13 (2012): R51.

4. C. Jacq, J. R. Miller, and G. G. Brownlee, "A Pseudogene Structure in 5S DNA of Xenopus Laevis," *Cell* 12 (1977): 109–120.

5. E. F. Vanin, "Processed Pseudogenes: Characteristics and Evolution," *Annual Review of Genetics* 19 (1985): 253.

6. H. Ali, F. Al-Mulla, N. Hussain, et al., "PKD1 Duplicated Regions Limit Clinical Utility of Whole Exome Sequencing for Genetic Diagnosis of Autosomal Dominant Polycystic Kidney Disease," *Sci Rep* 9, (2019): 4141.

7. M. D. Gallagher and A. S. Chen-Plotkin, "The Post-GWAS Era: From Association to Function," *The American Journal of Human Genetics* 102 (2018): 717–730.

8. M. Ghandi, F. W. Huang, J. Jane-Valbuena, et al., "Next-Generation Characterization of the Cancer Cell Line Encyclopedia," *Nature* 569 (2019): 503–508.

9. T. Hart, M. Chandrashekhar, M. Aregger, et al., "High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities," *Cell* 163 (2015): 1515–1526.

10. A. Tsherniak, F. Vazquez, P. G. Montgomery, et al., "Defining a Cancer Dependency Map," *Cell* 170 (2017): 564–576.

11. F. A. Karreth, M. Reschke, A. Ruocco, et al., "The BRAF Pseudogene Functions as a Competitive Endogenous RNA and Induces Lymphoma in Vivo," *Cell* 161 (2015): 319–332.

12. J. R. McCarrey, "Nucleotide Sequence of the Promoter Region of a Tissue-Specific Human Retroposon: Comparison With its Housekeeping Progenitor," *Gene* 61 (1987): 291–298.

13. J. R. McCarrey and K. Thomas, "Human Testis-Specific PGK Gene Lacks Introns and Possesses Characteristics of a Processed Gene," *Nature* 326 (1987): 501–505.

14. L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi, "A Coding-Independent Function of Gene and Pseudogene mRNAs Regulates Tumour Biology," *Nature* 465 (2010): 1033–1038.

15. C. Charrier, K. Joshi, J. Coutinho-Budd, et al., "Inhibition of SRGAP2 Function by its Human-Specific Paralogs Induces Neoteny During Spine Maturation," *Cell* 149 (2012): 923–935.

16. Y. Ma, S. Liu, J. Gao, et al., "Genome-Wide Analysis of Pseudogenes Reveals HBBP1's Human-Specific Essentiality in Erythropoiesis and Implication in β-Thalassemia," *Developmental Cell* 56 (2021): 478–493.

17. I. K. Suzuki, D. Gacquer, R. Van Heurck, et al., "Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis Through Delta/Notch Regulation," *Cell* 173 (2018): 1370–1384.

18. S. W. Cheetham, G. J. Faulkner, and M. E. Dinger, "Overcoming Challenges and Dogmas to Understand the Functions of Pseudogenes," *Nature Reviews Genetics* 21 (2020): 191–201.

19. Y. H. Lin, C. T. Yeh, C. Y. Chen, and K. H. Lin, "Pseudogene: Relevant or Irrelevant?," *Biomedical Journal* 48 (2025): 100790.

20. R. L. Troskie, G. J. Faulkner, and S. W. Cheetham, "Processed Pseudogenes: a Substrate for Evolutionary Innovation," *BioEssays* 43 (2021): e2100186.

21. Z. Zhang and M. Gerstein, "Large-Scale Analysis of Pseudogenes in the Human Genome," *Current Opinion in Genetics & Development* 14 (2004): 328–335.

22. D. Torrents, M. Suyama, E. Zdobnov, and P. Bork, "A Genome-Wide Survey of Human Pseudogenes," *Genome Research* 13 (2003): 2559–2567.

23. Z. Zhang, N. Carriero, D. Zheng, J. Karro, P. M. Harrison, and M. Gerstein, "PseudoPipe: an Automated Pseudogene Identification Pipeline," *Bioinformatics* 22 (2006): 1437–1439.

24. S. M. Searle, J. Gilbert, V. Iyer, and M. Clamp, "The Otter Annotation System," *Genome Research* 14 (2004): 963–970.

25. L. D. Hurst, "The Ka/Ks Ratio: Diagnosing the Form of Sequence Evolution," *Trends in Genetics* 18 (2002): 486–487.

26. K. Batcher, S. Varney, D. York, et al., "Recent, Full-Length Gene Retrocopies are Common in Canids," *Genome Research* 32 (2022): 1602–1611.

