

**ORIGINAL RESEARCH**

# Analysis of new retrogenes provides insight into dog adaptive evolution

Xiang Gao<sup>1</sup>  | Yan Li<sup>2</sup> | Adeyinka A. Adetula<sup>3</sup>  | Yu Wu<sup>4</sup> | Hong Chen<sup>5</sup><sup>1</sup>Center Laboratory, Renmin Hospital of Wuhan University, Wuhan, China<sup>2</sup>Department of Infectious Diseases, Zhongnan Hospital of Wuhan University, Wuhan, China<sup>3</sup>Key Laboratory of Agricultural Animal Genetics, Breeding, and Reproduction, Huazhong Agricultural University, Wuhan, China<sup>4</sup>Oilfield Community D-1-902, Wuhan, China<sup>5</sup>Department of Scientific Research, Renmin Hospital of Wuhan University, Wuhan, China**Correspondence**Yu Wu, Oilfield community D-1-902, No. 294 Parrot Avenue, Hanyang District, Wuhan 430050, China.  
Email: naida\_yin@whlhy.comHong Chen, Department of Scientific Research, Renmin Hospital of Wuhan University, Wuhan 430060, China.  
Email: rmchenhong@whu.edu.cn**Funding information**

The Hubei Provincial Natural Science Foundation of China, Grant/Award Number: 2015CFB350

**Abstract**

The origin and subsequent evolution of new genes have been considered as an important source of genetic and phenotypic diversity in organisms. Dog breeds show great phenotypic diversity for morphological, physiological, and behavioral traits. However, the contributions of newly originated retrogenes, which provide important genetic bases for dog species differentiation and adaptive traits, are largely unknown. Here, we analyzed the dog genome to identify new RNA-based duplications and comprehensively investigated their origin, evolution, functions in adaptive traits, and gene movement processes. First, we totally identified 3,025 retrocopies including 476 intact retrogenes, 2,518 retropseudogenes, and 31 chimerical retrogenes. Second, selective pressure along with ESTs expression analysis showed that most of the intact retrogenes were significantly under stronger purifying selection and subjected to more functional constraints when compared to retropseudogenes. Furthermore, a large number of retrocopies and chimerical retrogenes that occurred approximately 22 million years ago implied a burst of retrotransposition in the dog genome after the divergence time between dog and its closely related species red fox. Interestingly, GO and pathway analyses showed that new retrogenes had expanded in glutathione biosynthetic/metabolic process which likely provided important genetic basis for dogs' adaptation to scavenge human waste dumps. Finally, consistent with the results in human and mouse, a significant excess of functional retrogenes movement on and off the X chromosome in the dog confirmed a general pattern of gene movement process in mammals which was likely driven by natural selection or sexual antagonism. Together, these results increase our understanding that new retrogenes can reshape the dog genome and provide further exploration of the molecular mechanisms underlying the dogs' adaptive evolution.

**KEYWORDS**

adaptive evolution, gene movement, new genes, on and off the X chromosome, retrogene

## 1 | INTRODUCTION

Since the age of Darwin, biologists have been following an essential question: How do organisms evolve from their common ancestor to a rich variety of species? The origin and subsequent evolution of novel genes have been taken into account as a major contributor to adaptive evolution (Kaessmann, 2010). Novel genes provide important genetic novelties associated with biological diversity in organisms (Chen, Krinsky, & Long, 2013) and significantly contribute to the evolution of lineage-specific or species-specific phenotypic traits (Chen, Zhang, & Long, 2010). The molecular processes that can generate new genes mainly include de novo origin, exon or domain shuffling, gene duplication, retrotransposition, TE domestication, gene fission or fusion, gene lateral transfer, and so on (Chen et al., 2013). These mechanisms also cooperate in creating a new gene (Chen et al., 2013; Kaessmann, 2010; Long, Betran, Thornton, & Wang, 2003; Long, VanKuren, Chen, & Vibranovski, 2013).

Retrotransposition is a special RNA-based duplication mechanism in which transcribed and spliced mRNA is occasionally reverse transcribed and integrated into a new DNA locus to form a retrogene (Betran, Thornton, & Long, 2002). Retrogenes have long been regarded as nonexpressed pseudogenes and evolutionary dead ends due to the lack of regulatory sequences such as promoters. However, extensive structural variations in retrocopies have been speculated as “evolutionary seeds” for the evolution of new genes with novel functions if they casually acquired new regulatory elements or coding sequences by expression (Pan & Zhang, 2009). Retrotransposition is an important mechanism of gene duplication and has produced a large number of functional genes in mammalian genomes (Marques, Dupanloup, Vinckenbosch, Reymond, & Kaessmann, 2005). Recent studies have revealed that a substantial number of “processed genes” or “retrogenes” with novel functions are derived from the mRNA of various intron-containing genes (Chen et al., 2012; Long & Langley, 1993; Parker et al., 2009; Rosso et al., 2008). For example, *Jingwei*, a new chimeric retrogene only in *Drosophila teissieri* and *Drosophila yakuba*, plays an important role in the metabolism of recruitment pheromones and juvenile hormones (Long & Langley, 1993; Zhang, Yang, Long, Li, & Dean, 2010). *Sphinx*, a *D. melanogaster*-specific non-coding RNA gene by retrotranscription, has an impact on male courtship behaviors (Dai et al., 2008). The hominoid-specific *CDC14Bretro* gene is subjected to a short period of intense positive selection in the African ape ancestor 7–12 Mya and leads to rapid redistribution from microtubules to a new cellular location (Rosso et al., 2008). For RNA-based duplication pairs, ancestral genes and newly originated retrogenes are easily identified by inspecting their gene structures. Because the ancestral gene contains multiple protein-coding exons, the retrogene contains a single protein-coding exon. Therefore, retrogenes provide a convenient and feasible resource to investigate the origin and evolution of new genes.

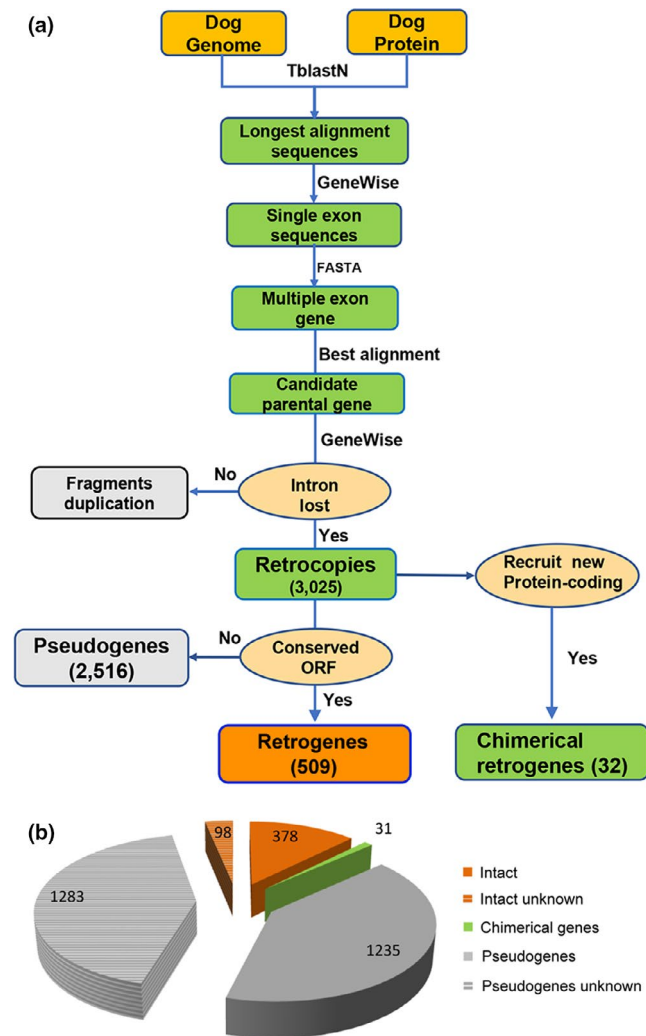
The dogs occupy a special niche in mammalian genomics because they provide important evolutionary information within the mammal. Existing dog breeds show great phenotypic diversity for morphological, physiological, and behavioral traits (Lindblad-Toh

et al., 2005). The dogs are also the first organism that humans domesticated before any plant and any other animal. Previous evidence has revealed that ancient dogs gradually separated from other gray wolves and were involved into humans' life as early as 20,000–40,000 years before present (YBP; Figure 1; Botigue et al., 2017; Skoglund, Ersmark, Palkopoulou, & Dalen, 2015). Constantly following human migration, dogs' diversity has been greatly influenced in many aspects such as breeding, migration, hybridization, invasion, and decimation or assimilation of local populations. The domestic dogs, then, play important roles for geneticists in studying genes and sequence-level variations which are associated with morphology, susceptibility to simple and complex diseases, and behavior (Malmstrom et al., 2008). For example, genetic variants of IGF1 contribute to a small size phenotype associated with domestication (Sutter et al., 2007). A series of genetic variations have been reported to be associated with the phenotype of coat color in the dog (Berryere, Kerns, Barsh, & Schmutz, 2005; Clark, Wahl, Rees, & Murphy, 2006; Karlsson et al., 2007). A variant in  $\beta$ -defensin 103 (CBD103), which was previously associated with immune function, results in the dominant phenotype of black coat color in the dog (Candille et al., 2007). SLC1A2 gene is significantly associated with aggression toward strangers (Takeuchi et al., 2009). Interestingly, the expression of a newly originated retrogene fibroblast growth factor 4 (*fgf4*) is strongly associated with chondrodysplasia, a short-legged phenotype that defined at least 19 dog breeds which include dachshund, corgi, and basset hound (Parker et al., 2009). These results indicate that a single gene can play critical roles in constraining and directing phenotypic diversity in the dog. However, there are very few reports about the sustainable evolution and innovative functions of new genes in the dog.

In this study, we used dog genome sequence, protein sequence, and ESTs data to identify many retrogenes and chimerical retrogenes



**FIGURE 1** The Labrador retriever



**FIGURE 2** Identification of dog retrogenes. (a), The main procedure for identifying retrogenes. (b), Identification of different categories of retrocopies

that originated ~22 million years ago (Mya) after the divergence time between dog and its closely related species red fox. By analyzing their selective pressure and functions, we found that new retrogenes were likely available for improving dogs' adaptation to scavenge human waste dumps. Furthermore, we also detected a biased movement of functional retrogenes on and off the X chromosomes, which was likely driven by natural selection or sexual antagonism. This study will not only increase our understanding that new retrogenes can reshape the dog genome but also allow further exploration of the molecular mechanisms underlying the dogs' adaptive evolution.

## 2 | MATERIALS AND METHODS

### 2.1 | Identification of dog retrogenes

To identify parental–retrogene pairs in the dog, we used a similar algorithm that was presented in previous studies (Figure 2a; Fu,

Chen, Zou, Long, & He, 2010; Marques et al., 2005). Dog genome sequence and all dog annotated peptide sequences were downloaded from the Ensembl database (Release 90). To detect retrogenes, we aligned the 25,882 annotated protein sequences to the whole dog genome sequence using TblastN with an  $E$ -value threshold at  $10^{-3}$ . The homology sequences (with identity >50%, overlap >50% and a minimum length of 50 amino acids) were maintained for assurance of high-quality alignments. GeneWise software was used to define the intron–exon boundary of the merged target sequences with default settings (score >35). The single exon sequences were considered as candidate copies. Then, the FASTA software was used to perform similarity searches of candidate copies against all protein genes. If the best alignment of a single exon candidate copy was a protein gene with multiple coding exons, it was regarded as a candidate retrocopy. We then checked whether the introns of the parental gene (the best hit) had been lost in the retrocopies. If introns were retained, this retrocopy may be false positive and would be discarded. The retrocopies with either premature stop codons or frameshift mutations were defined as retrospseudogenes. If one retrocopy could recruit novel regulatory elements and new protein-coding exons and evolve into a functional retrogene, it was defined as a chimerical retrogene.

### 2.2 | Divergence time and functionality analysis

We used the PAML35 package to calculate  $K_a$  (nonsynonymous substitution rate of nonsynonymous sites),  $K_s$  (synonymous substitution rate of synonymous sites), and  $K_a/K_s$  (i.e.,  $\omega$ ) between the parental gene and its retrocopy (Yang, 2007). According to the formula  $T = K_s/2\lambda$  ( $\lambda$ : synonymous mutation rate), we estimated the origin time of retrocopies based on the  $K_s$  value. In addition, the  $\omega$  value between paternal gene and retrogene pair was used to estimate functional constraints with  $\omega < 1$  representing the purifying selection,  $\omega = 1$  for the neutral selection, and  $\omega > 1$  for the positive selection. We applied a more stringent criterion ( $\omega < 0.5$ ) to examine the functional constraints of gene pairs. If  $\omega$  value was <0.5, the gene was probably subjected to purifying selection and functional constraints.

### 2.3 | Transcription analysis of the retrogenes

A total of dog ESTs were downloaded from the NCBI database and processed for various contaminants, low-quality and low-complexity sequences using the SeqClean program (<https://sourceforge.net/projects/seqclean/files/>). Then, the clean data were mapped to the dog genome with BLAT software. The alignments, with identity >95% and overlap >90%, were retained for further analysis. As we know, the new parental gene and retrogene pair generally shared a high identity in protein-coding regions. Thus, the ESTs that were mapped to a unique location on the genome and with alignment sequence >100 bp and nucleotide identity of >97% were retained to analyze the transcription abilities of retrogenes.



**FIGURE 3** Three types of retrotransposition events. (a), (b), and (c) represent three types of retrotransposition events, respectively. Green boxes indicate coding sequence regions; orange boxes indicate UTR regions; blue boxes indicate new recruited coding sequence regions; gray box indicates a retrocopy inserted into the intron region of another gene

## 2.4 | GO enrichment and statistical analyses

Considering that the new parental–retrotransposon pair usually shared a high degree of identity in protein-coding regions, it was viable to predict the retrotransposon's function through its corresponding parental gene's function. The dog annotation was relatively limited, so all genes were firstly converted into human homologous genes by HGNC symbol ID in the Ensembl BioMart database. Then, GO terms and pathway enrichment analyses were performed using Metascape (<http://metascape.org>) with the default parameters. And the top 20 significantly enriched clusters were reported (Benjamini–Hochberg corrected  $q < .05$ ). The network graphs were edited using Cytoscape (Shannon et al., 2003). Gene movements among chromosomes were visualized by Circos software package (Krzywinski et al., 2009). All statistical analyses were carried out using the R programming language (version 3.2.0).