27. X. Feng and H. Li, "Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies," *Molecular Biology and Evolution* 38 (2021): 2958–2966.

28. C. Sisu, "GENCODE Pseudogenes," *Methods Mol Biol* 2324 (2021): 67-82, https://doi.org/10.1007/978-1-0716-1503-4_5.

29. J. E. Karro, Y. Yan, D. Zheng, et al., "Pseudogene.Org: a Comprehensive Database and Comparison Platform for Pseudogene Annotation," *Nucleic Acids Res* 35 (2007): D55.

30. Y. Shao, C. Chen, H. Shen, et al., "GenTree, an Integrated Resource for Analyzing the Evolution and Function of Primate-Specific Coding Genes," *Genome Research* 29 (2019): 682–696.

31. S. Balasubramanian, D. Zheng, Y.-J. Liu, et al., "Comparative Analysis of Processed Ribosomal Protein Pseudogenes in Four Mammalian Genomes," *Genome Biology* 10 (2009): R2.

32. Y.-J. Liu, D. Zheng, S. Balasubramanian, et al., "Comprehensive Analysis of the Pseudogenes of Glycolytic Enzymes in Vertebrates: The Anomalously High Number of GAPDH Pseudogenes Highlights a Recent Burst of Retrotrans-Positional Activity," *BMC Genomics* 10 (2009): 480.

33. A. C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann, "Emergence of Young Human Genes After a Burst of Retroposition in Primates," *PLoS Biology* 3 (2005): e357.

34. C. Esnault, J. Maestre, and T. Heidmann, "Human LINE Retrotransposons Generate Processed Pseudogenes," *Nature Genetics* 24 (2000): 363–367.

35. R. Baertsch, M. Diekhans, W. J. Kent, D. Haussler, and J. Brosius, "Retrocopy Contributions to the Evolution of the Human Genome," *BMC Genomics* 9 (2008): 466.

36. N. Vinckenbosch, I. Dupanloup, and H. Kaessmann, "Evolutionary Fate of Retroposed Gene Copies in the Human Genome," *Proceedings of the National Academy of Sciences* 103 (2006): 3220–3225.

37. R. Jorquera, R. Ortiz, F. Ossandon, et al., "SinEx DB: a Database for Single Exon Coding Sequences in Mammalian Genomes," *Database (Oxford)* 2016 (2016): baw095.

38. Z. D. Zhang, A. Frankish, T. Hunt, J. Harrow, and M. Gerstein, "Identification and Analysis of Unitary Pseudogenes: Historic and Contemporary Gene Losses in Humans and Other Primates, Genome Biology," *Genome Biology* 11 (2010): R26.

39. A. Frankish, M. Diekhans, A. M. Ferreira, et al., "GENCODE Reference Annotation for the Human and Mouse Genomes," *Nucleic Acids Res* 47 (2019): D766.

40. R. Dong, X. O. Zhang, Y. Zhang, X. K. Ma, L. L. Chen, and L. Yang, "CircRNA-Derived Pseudogenes," *Cell Research* 26 (2016): 747–750.

41. J. Liu, T. Liu, X. Wang, and A. He, "Circles Reshaping the RNA World: from Waste to Treasure," *Molecular Cancer* 16 (2017): 58.

42. A. Marquis, V. Hubing, C. Ziemann, E. N. Moriyama, and L. Zhang, "The Primate-Specific Presence of Interferon Regulatory Factor-5 Pseudogene 1," *Journal of Medical Virology* 96 (2024): e29879.

43. F. N. Carelli, T. Hayakawa, Y. Go, H. Imai, M. Warnefors, and H. Kaessmann, "The Life History of Retrocopies Illuminates the Evolution of New Mammalian Genes," *Genome Research* 26 (2016): 301–314.

44. Y. Zhang, S. Li, A. Abyzov, and M. B. Gerstein, "Landscape and Variation of Novel Retroduplications in 26 Human Populations," *PLOS Computational Biology* 13 (2017): e1005567.

45. Y. Bai, C. Casola, C. Feschotte, and E. Betran, "Comparative Genomics Reveals a Constant Rate of Origination and Convergent Acquisition of Functional Retrogenes in Drosophila," *Genome Biology* 8 (2007): R11.

46. M. Fablet, M. Bueno, L. Potrzebowski, and H. Kaessmann, "Evolutionary Origin and Functions of Retrogene Introns," *Molecular Biology and Evolution* 26 (2009): 2147–2156.

47. K. Okamura and K. Nakai, "Retrotransposition as a Source of New Promoters," *Molecular Biology and Evolution* 25 (2008): 1231–1238.