## 3 | RESULTS AND DISCUSSION

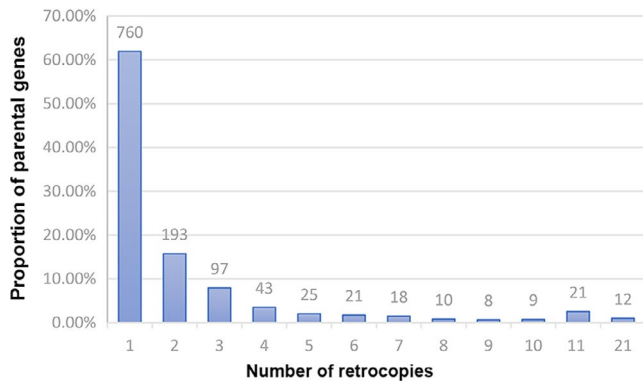
### 3.1 | Identification of dog retrotransposons

The origin of new RNA-based duplication genes has been a significant driver of organismal evolution (Chen et al., 2013; Long et al., 2013). The main principle for identifying retrotransposons was to search gene pairs where one copy was a multiple coding exons gene, whereas the other copy was a single coding exon gene (Figure 2a). In the gene pair, the single coding exon gene was defined as a retrotransposon while the multiple coding exons gene was defined as its parental gene. First, a total of 25,157 annotated dog protein sequences were mapped onto the dog genome with Tblastn software. We obtained the longest alignment sequences that mapped to the proteins. Then, we used the GeneWise program to identify single exon sequences. Only 10,544 single exon sequences were retained as candidate

retrocopies (probably encode single exon protein). Second, we performed similarity searches of the candidate sequences against all proteins using FASTA (Akram et al., 2011) to find their corresponding parental genes (multiple exon genes). The best alignment protein of the candidate sequence was considered as a candidate parental protein (with identity >50%, overlap >50%, and more than 50 amino acids). Third, we used GeneWise program once again to check the gene pair where the parental gene should contain multiple coding exons and the retrotransposon should be a single coding exon gene. For example, a multiple coding exons gene ENSCAFP00000012206.3 was involved in a retrotransposition event and integrated into a new DNA locus. It recruited novel regulatory sequences in the 5' and 3' flanking regions and then generated a new single exon gene ENSCAFG000000031930 (Figure 3a). This single exon gene was defined as a putative retrotransposon. Finally, we totally identified 3,025 candidate retrocopies which originated from 1,227 parental genes (Table S1), indicating that one parental gene could produce about 2.41 retrocopies at an average level.

The distribution of retrocopy numbers (being produced by each parental gene) indicated that most of the parental genes (~62%, 760/1,227) could only generate one retrocopy, while a portion of parental genes could generate more than 10 retrocopies (Figure 4). Astonishingly, we analyzed the top 20 high-yield parental genes and found that ENSCAFG00000005101 (RPSA) generated the most retrocopies ~41, followed by ENSCAFG00000017680 (RPL7A) producing 39 and ENSCAFG00000015077 (GAPDH) producing 37 (Table 1). Interestingly, the top 20 high-yield parental genes were predominantly composed of ribosomal protein genes (60%, 12/20). For one thing, on average of 400 copies of rRNA genes existed in the mammalian genome and distributed in variable orientation on several chromosomes (Moss, Langlois, Gagnon-Kugler, & Stefanovsky, 2007). High quantity of rRNA genes probably increased the opportunity of being involved in retrotransposition event. For another,





**FIGURE 4** Distribution of retrocopy numbers

**TABLE 1** The top 20 high-yield parental genes

Gene ID	Symbol	Retrocopy number
ENSCAFG00000005101	RPSA	41
ENSCAFG00000017680	RPL7A	39
ENSCAFG00000015077	GAPDH	37
ENSCAFG00000008021	RPL7	37
ENSCAFG00000001615	RPS6	33
ENSCAFG00000019472	RPS2	31
ENSCAFG00000005764	RPL15	31
ENSCAFG00000001818	DNAJA1	29
ENSCAFG00000006580	HNRNPA1	28
ENSCAFG00000008640	TUBA1B	25
ENSCAFG00000001318	RPL10A	23
ENSCAFG00000006597	HMGB1	21
ENSCAFG00000020136	RPL12	20
ENSCAFG00000008212	BTF3	20
ENSCAFG00000007990	RPS3A	19
ENSCAFG00000012195	LDHB	18
ENSCAFG000000032728	/	17
ENSCAFG00000003631	RPS11	16
ENSCAFG00000008873	RPL6	15
ENSCAFG00000018174	RPS15A	14

Note: / represents no annotation gene.

the retrotransposon long interspersed element 1 (LINE1) composed the largest proportion of TE-derived sequences. In embryos, LINE1 RNA could mediate binding of nucleolin and Kap1 to rDNA to regulate rRNA synthesis (Percharde et al., 2018). Thus, high-response frequency with LINE1 would contribute to the formation of rDNA retrocopies.

Generally, the retrocopy only maintains the coding region of its parent gene and lose the corresponding regulatory elements such as promoters and enhancers, thus leading to accumulation of nonsense mutations and eventually evolving into a pseudogene. To obtain the expression ability, it should recruit new regulatory elements nearby to become a functional transcriptional gene.

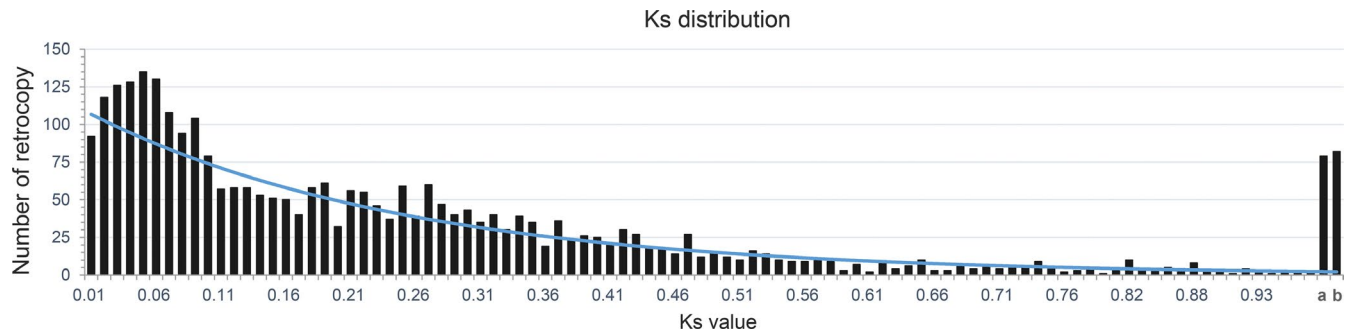
According to our GeneWise analysis, only ~16.8% (507/3,025) retrocopies retained their parent genes' open reading frame (ORF) sequence, which was considered as the "intact retrogenes." The remaining 2,518 retrocopies were defined as pseudogenes as they exhibited either frameshift mutations or premature termination codons. This result was nearly equivalent to the proportion of intact retrogenes in human and rat genome (Kabza, Ciomborowska, & Makalowska, 2014; Marques et al., 2005). However, the total number of retrocopies in the dog was much higher than that in most other animals. One possibility was that a large number of LINEs and SINEs were observed to be overactive in the dog (Mamedov, Arzumanyan, Amosova, Lebedev, & Sverdlov, 2005; Wang & Kirkness, 2005). LINEs and SINEs are the major sources of insertional mutagenesis such as retrotransposition events (Esnault, Maestre, & Heidmann, 2000; Gogvadze & Buzdin, 2005; Schmid, 1998). They can increase the variability and instability of the dog genome and led to generate higher content of retrocopies. Even so, the majority of these retrocopies will be gradually eliminated during the evolutionary process in organisms, while a small part of them evolve into new functional genes by producing truncated protein or expressed lncRNA (Chen et al., 2013; Hirotsune et al., 2003; Podlaha & Zhang, 2004) and subsequently be retained under a long-term natural selection.