48. M. Sorourian, M. M. Kunte, S. Domingues, et al., "Relocation Facilitates the Acquisition of Short Cis-Regulatory Regions That Drive the Expression of Retrogenes during Spermatogenesis in Drosophila," *Molecular Biology and Evolution* 31 (2014): 2170–2180.

49. H. Kaessmann, N. Vinckenbosch, and M. Long, "RNA-Based Gene Duplication: Mechanistic and Evolutionary Insights," *Nature Reviews Genetics* 10 (2009): 19–31.

50. R. L. Troskie, Y. Jafrani, T. R. Mercer, A. D. Ewing, G. J. Faulkner, and S. W. Cheetham, "Long-Read cDNA Sequencing Identifies Functional Pseudogenes in the Human Transcriptome," *Genome Biology* 22 (2021): 146.

51. R. S. Illingworth, U. Gruenewald-Schneider, S. Webb, et al., "Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome," *PLoS Genetics* 6 (2010): e1001134.

52. T. K. Kundu and M. R. Rao, "CpG Islands in Chromatin Organization and Gene Expression," *Journal of Biochemistry* 125 (1999): 217.

53. B. V. Chakravarthi, P. Dedigama-Arachchige, S. Carskadon, et al., "Pseudogene Associated Recurrent Gene Fusion in Prostate Cancer," *Neoplasia* 21 (2019): 989–1002.

54. M. Long and C. H. Langley, "Natural Selection and the Origin of Jingwei, a Chimeric Processed Functional Gene in Drosophila," *Science* 260 (1993): 91–95.

55. D. R. Schrider, F. C. P. Navarro, P. A. F. Galante, et al., "Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans," *PLoS Genetics* 9 (2013): e1003242.

56. W. Wang, F. G. Brunet, E. Nevo, and M. Long, "Origin of Sphinx, a Young Chimeric RNA Gene in Drosophila Melanogaster," *Proceedings of the National Academy of Sciences* 99 (2002): 4448–4453.

57. J. Lai, M. L. Lehman, M. E. Dinger, et al., "A Variant of the KLK4 Gene is Expressed as a Cis Sense–antisense Chimeric Transcript in Prostate Cancer Cells," *RNA* 16 (2010): 1156–1166.

58. M. R. Kubiak, M. W. Szcześniak, and I. Makałowska, "Complex Analysis of Retroposed Genes' Contribution to Human Genome, Proteome and Transcriptome," *Genes* 11 (2020): 542.

59. R. Denzler, V. Agarwal, J. Stefano, D. P. Bartel, and M. Stoffel, "Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance," *Molecular Cell* 54 (2014): 766–776.

60. L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, "A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language?," *Cell* 146 (2011): 353–358.

61. D. W. Thomson and M. E. Dinger, "Endogenous microRNA Sponges: Evidence and Controversy," *Nature Reviews Genetics* 17 (2016): 272–283.

62. E. Chiefari, S. Iiritano, F. Paonessa, et al., "Pseudogene-Mediated Posttranscriptional Silencing of HMGA1 Can Result in Insulin Resistance and Type 2 Diabetes," *Nature Communications* 1 (2010): 40.

63. J. J. Chiang, K. M. J. Sparrer, M. van Gent, et al., "Viral Unmasking of Cellular 5S rRNA Pseudogene Transcripts Induces RIG-I-Mediated Immunity," *Nature Immunology* 19 (2018): 53–62.

64. P. Johnsson, A. Ackley, L. Vidarsdottir, et al., "A Pseudogene Long-Noncoding-RNA Network Regulates PTEN Transcription and Translation in Human Cells," *Nature Structural & Molecular Biology* 20 (2013): 440–446.

65. P. G. Hawkins and K. V. Morris, "Transcriptional Regulation of Oct4 by a Long Non-Coding RNA Antisense to Oct4-Pseudogene 5," *Transcription* 1 (2010): 165–175.

66. S. A. Korneev, J. H. Park, and M. O'Shea, "Neuronal Expression of Neural Nitric Oxide Synthase (nNOS) Protein is Suppressed by an Antisense RNA Transcribed from an NOS Pseudogene," *The Journal of Neuroscience* 19 (1999): 7711–7720.

67. W.-L. Chan, C.-Y. Yuo, W.-K. Yang, et al., "Transcribed Pseudogene ψPPM1K Generates Endogenous siRNA to Suppress Oncogenic Cell Growth in Hepatocellular Carcinoma," *Nucleic Acids Research* 41 (2013): 3734–3747.

68. X. Guo, M. Lin, S. Rockowitz, H. M. Lachman, and D. Zheng, "Characterization of Human Pseudogene-Derived Non-Coding RNAs for Functional Potential," *PLoS ONE* 9 (2014): e93972.