If a retrocopy has inserted into the exon/intron structure of another gene (Figure 3b), it may disrupt the existing protein-coding regions or genetic topology and result in deleterious mutations. Occasionally, a portion of retrocopies recruit new regulatory elements and protein-coding exons nearby and then evolve into functional chimerical retrogenes. For example, ENSCAFG000000031782 was originated as insertion of the retrotransposition sequence of ENSCAFP00000011052.3 and recruited nearby novel regulatory sequences and protein-coding regions to form a chimerical structure (Figure 3c). Out of 507 intact retrogenes, we also identified 31 chimerical retrogenes. By recruiting new protein-coding regions, chimerical retrogenes are likely to evolve into novel protein genes and drive genetic innovation and adaptive evolution (Long et al., 2003). These 31 chimerical retrogenes were regarded as intact retrogenes in the following analysis.

In Ensembl release 90, the dog genome annotation GTF file contained ~32,704 genes and ~39,074 transcripts. In contrary, in our retrocopies data set, we detected more than 1,300 retrocopies that did not overlap with any annotated gene, suggesting that they were likely newly annotated genes or nonfunctional fragments (Figure 2b). Nevertheless, 98 of them were intact retrogenes and maintained the conserved ORF sequences of their parental genes. This result indicated that newly annotated retrogenes could not only shape the dog genome but also complement the blank position without annotation.

### 3.2 | A burst of new retrogenes and chimerical retrogenes in the dog genome

In the dog, neutral mutation rate is estimated  $\sim 0.4 \times 10^{-8}$  per generation (Frantz et al., 2016; Skoglund et al., 2015). We used a 3-year



**FIGURE 5**  $K_s$  distribution for 3,025 retrocopies. Retrocopies with  $K_s$  between 1.0–2.0 and  $\geq 2.0$  were, respectively, pooled in a single bin. a, with  $K_s$  between 1.0–2.0; b, with  $K_s \geq 2.0$ . The peak of  $K_s$  in the range between 0.02–0.06 revealed a burst of retrotransposition in dog genome and mirrored the divergent time between dog and its closely related species red fox (estimated in <http://www.timetree.org/>, about 7.5–22.5 Mya)

generation time (corresponding to  $1.33 \times 10^{-9}$  mutations per year) to calculate the divergence time between retrocopies and corresponding parental genes ( $T = K_s/2\lambda$ ). The  $K_s$  distribution of retrocopies showed that a burst of retrotransposition (729, 24%) reached its peak region in the range of  $K_s$  0.02–0.06 (Figure 5), approximately 7.5–22.5 million years ago (Mya), suggesting that a large number of new retrogenes originated during a short evolutionary history. Interestingly, the burst of retrocopies was corresponding to the divergence time between dog and its closely related species red fox (7.0–22.0 Mya from <http://www.timetree.org/>), implying that the retrotransposition was likely to drive the origin of new genes and consequently contributed to their differentiation or phenotypic traits. Consistent with previous studies in other mammals (Marques et al., 2005; Pan & Zhang, 2009), fish (Fu et al., 2010), and plant (Wang et al., 2006), this result confirmed that retrotransposition could play significant roles in both genome remodeling and lineage or species differentiation.

Interestingly, we detected 14 intact retrogenes with  $K_s$  values of 0 which might be indicative of very recent retrogenes. And further investigation of these genes might shed light on new retrogenes which might be involved in dog-specific phenotypic traits. Nevertheless, this study was entirely based on the analysis of the genome of domestic dog and the origin time of new retrogenes was estimated by  $K_s$  value. Thus, a further phylogenetic analysis, by comparing the other genomes such as the gray wolf and the red fox (Gopalakrishnan et al., 2017; Kukekova et al., 2018), would be helpful to validate dog-specific or *Canini*-specific candidate retrogenes. Function analysis of these dog-specific new retrogenes would be highly valuable in revealing the molecular mechanism of artificial domestication.

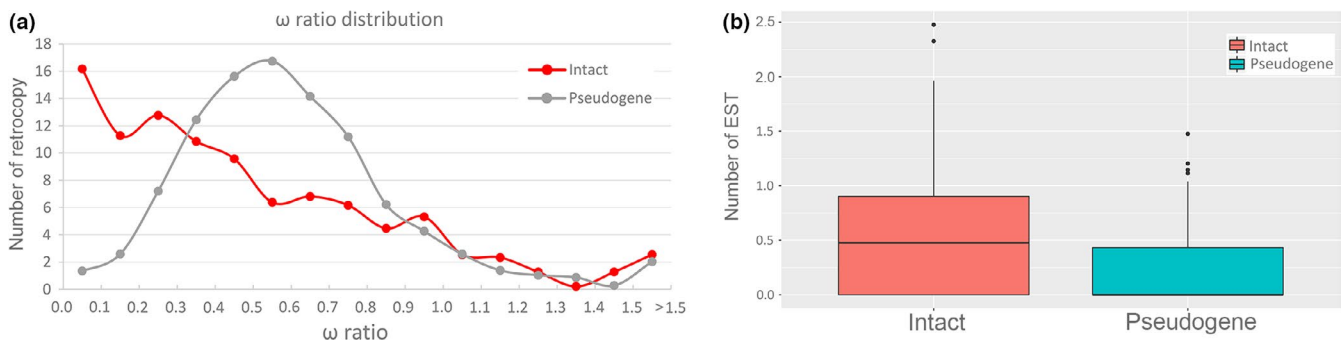
In addition, we calculated the  $K_s$  distribution of 31 chimerical retrogenes and estimated their divergence time (Table S1). Interestingly, nine of them had  $K_s$  values lower than 0.06, implying that these nine new chimerical retrogenes formed in the last 22.5 million years ( $\sim 0.4$  chimerical retrogene per million years). The rate of formation of new chimerical retrogenes in dogs was more than three times higher than that in humans ( $\sim 0.14$  chimerical retrogenes per million years) while lower than the proportion of new chimerical retrogene in zebrafish

(Fu et al., 2010) and rice (Wang et al., 2006). These results also indicated that the formation of dog chimerical retrogenes occurred at a remarkably rapid evolutionary rate.

### 3.3 | The functionality of retrogenes

We applied a two-step strategy to survey the functionality of retrocopies: (a) to compare nonsynonymous ( $K_a$ ) and synonymous ( $K_s$ ) substitution rates ( $\omega$  ratios) between retrogenes and their parental genes; (b) to estimate the transcriptional ability of retrocopies.

First, we tested the functional constraints at the protein level between parental genes and retrocopies by using the nonsynonymous to synonymous substitution rate ( $\omega = K_a/K_s$ ). The values  $\omega < 1$ ,  $\omega = 1$ , and  $\omega > 1$  represent purifying selection, neutral evolution, and positive selection, respectively (Yang & Nielsen, 2002). We used a stricter criterion ( $\omega < 0.5$ ) to estimate the functionality of retrogenes (Emerson, Kaessmann, Betran, & Long, 2004). After removing 46 retrocopies with  $K_s = 0$  ( $\omega$  value is unable to evaluate), we obtained 470 intact retrogenes and 2,509 retropseudogenes for further analysis. Obviously, intact retrogenes significantly showed much lower  $\omega$  ratio values than that in retropseudogenes (Figure 6a,  $p < .01$ ). About 60.6% (285/470) of intact retrogenes had lower  $\omega$  than 0.5, compared with 39.2% (984/2,509) of retropseudogenes (Table 2). Thus, these results indicated that more than half of intact retrogenes were under stronger purifying selection and tended to be subjected to more functional constraints when compared to retropseudogenes. This revealed a similar proportion of functional constraint in pig retrocopies ( $\sim 64.7\%$  of intact retrogenes and  $\sim 39.6\%$  of retropseudogenes were subjected to functional constraint; Fang et al., 2018), suggesting that intact retrogenes were more likely to evolve into functional genes and subsequently be reserved by natural/artificial selection in mammals during a long evolutionary history. In addition, we detected about 25 retrogenes with  $\omega$  ratios significantly higher than 1, suggesting that they were likely reserved by positive selection and evolved into novel functions. New genes have been considered to be an important driver to phenotypic evolution innovation (Kaessmann, 2010). They evolve important genetic novelties to facilitate biological diversity and contribute to the evolution