69. O. H. Tam, A. A. Aravin, P. Stein, et al., "Pseudogene-Derived Small Interfering RNAs Regulate Gene Expression in Mouse Oocytes," *Nature* 453 (2008): 534–538.

70. T. Watanabe, Y. Totoki, A. Toyoda, et al., "Endogenous siRNAs from Naturally Formed dsRNAs Regulate Transcripts in Mouse Oocytes," *Nature* 453 (2008): 539–543.

71. X. Hu, L. Yang, and Y. Y. Mo, "Role of Pseudogenes in Tumorigenesis," *Cancers (Basel)* 10 (2018): 256.

72. M. Brosch, G. I. Saunders, A. Frankish, et al., "Shotgun Proteomics Aids Discovery of Novel Protein-Coding Genes, Alternative Splicing, and "Resurrected" Pseudogenes in the Mouse Genome," *Genome Research* 21 (2011): 756–767.

73. M.-L. Dubois, A. Meller, S. Samandi, et al., "UBB Pseudogene 4 Encodes Functional Ubiquitin Variants," *Nature Communications* 11 (2020): 1306.

74. Z. Ji, R. Song, A. Regev, and K. Struhl, "Many lncRNAs, 5′UTRs, and Pseudogenes are Translated and Some are Likely to Express Functional Proteins," *eLife* 4 (2015): e08890.

75. M.-S. Kim, S. M. Pinto, D. Getnet, et al., "A Draft Map of the Human Proteome," *Nature* 509 (2014): 575–581.

76. J. St-Germain, M. R. Khan, V. Bavykina, et al., "Functional Characterization of a Phf8 Processed Pseudogene in the Mouse Genome," *Genes* 14 (2023): 172.

77. S. van Heesch, F. Witte, V. Schneider-Lunitz, et al., "The Translational Landscape of the Human Heart," *Cell* 178 (2019): 242–260.

78. M. Y. Dennis, X. Nuttle, P. H. Sudmant, et al., "Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication," *Cell* 149 (2012): 912–922.

79. I. T. Fiddes, G. A. Lodewijk, M. Mooring, et al., "Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis," *Cell* 173 (2018): 1356–1369.

80. H. Hayashi, T. Arao, Y. Togashi, et al., "The OCT4 Pseudogene POU5F1B is Amplified and Promotes an Aggressive Phenotype in Gastric Cancer," *Oncogene* 34 (2015): 199–208.

81. J. Zhang, X. Wang, M. Li, et al., "NANOGP8 is a Retrogene Expressed in Cancers," *The FEBS Journal* 273 (2006): 1723–1730.

82. M. Kandouz, A. Bier, G. D. Carystinos, M. A. Alaoui-Jamali, and G. Batist, "Connexin43 pseudogene is Expressed in Tumor Cells and Inhibits Growth," *Oncogene* 23 (2004): 4763–4770.

83. G. Liu, X. Cui, H. Li, and L. Cai, "Evolutionary Direction of Processed Pseudogenes," *Science China Life Sciences* 59 (2016): 839–849.

84. A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D. A. Hume, "Mammalian RNA Polymerase II Core Promoters: Insights from Genome-Wide Studies," *Nature Reviews Genetics* 8 (2007): 424–436.

85. P. Huang, C. A. Keller, B. Giardine, et al., "Comparative Analysis of Three-Dimensional Chromosomal Architecture Identifies a Novel Fetal Hemoglobin Regulatory Element," *Genes & Development* 31 (2017): 1704–1713.

86. J. M. Bischof, A. P. Chiang, T. E. Scheetz, et al., "Genome-Wide Identification of Pseudogenes Capable of Disease-Causing Gene Conversion," *Human Mutation* 27 (2006): 545.

87. P. Concolino and A. Costella, "Congenital Adrenal Hyperplasia (CAH) due to 21-Hydroxylase Deficiency: a Comprehensive Focus on 233 Pathogenic Variants of CYP21A2 Gene," *Molecular Diagnosis & Therapy* 22 (2018): 261–280.

88. A. M. Rygiel, S. Beer, P. Simon, et al., "Gene Conversion between Cationic Trypsinogen (PRSS1) and the Pseudogene Trypsinogen 6 (PRSS3P2) in Patients with Chronic Pancreatitis," *Human Mutation* 36 (2015): 350.

89. Vanita, V. Sarhadi, A. Reis, et al., "A Unique Form of Autosomal Dominant Cataract Explained by Gene Conversion Between β-Crystallin B2 and its Pseudogene," *Journal of Medical Genetics* 38 (2001): 392.