**FIGURE 6** Functional analysis of retrogenes. (a), distribution of  $\omega$  ratio between intact retrogenes and retropseudogenes. Retrocopies with  $\omega > 1.5$  were pooled in a single bin. (b), Transcription ability of intact retrogenes and retropseudogenes supported by ESTs

**TABLE 2** Selective pressure of retrocopies

Gene	$\omega < 0.5$	$\omega = 0.5-1.2$	$\omega > 1.2$
Intact gene	285/470 (60.6%)	160/470 (34.1%)	25/470 (5.3%)
Retropseudogene	984/2,509 (39.2%)	1,419/2,509 (56.6%)	106/2,509 (4.2%)

Note: Intact gene: excluding 14 retrocopies with  $K_s = 0$  and  $\omega = \text{NA}$ ; retropseudogene: excluding 1 retrocopy with  $K_s = 0$  and  $\omega = \text{NA}$ .

**TABLE 3** ESTs support of retrocopies

Raw ESTs	Clean ESTs	Uniq ESTs	Intact	Retropseudo
382,638	375,093	242,251	165/507	126/2,518

Note: 165/507 indicates that 165 intact retrogenes were supported by at least one.

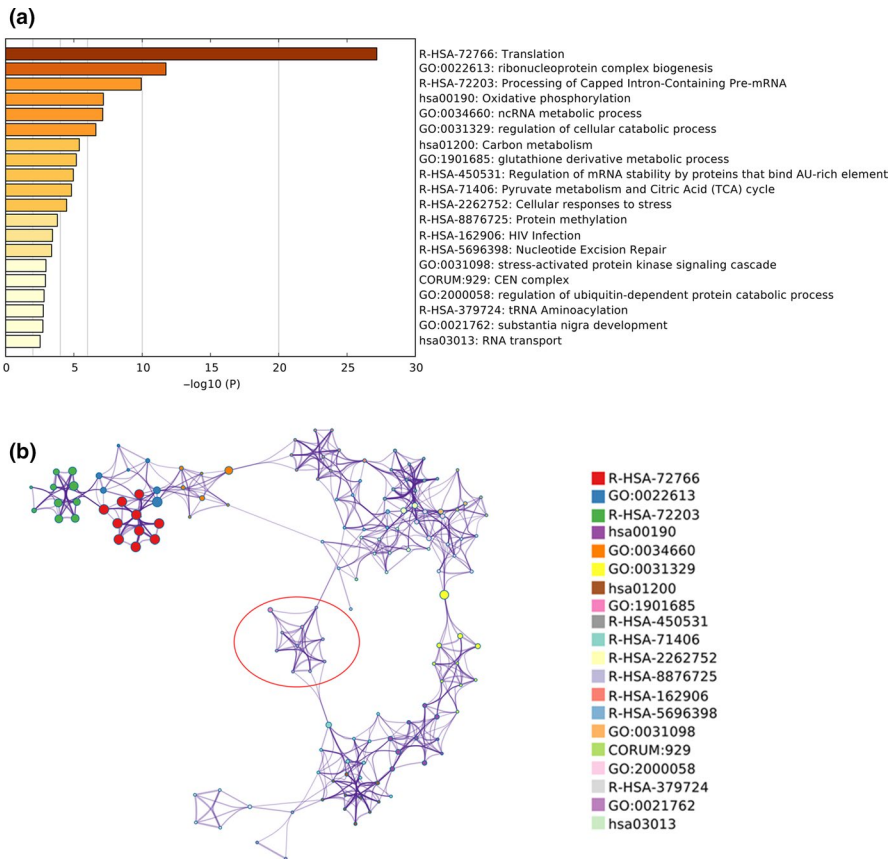
of a lineage or species-specific phenotypic traits (Chen et al., 2013, 2010; Long et al., 2013). Thus, dog retrogenes may play important roles in promoting adaptive evolution and species differentiation. Interestingly, a portion of retropseudogenes was under purifying selection or positive selection as well, suggesting that they were either function in splicing out the disabled mutations to produce novel proteins or had been functional but recently became pseudogenes.

Moreover, transcription ability is another valuable way to validate gene functions. Considering that a retrogene shared high identity with its corresponding parental gene in coding regions, we used a more credible ESTs data to estimate the transcription ability instead of short-tag expression sequences or hybridization-based data (Fu et al., 2010; Harbers & Carninci, 2005). First, a total of 382,638 dog ESTs were downloaded from NCBI and processed for various contaminants such as vectors, adaptors, primers, or linker sequences using the SeqClean program. Then, about 375,093 clean ESTs were mapped onto the dog genome using BLAT software with default parameters. Only 242,251 high identity ESTs that aligned to a unique location in the dog genome were retained for further analyses. By analyzing the overlap region between retrocopy and ESTs, we found that intact retrogenes significantly exhibited much higher transcription ability than retropseudogenes (Figure 6b,  $p < .01$ ). About 32.5% (165/507) of intact retrogenes were covered by at least one EST compared with only 5.0% (126/2,518) in retropseudogenes (Table 3). In the human genome, the proportion of expressed intact retrogenes

has been reported about ~30.1% (Vinckenbosch, Dupanloup, & Kaessmann, 2006), indicating that a large proportion of retrocopies are transcribed and functionally reserved in mammals. In addition, 19 of 126 retropseudogenes were covered by more than four ESTs. Reading frameshifts were considered as one of the crucial molecular processes that could generate new gene structures. About 470 human gene duplicates had subjected to frameshift mutations and generate new coding sequences (Okamura, Feuk, Marques-Bonet, Navarro, & Scherer, 2006). Thus, these 19 expressed retropseudogenes were likely functional transcript and performed reading frameshifts in a protein-coding gene to generate novel proteins. This result supported the previous' result that a portion of retropseudogenes was subjected to stringent functional constraints.

### 3.4 | New retrogenes improved dogs' adaptive evolution

Origin of new genes provides critical genetic novelties for biological diversity and contribute to the evolution of lineage- or species-specific phenotypic traits (Betran, Wang, Jin, & Long, 2002; Chen et al., 2013, 2010). As a domestic species, the dog is a part of the largest groups of phenotypic diversity (Beck, 2002). We performed functional enrichment analyses of 207 new retrogenes ( $K_s < 0.06$  and  $\omega < 0.5$ ) to investigate their contributions to dogs' adaptive evolution. Top 20 clusters with their representative enriched terms from Metascape pathway enrichment were performed with the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, and CORUMA (Figure 7a). The most significant GO terms mainly consisted of biological processes which were involved in ribonucleoprotein complex biogenesis, ncRNA metabolic process, regulation of cellular catabolic process, and glutathione derivative metabolic process.