90. J. M. Dowen, Z. P. Fan, D. Hnisz, et al., "Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes," *Cell* 159 (2014): 374–387.

91. G. Li, M. J. Fullwood, H. Xu, et al., "ChIA-PET Tool for Comprehensive Chromatin Interaction Analysis with Paired-End Tag Sequencing, Genome Biology," *Genome Biology* 11 (2010): R22.

92. S. S. P. Rao, M. H. Huntley, N. C. Durand, et al., "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell* 159 (2014): 1665–1680.

93. Y. Du and Y. Gao, "Development and Validation of a Novel Pseudogene Pair-Based Prognostic Signature for Prediction of Overall Survival in Patients with Hepatocellular Carcinoma," *BMC Cancer* 20 (2020): 887.

94. J. Kwon, Y. V. Liu, C. Gao, et al., "Pseudogene-Mediated DNA Demethylation Leads to Oncogene Activation," *Science Advances* 7 (2021): eabg1695.

95. B. Nsengimana, F. A. Khan, U. A. Awan, et al., "Pseudogenes and Liquid Phase Separation in Epigenetic Expression," *Frontiers in Oncology* 12 (2022): 912282.

96. C. Sisu, "Pseudogenes as Biomarkers and Therapeutic Targets in Human Cancers," *Methods in Molecular Biology* 2324 (2021): 319.

97. L. Tan, Y. Qin, R. Xie, et al., "N6-Methyladenosine-Associated Prognostic Pseudogenes Contribute to Predicting Immunotherapy Benefits and Therapeutic Agents in Head and Neck Squamous Cell Carcinoma," *Theranostics* 12 (2022): 7267–7288.

98. L. Gao, W. Ren, L. Zhang, et al., "PTENp1, a Natural Sponge of miR-21, Mediates PTEN Expression to Inhibit the Proliferation of Oral Squamous Cell Carcinoma," *Molecular Carcinogenesis* 56 (2017): 1322–1334.

99. T. Gong, S. Zheng, S. Huang, et al., "PTENP1 inhibits the Growth of Esophageal Squamous Cell Carcinoma by Regulating SOCS6 Expression and Correlates with Disease Prognosis," *Molecular Carcinogenesis* 56 (2017): 2610–2619.

100. J. Liu, Y. Xing, L. Xu, W. Chen, W. Cao, and C. Zhang, "Decreased Expression of Pseudogene PTENP1 Promotes Malignant Behaviours and is Associated with the Poor Survival of Patients with HNSCC," *Scientific Reports* 7 (2017): 41179.

101. L. Poliseno, A. Haimovic, P. J. Christos, et al., "Deletion of PTENP1 Pseudogene in Human Melanoma," *Journal of Investigative Dermatology* 131 (2011): 2497–2500.

102. S. Yndestad, E. Austreid, K. O. Skaftnesmo, P. E. Lønning, and H. P. Eikesdal, "Divergent Activity of the Pseudogene PTENP1 in ER-Positive and Negative Breast Cancer," *Molecular Cancer Research* 16 (2018): 78–89.

103. M. Zou, E. Y. Baitei, A. S. Alzahrani, et al., "Oncogenic Activation of MAP Kinase by BRAF Pseudogene in Thyroid Tumors," *Neoplasia* 11 (2009): 57–63.

104. A. M. Habib, A. L. Okorokov, M. N. Hill, et al., "Microdeletion in a FAAH Pseudogene Identified in a Patient with High Anandamide Concentrations and Pain Insensitivity," *British Journal of Anaesthesia* 123 (2019): e249–e253.

105. E. A. Brown, P. J. Dickinson, T. Mansour, et al., "FGF4 retrogene on CFA12 is Responsible for Chondrodystrophy and Intervertebral Disc Disease in Dogs," *Proceedings of the National Academy of Sciences* 114 (2017): 11476–11481.

106. H. G. Parker, B. M. VonHoldt, P. Quignon, et al., "An Expressed fgf4 Retrogene is Associated with Breed-Defining Chondrodysplasia in Domestic Dogs," *Science* 325 (2009): 995–998.

107. M. Sun, L. Chang, L. He, et al., "Combining Single-Cell Profiling and Functional Analysis Explores the Role of Pseudogenes in Human Early Embryonic Development," *Journal of Genetics and Genomics* 51 (2024): 1173–1186.

108. N. A. Rapicavoli, K. Qu, J. Zhang, M. Mikhail, R. M. Laberge, and H. Y. Chang, "A Mammalian Pseudogene lncRNA at the Interface of Inflammation and Anti-Inflammatory Therapeutics," *eLife* 2 (2013): e00762.

109. ENCODE Project Consortium, "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature* 489 (2012): 57.

110. P. Kapranov, J. Cheng, S. Dike, et al., "RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription," *Science* 316 (2007): 1484–1488.

111. A. M. Garrett, P. J. Bosch, D. M. Steffen, et al., "CRISPR/Cas9 Interrogation of the Mouse Pcdhg Gene Cluster Reveals a Crucial Isoform-Specific Role for Pcdhgc4," *PLOS Genetics* 15 (2019): e1008554.

112. L. Duret, C. Chureau, S. Samain, J. Weissenbach, and P. Avner, "The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene," *Science* 312 (2006): 1653–1665.

113. H. H. Stedman, B. W. Kozyak, A. Nelson, et al., "Myosin Gene Mutation Correlates with Anatomical Changes in the Human Lineage," *Nature* 428 (2004): 415–418.

114. J. A. Carlisle, A. L. Bickell, and M. W. Hart, "Sexual Selection and Pseudogenization in Primate Fertilization," *Evolution* 79 (2025): 1670–1680.

115. J. A. Carlisle, D. H. Gurbuz, and W. J. Swanson, "Recurrent Independent Pseudogenization Events of the Sperm Fertilization Gene ZP3r in Apes and Monkeys," *Journal of Molecular Evolution* 92 (2024): 695–702.

116. G. Drouin, J. R. Godin, and B. Pagé, "The Genetics of Vitamin C Loss in Vertebrates," *Current Genomics* 12 (2011): 371–378.

117. A. A. M. Gad and A. Sirko, "L-Gulono-γ-Lactone Oxidase, the Key Enzyme for L-Ascorbic Acid Biosynthesis," *Current Issues in Molecular Biology* 46 (2024): 11057–11074.

118. Y. Inai, Y. Ohta, and M. Nishikimi, "The Whole Structure of the Human Nonfunctional L-Gulono-.GAMMA.-Lactone Oxidase Gene-the Gene Responsible for Scurvy-and the Evolution of Repetitive Sequences Thereon," *Journal of Nutritional Science and Vitaminology* 49 (2003): 315–319.

119. J. T. Kratzer, M. A. Lanaspa, M. N. Murphy, et al., "Evolutionary History and Metabolic Insights of Ancient Mammalian Uricases," *Proceedings of the National Academy of Sciences* 111 (2014): 3763–3768.

120. A. V. Yeldandi, V. Yeldandi, S. Kumar, et al., "Molecular Evolution of the Urate Oxidase-Encoding Gene in Hominoid Primates: Nonsense Mutations," *Gene* 109 (1991): 281–284.

121. M. Behrens, "Pharmacology of TAS1R2/TAS1R3 Receptors and Sweet Taste," *Handb Exp Pharmacol* 275 (2022): 155.

122. X. Li, W. Li, H. Wang, et al., "Pseudogenization of a Sweet-Receptor Gene Accounts for Cats' indifference toward Sugar," *PLoS Genet* 1 (2005): 27.

123. E. Z. Kvon, O. K. Kamneva, U. S. Melo, et al., "Progressive Loss of Function in a Limb Enhancer during Snake Evolution," *Cell* 167 (2016): 633–642.

124. D. M. Martill, H. Tischlinger, and N. R. Longrich, "A Four-Legged Snake from the Early Cretaceous of Gondwana," *Science* 349 (2015): 416–419.

125. A. V. Gore, K. A. Tomins, J. Iben, et al., "An Epigenetic Mechanism for Cavefish Eye Degeneration," *Nature Ecology & Evolution* 2 (2018): 1155–1160.

126. L. Ma, A. V. Gore, D. Castranova, et al., "A Hypomorphic Cystathionine Ss-Synthase Gene Contributes to Cavefish Eye Loss by Disrupting Optic Vasculature," *Nature Communications* 11 (2020): 2772.

127. S. Chen, B. H. Krinsky, and M. Long, "New Genes as Drivers of Phenotypic Evolution," *Nature Reviews Genetics* 14 (2013): 645–660.

128. W. Gilbert, S. J. de Souza, and M. Long, "Origin of Genes," *Proceedings of the National Academy of Sciences* 94 (1997): 7698–7703.

129. J. Zhang, "Evolution by Gene Duplication: an Update," *Trends in Ecology & Evolution* 18 (2003): 292–298.

130. D. Kanber, K. Buiting, C. Roos, et al., "The Origin of the RB1 Imprint," *PLoS ONE* 8 (2013): e81502.

131. H. Hezroni, R. Ben-Tov Perry, Z. Meir, G. Housman, Y. Lubelsky, and I. Ulitsky, "A Subset of Conserved Mammalian Long Non-Coding RNAs are Fossils of Ancestral Protein-Coding Genes," *Genome Biology* 18 (2017): 162.