**FIGURE 7** GO and pathway enrichment analysis of new retrogenes. (a), Heatmap of enriched terms of new retrogenes. Only top 20 enrichment with  $p < .01$  are listed. (b) Network of enriched terms by cluster ID, where nodes that share the same cluster ID are typically close to each other. Glutathione biosynthetic/metabolic process and response to toxic substance is enriched in the red circle

Moreover, pathways enrichment analysis was mainly involved in translation, oxidative phosphorylation, carbon metabolism, and RNA transport. Interestingly, we found that some new retrogenes were significantly enriched in glutathione biosynthetic/metabolic process (GSTA4, GSTP1, MGST3, AKR1A1, GPX8, and EEF1G) and response to toxic substances (ACTB, CCNB1, GOT2, GSTP1, HNRNPA1, MGST3, RPL10A, TXNL1, and GPX8; Figure 7b). For example, glutathione S-transferases (GSTs, including GSTA4, GSTP1, and MGST3) was defined as the most important intracellular non-enzymatic antioxidant for their ability to catalyze the conjugation of the reduced form of glutathione (GSH) to xenobiotic substrates. GSTs exist extensively in plants and animals (Pompella, Visvikis, Paolicchi, De Tata, & Casini, 2003; Sies, 1999) and play critical roles in preventing damage to important cellular components that are caused by reactive oxygen species such as free radicals, peroxides, methylglyoxal, and lipid peroxides (Aoyama & Nakaki, 2015). It also effects in the detoxification of a variety of xenobiotics such as chemical carcinogens, environmental pollutants, and antitumor agents, which interact with glutathione and are ultimately excreted in the urine or feces in the form of mercapturic acids (Hernandez et al., 2015). During the early phase of the agricultural revolution, humans changed from a nomadic lifestyle to a sedentary lifestyle and ancestral wolves may have been attracted by dumps and spoiled food sources near early human settlements. Deterioration of food, especially lipids and proteins in the meat, could give rise to infectious organisms or toxic products such as bacteria, viruses, and lipid peroxidation. Thus, new genes that involved in glutathione

biosynthetic/metabolic process and response to toxic substances pathways were likely helpful for wolves/dogs to detoxify environmental pollutants in intestinal. These genetic novelties allowed wolves/dogs groups for efficient use of dumped food by human and probably contributed to their adaptation to early human settlements. This is one possibility why ancestors of wolves/dogs were kept by the human.

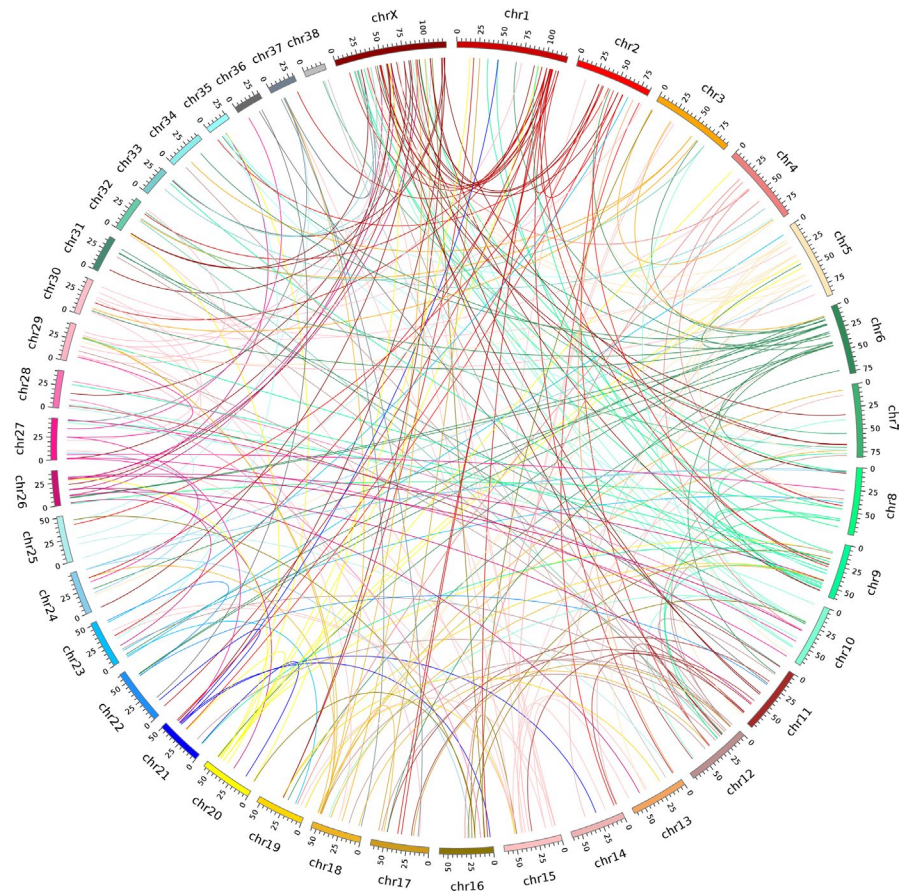
In addition, we checked the transcription ability of these new retrogenes that significantly enriched in glutathione biosynthetic/metabolic process and response to the toxic substance. These 12 parental genes totally produced about 95 retrocopies (Table S2), of which 13 were defined as functional new retrogenes (with  $K_s < 0.06$  and  $\omega < 0.5$ ). Moreover, eight functional new retrogenes were supported by at least one ESTs, suggesting that they were likely expressed and contributed to detoxify environmental pollutants in intestinal.

### 3.5 | Extensive gene traffic on the X chromosome

Mammalian sex chromosomes (X and Y) change profoundly in their differentiation from ancestral autosomes. Throughout this process, the selective relocation of new genes can be driven by gene duplication (Emerson et al., 2004; Vibranovski, Zhang, & Long, 2009). Retrotransposition reshapes genome through the reverse transcription of mRNA and insertion of DNA into a new locus. Due to the distinguishing features between a retrocopy and its ancestral parental gene, it is feasible to identify the direction of the RNA-duplicated



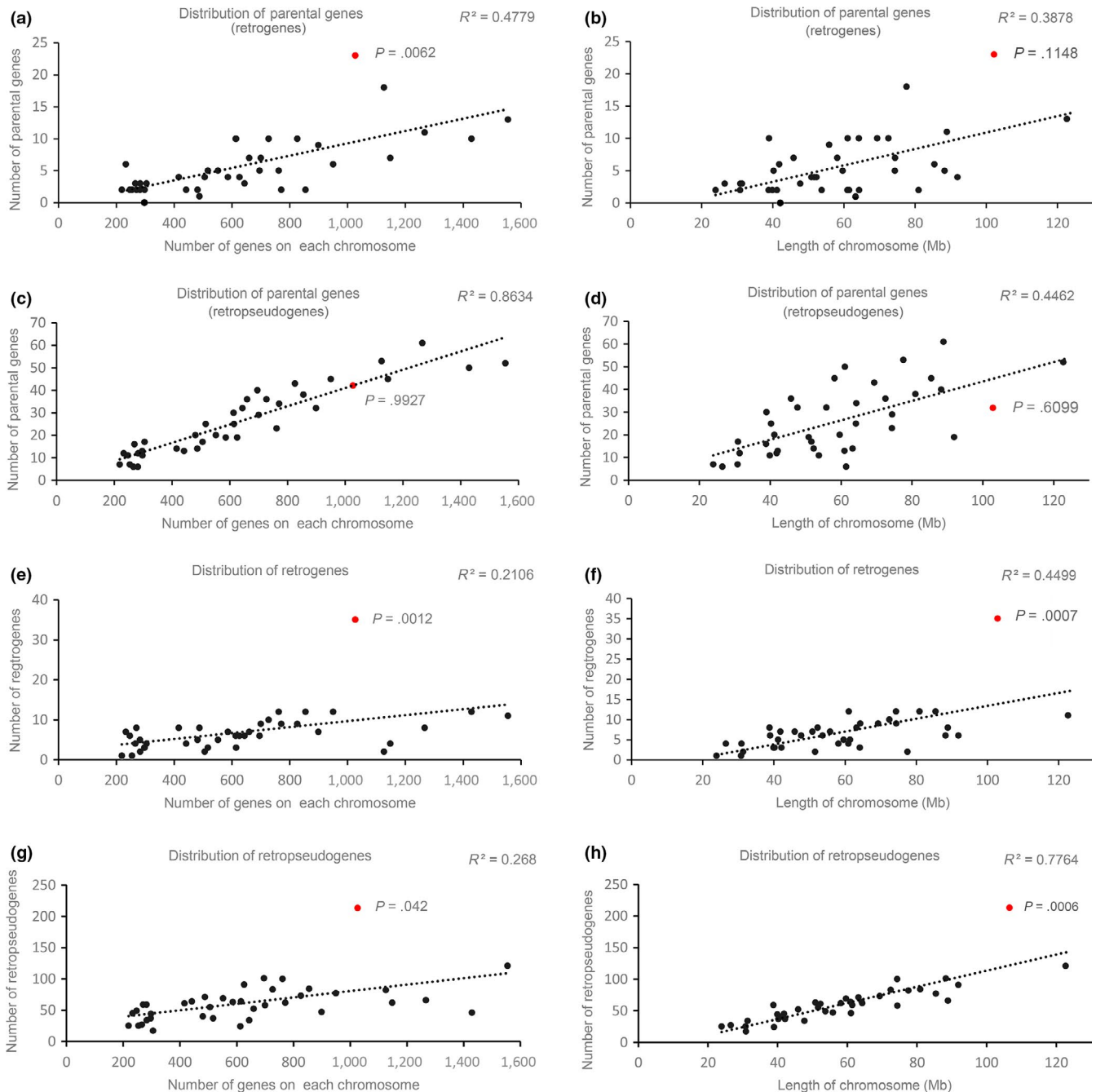
**FIGURE 8** Gene interchromosomal movement. Different color bars represent different chromosomes. The line indicates gene interchromosomal movement, and its color indicate the movement direction