132. C. Bekpen, T. Marques-Bonet, C. Alkan, et al., "Death and Resurrection of the Human IRGM Gene," *PLoS Genetics* 5 (2009): e1000403.

133. C. Bekpen, R. J. Xavier, and E. E. Eichler, "Human IRGM Gene "to be or Not to be, Human IRGM Gene "to be or Not to be"," *Seminars in Immunopathology* 32 (2010): 437–444.

134. A. Anand, C. A. Olson, L. Yang, et al., "Pseudogene Repair Driven by Selection Pressure Applied in Experimental Evolution," *Nature Microbiology* 4 (2019): 386–389.

135. K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, et al., "Predicting Splicing from Primary Sequence with Deep Learning," *Cell* 176 (2019): 535–548.

136. M. Pertea, A. Shumate, G. Pertea, et al., "CHESS: a New Human Gene Catalog Curated from Thousands of Large-Scale RNA Sequencing Experiments Reveals Extensive Transcriptional Noise," *Genome Biology* 19 (2018): 208.

137. I. Rivas-González, M. Rousselle, F. Li, et al., "Pervasive Incomplete Lineage Sorting Illuminates Speciation and Selection in Primates," *Science* 380 (2023): eabn4409.

138. I. Shinder, R. Hu, H. J. Ji, K. H. Chao, and M. Pertea, "EASTR: Identifying and Eliminating Systematic Alignment Errors in Multi-Exon Genes," *Nature Communications* 14 (2023): 7223.

139. A. Varabyou, M. J. Sommer, B. Erdogdu, et al., "CHESS 3: an Improved, Comprehensive Catalog of Human Genes and Transcripts Based on Large-Scale Expression Data, Phylogenetic Analysis, and Protein Structure," *Genome Biology* 24 (2023): 249.

140. M. N. Cabili, C. Trapnell, L. Goff, et al., "Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses," *Genes & Development* 25 (2011): 1915–1972.

141. N. Cloonan, A. R. R. Forrest, G. Kolle, et al., "Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing," *Nature Methods* 5 (2008): 613–619.

142. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq," *Nature Methods* 5 (2008): 621–628.

143. S. Kalyana-Sundaram, C. Kumar-Sinha, S. Shankar, et al., "Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers," *Cell* 149 (2012): 1622–1634.

144. E. K. Gustavsson, S. Sethi, Y. Gao, et al., "The Annotation of GBA1 Has Been Concealed by its Protein-Coding Pseudogene GBAP1," *Science Advances* 10 (2024): eadk1296.

145. K. F. Au, V. Sebastiano, P. T. Afshar, et al., "Characterization of the Human ESC Transcriptome by Hybrid Sequencing," *Proc Natl Acad Sci U S A* 110 (2013): E4821.

146. D. R. Garalde, E. A. Snell, D. Jachimowicz, et al., "Highly Parallel Direct RNA Sequencing on an Array of Nanopores," *Nature Methods* 15 (2018): 201–206.

147. D. A. Glinos, G. Garborcauskas, P. Hoffman, et al., "Transcriptome Variation in Human Tissues Revealed by Long-Read Sequencing," *Nature* 608 (2022): 353–359.

148. S. Oikonomopoulos, Y. C. Wang, H. Djambazian, D. Badescu, and J. Ragoussis, "Benchmarking of the Oxford Nanopore MinION Sequencing for Quantitative and Qualitative Assessment of cDNA Populations," *Scientific Reports* 6 (2016): 31602.

149. F. J. Pardo-Palacios, D. Wang, F. Reese, et al., "Systematic Assessment of Long-Read RNA-seq Methods for Transcript Identification and Quantification," *Nature Methods* 21 (2024): 1349–1363.

150. R. I. Kuo, Y. Cheng, R. Zhang, et al., "Illuminating the Dark Side of the Human Transcriptome with Long Read Transcript Sequencing," *BMC Genomics* 21 (2020): 751.

151. S. H. Qian, L. Chen, Y. L. Xiong, and Z. X. Chen, "Evolution and Function of Developmentally Dynamic Pseudogenes in Mammals," *Genome Biology* 23 (2022): 235.

152. N. A. Salem, A. H. Mahnke, A. M. Tseng, et al., "A Novel Oct4/Pou5f1-Like Non-Coding RNA Controls Neural Maturation and Mediates Developmental Effects of Ethanol," *Neurotoxicology and Teratology* 83 (2021): 106943.