gene in chromosomal movement. To elucidate gene movements and chromosome bias in the dog genomes, we analyzed the relocation distribution of retrogenes and their parental genes. By screening 3,025 retrotransposition events, 2,902, 77, and 46 of which were involved into interchromosomal, intrachromosomal, and scaffold/MT movements, respectively (corresponding to 95.93%, 2.55%, and 1.52%, respectively, Table S1), suggesting that retrocopies were inclined to distribute in new chromatin environments.

First, we estimated the relationship between functional retrogenes' movement bias and chromosomal characteristics (including the number of genes on each chromosome and the size of chromosomal). To avoid overestimating the number of functional retrogenes, we performed a stringent functionality criterion based on both selective constraints, with  $\omega < 0.5$ , and conserved ORF region without frameshift mutations or premature stop codons. In total, we identified 272 functional retrogene pairs where the retrogenes did not share chromosome linkage with their parental copy (Figure 8). Analysis of the chromosomal location of their parental genes showed that the numbers of parental genes on each chromosome were positively associated with both the number of genes ( $r = .69$ , Figure 9a) and the size of chromosomes ( $r = .62$ , Figure 9b). One possibility was the case that the longer chromosomes usually contained more genes, which provided greater opportunities for being involved into retrotransposition event. However, as the only outlier, the X chromosome exhibited a significant excess of parental genes when compared with the autosomes (red dot,  $p = .0062$ , Figure 9a), indicating

that parental genes escaped from X chromosome to autosomes. Previous study indicated that retrotransposition events occurred randomly between chromosomes (Zhang, Harrison, & Gerstein, 2002). Thus, we tested whether this asymmetric pattern of gene escaping event was consistent for nonfunctional retroseudogenes? To survey occasionality of abnormal movement out of the X chromosome into autosomes, we analyzed neutral patterns of parental gene distribution by using their nonfunctional retroseudogenes ( $0.5 < \omega < 1.0$ , and with frameshift mutations or premature stop codons). A total of 1,268 retroseudogenes and parental genes pairs, which were involved in interchromosomal movements, were used to study the relationship between parental gene bias and chromosomal characteristics. Consistently, it was observed that the numbers of retroseudogenes' parental genes on each chromosome were positively associated with both the number of genes and the length of chromosome ( $r = .93$ , Figure 9c and  $r = .67$ , Figure 9d) and the X chromosome did not show significant different gene movements (red dot,  $p = .9927$  and  $p = .6098$ ). These results indicated that an excess of escaping event from the X chromosome to the autosomes was a widespread mechanism of how functional RNA-duplicated genes originated and survived. In agreement with previous studies in human, mouse, and *Drosophila* (Betran, Thornton, et al., 2002; Emerson et al., 2004), the excess of parental genes escaping from the X chromosome to the autosomes may be explained by natural selection or sexual antagonism (Charlesworth, Coyne, & Barton, 1987; Engelstadter & Haig, 2008; Vibranovski et al., 2009; Wu & Xu,



**FIGURE 9** Analysis of retrocopies relocations. We estimate the relationship between genes movement bias and chromosomal characteristics (including the number of genes on each chromosome and the size of chromosomal). Regressions for the parental genes of retrogenes (a, b) and retropseudogenes (c, d). Regressions for retrogenes (e, f) and retropseudogenes (g, h). Red point represents X chromosome, and  $p < .01$  indicates a significant excess movement out of/into X chromosome.  $R$  value represents the related coefficient

2003). In the male germline, expression of parental genes located on inactivation X chromosome will be inhibited. By producing retrogenes in autosomes, the male-biased genes avoid X-linkage inactivation and become expressed genes in testis. Retrogenes that escaped from the X chromosome have a higher potential to be expressed in testis and probably evolve into new genes beneficial for males. In the sexual antagonism hypothesis, X-inactivation has been supposed to attain equal gene expression between males and females (dosage compensation; Charlesworth, 1978; Engelstadter & Haig,

2008). And these genes which were beneficial for males but costly for females are probably to be fixed on the autosomes. In addition, retrogenes that originated from the X chromosome usually exhibit testis expression patterns in *Drosophila* (Bai, Casola, Feschotte, & Betran, 2007; Betran, Thornton, et al., 2002), confirming that the fixation of male-biased beneficial mutations on the autosomes is a long natural selection bias.

Second, we surveyed the distribution of retrocopies on the chromosome to test functional/nonfunctional retrocopies' insertion bias. For

functional retrogenes, the number of retrocopies was positively correlated with the length of each chromosome ( $r = .67$ , Figures 8f and 9e). Consistent with the relocation of parental genes, the X chromosome was an outlier with an excess of retrogenes recruitment from autosomes (red dot,  $p = .0007$ , Figure 9f). For nonfunctional retropseudogenes, a similar relocation bias has been observed in the X chromosome (red dot,  $r = .88$ ,  $p = .0006$ , Figure 9h). These results implied that the bias in disproportionate insertions of both functional and nonfunctional retrocopies existed for the X chromosome. One possibility was that the X chromosome was enriched twofold for L1 repetitive elements (Bailey, Carrel, Chakravarti, & Eichler, 2000), which probably contributed to a bias in extra insertions or fewer deletions of retrocopies.

## 4 | CONCLUSION

In this study, we identified a large number of new retrogenes and chimerical retrogenes originated within ~22 Mya. By analyzing selective pressure along with ESTs expression, most of the intact retrogenes were significantly under stronger purifying selection and subjected to the more functional constraint when compared to retropseudogenes. Interestingly, new retrogenes likely provided important genetic basis for dogs' adaptation to scavenge human waste dumps. Furthermore, we also found a biased movement of functional retrogenes from the X to A chromosomes and A to X chromosomes, which were likely driven by natural selection or sexual antagonistic. Together, our analysis demonstrates that retrotransposition is an important mechanism that can reshape the dog genome and probably contribute to its adaptive evolution.