153. X. Yang, S. Yang, Y. Luo, et al., "Single-Cell Transcriptional Footprint for Pseudogene SsCLEC9A is Associated with Antigen Processing and Presentation in Sus scrofa," *International Journal of Biological Macromolecules* 302 (2025): 140629.

154. C. Hu, T. Li, Y. Xu, et al., "CellMarker 2.0: an Updated Database of Manually Curated Cell Markers in Human/Mouse and Web Tools Based on scRNA-seq Data," *Nucleic Acids Res* 6, no. 51 (2023): D870.

155. GTEx Consortium, "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues," *Science* 369 (2020): 1318.

156. L. Han, Y. Yuan, S. Zheng, et al., "The Pan-Cancer Analysis of Pseudogene Expression Reveals Biologically and Clinically Relevant Tumour Subtypes," *Nature Communications* 5 (2014): 3963.

157. K. J. Karczewski, L. C. Francioli, G. Tiao, et al., "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans," *Nature* 581 (2020): 434–443.

158. A. W. Cheng, H. Wang, H. Yang, et al., "Multiplexed Activation of Endogenous Genes by CRISPR-on, an RNA-Guided Transcriptional Activator System," *Cell Research* 23 (2013): 1163–1171.

159. L. Cong, F. A. Ran, D. Cox, et al., "Multiplex Genome Engineering Using CRISPR/Cas Systems," *Science* 339 (2013): 819–823.

160. L. A. Gilbert, M. A. Horlbeck, B. Adamson, et al., "Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation," *Cell* 159 (2014): 647–661.

161. N. C. Yeo, A. Chavez, A. Lance-Byrne, et al., "An Enhanced CRISPR Repressor for Targeted Mammalian Gene Regulation," *Nature Methods* 15 (2018): 611–616.

162. N. Kawamura, K. Nimura, H. Nagano, S. Yamaguchi, N. Nonomura, and Y. Kaneda, "CRISPR/Cas9-Mediated Gene Knockout of NANOG and NANOGP8 Decreases the Malignant Potential of Prostate Cancer Cells," *Oncotarget* 6 (2015): 22361–22374.

163. M. Vitiello, M. Evangelista, Y. Zhang, L. Salmena, P. P. Pandolfi, and L. Poliseno, "PTENP1 is a ceRNA for PTEN: It's CRISPR Clear," *Journal of Hematology & Oncology* 13 (2020): 73.

164. M. Sun, Y. Wang, C. Zheng, et al., "Systematic Functional Interrogation of Human Pseudogenes Using CRISPRi," *Genome Biology* 22 (2021): 240.

165. J. S. Lagas, M. F. Sentmanat, and X. Cui, "Efficient GBA1 Editing via HDR with ssODNs by Outcompeting Pseudogene-Mediated Gene Conversion Upon CRISPR/Cas9 Cleavage," *Frontiers in Genome Editing* 7 (2025): 1581743.

166. A. Levasseur and P. Pontarotti, "The Role of Duplications in the Evolution of Genomes Highlights the Need for Evolutionary-Based Approaches in Comparative Genomics," *Biology Direct* 6 (2011): 11.

167. M. Arenas and D. Posada, "Simulation of Genome-Wide Evolution Under Heterogeneous Substitution Models and Complex Multispecies Coalescent Histories," *Molecular Biology and Evolution* 31 (2014): 1295–1301.

Biographies



Dr. Jia Yu is a Professor at Peking Union Medical College. His work focuses on hematopoietic regulation, stem cell fate, and translational medicine. He has achieved research outcomes in areas including cellular atlases of tissue and organ development and disease, novel mechanisms of non-coding RNA biogenesis and processing, new regulatory mechanisms of stem cell fate, and translational medicine research. As a corresponding author, he has published over 40 SCI-indexed papers in internationally recognized

academic journals including *Nature Cell Biology* and so on, with his publications receiving a total of more than 5,000 citations.



Dr. Yanni Ma is a Professor at Peking Union Medical College. Her research centers on RNA and hematopoietic regulation, where she investigates fundamental mechanisms of stem cell fate transitions, hematopoietic development and differentiation, and disease pathogenesis from the perspective of RNA metabolism, while also developing RNA-based therapeutics for diseases. She has made several important discoveries and has published multiple papers as corresponding or first author in academic journals including *Nature Cell Biology* and *Cell Stem Cell*.



Dr. Mengyao Sun is a Postdoctoral Fellow at Peking Union Medical College. Her primary research focuses on non-coding genes and stem cell biology, where she investigates the regulatory functions of pseudogenes at multiple molecular levels and studies the mechanisms underlying stemness maintenance in embryonic stem cells and hematopoietic stem cells.