## ACKNOWLEDGMENTS

This work was supported by the Hubei Provincial Natural Science Foundation of China (No. 2015CFB350). We also thank the staff and faculty at the College of Animal Science and Veterinary Medicine at Huazhong Agricultural University for supporting computing platform, and Ph.D Dustin Kenneth Goncharoff (Northwestern University Feinberg School of Medicine) to improve the writing.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## AUTHOR CONTRIBUTIONS

X.G. and Y.L. performed the data analysis and wrote the first draft of the manuscript; A.A.A., Y.W., and H.C. reviewed and edited the manuscript; X.G., Y.W., and H.C. developed the original idea and supervised the study.

## DATA AVAILABILITY STATEMENT

All relevant data are within the paper and its Supporting Information files (Tables S1 and S2).

## ORCID

Xiang Gao  <https://orcid.org/0000-0002-3667-8900>

Adeyinka A. Adetula  <https://orcid.org/0000-0002-1865-0007>

## REFERENCES

- Akram, M., Jamil, M. S., Mehmood, Z., Akram, M., Waqas, M. K., Iqbal, Z., ... Asif, H. M. (2011). Fast alignment (FASTA): A review article. *Journal of Medicinal Plants Research*, 5, 6931–6933.
- Aoyama, K., & Nakaki, T. (2015). Glutathione in cellular redox homeostasis: Association with the excitatory amino acid carrier 1 (EAAC1). *Molecules*, 20, 8742–8758. <https://doi.org/10.3390/molecules20058742>
- Bai, Y. S., Casola, C., Feschotte, C., & Betran, E. (2007). Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biology*, 8, R11.
- Bailey, J. A., Carrel, L., Chakravarti, A., & Eichler, E. E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 6634–6639. <https://doi.org/10.1073/pnas.97.12.6634>
- Beck, A. M. (2002). Dogs: A startling new understanding of canine origin, behavior, and evolution. *Society & Animals*, 10, 217–219.
- Berryere, T. G., Kerns, J. A., Barsh, G. S., & Schmutz, S. M. (2005). Association of an Agouti allele with fawn or sable coat color in domestic dogs. *Mammalian Genome*, 16, 262–272. <https://doi.org/10.1007/s00335-004-2445-6>
- Betran, E., Thornton, K., & Long, M. (2002). Retroposed new genes out of the X in *Drosophila*. *Genome Research*, 12, 1854–1859. <https://doi.org/10.1101/gr.6049>
- Betran, E., Wang, W., Jin, L., & Long, M. (2002). Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Molecular Biology and Evolution*, 19, 654–663. <https://doi.org/10.1093/oxfordjournals.molbev.a004124>
- Botigue, L. R., Song, S., Scheu, A., Gopalan, S., Pendleton, A. L., Oetjens, M., ... Veeramah, K. R. (2017). Ancient European dog genomes reveal continuity since the Early Neolithic. *Nature Communications*, 8, 16082. <https://doi.org/10.1038/ncomms16082>
- Candille, S. I., Kaelin, C. B., Cattanaach, B. M., Yu, B., Thompson, D. A., Nix, M. A., ... Barsh, G. S. (2007). A beta-defensin mutation causes black coat color in domestic dogs. *Science*, 318, 1418–1423. <https://doi.org/10.1126/science.1147880>
- Charlesworth, B. (1978). Model for evolution of Y-chromosomes and dosage compensation. *Proceedings of the National Academy of Sciences of the United States of America*, 75, 5618–5622. <https://doi.org/10.1073/pnas.75.11.5618>
- Charlesworth, B., Coyne, J. A., & Barton, N. H. (1987). The relative rates of evolution of sex-chromosomes and autosomes. *The American Naturalist*, 130, 113–146. <https://doi.org/10.1086/284701>
- Chen, S. D., Krinsky, B. H., & Long, M. Y. (2013). New genes as drivers of phenotypic evolution. *Nature Reviews Genetics*, 14, 645–660. <https://doi.org/10.1038/nrg3521>
- Chen, S. D., Ni, X. C., Krinsky, B. H., Zhang, Y. E., Vibranovski, M. D., White, K. P., & Long, M. Y. (2012). Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. *The EMBO Journal*, 31, 2798–2809. <https://doi.org/10.1038/emboj.2012.108>
- Chen, S. D., Zhang, Y. E., & Long, M. Y. (2010). New genes in *Drosophila* quickly become essential. *Science*, 330, 1682–1685. <https://doi.org/10.1126/science.1196380>
- Clark, L. A., Wahl, J. M., Rees, C. A., & Murphy, K. E. (2006). Retrotransposon insertion in SILV is responsible for merle patterning



- of the domestic dog. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 1376–1381.
- Dai, H. Z., Chen, Y., Chen, S. D., Mao, Q. Y., Kennedy, D., Landback, P., ... Long, M. Y. (2008). The evolution of courtship behaviors through the origination of a new gene in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 7478–7483. <https://doi.org/10.1073/pnas.0800693105>
- Emerson, J. J., Kaessmann, H., Betran, E., & Long, M. (2004). Extensive gene traffic on the mammalian X chromosome. *Science*, 303, 537–540. <https://doi.org/10.1126/science.1090042>
- Engelstadter, J., & Haig, D. (2008). Sexual antagonism and the evolution of X chromosome inactivation. *Evolution*, 62, 2097–2104. <https://doi.org/10.1111/j.1558-5646.2008.00431.x>
- Esnault, C., Maestre, J., & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, 24, 363–367. <https://doi.org/10.1038/74184>
- Fang, C., Zou, C., Fu, Y., Li, J., Li, Y., Ma, Y., ... Li, C. (2018). DNA methylation changes and evolution of RNA-based duplication in *Sus scrofa*: Based on a two-step strategy. *Epigenomics*, 10, 199–218.
- Frantz, L. A., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., ... Larson, G. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352, 1228–1231. <https://doi.org/10.1126/science.aaf3161>
- Fu, B., Chen, M., Zou, M., Long, M., & He, S. (2010). The rapid generation of chimerical genes expanding protein diversity in zebrafish. *BMC Genomics*, 11, 657. <https://doi.org/10.1186/1471-2164-11-657>
- Gogvadze, E. V., & Buzdin, A. A. (2005). New mechanism of retrogene formation in mammalian genomes: In vivo recombination during RNA reverse transcription. *Molekuljarna Biologija*, 39, 364–373.
- Gopalakrishnan, S., Samaniego Castruita, J. A., Sinding, M. S., Kuderna, L. F. K., Raikkonen, J., Petersen, B., ... Gilbert, M. T. P. (2017). The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. *BMC Genomics*, 18, 495. <https://doi.org/10.1186/s12864-017-3883-3>
- Harbers, M., & Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, 2, 495–502. <https://doi.org/10.1038/nmeth768>
- Hernandez, L. E., Sobrino-Plata, J., Montero-Palmero, M. B., Carrasco-Gil, S., Flores-Caceres, M. L., Ortega-Villasante, C., & Escobar, C. (2015). Contribution of glutathione to the control of cellular redox homeostasis under toxic metal and metalloid stress. *Journal of Experimental Botany*, 66, 2901–2911. <https://doi.org/10.1093/jxb/erv063>
- Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., ... Yoshiki, A. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, 423, 91–96. <https://doi.org/10.1038/nature01535>
- Kabza, M., Ciomborowska, J., & Makalowska, I. (2014). RetrogeneDB—A database of animal retrogenes. *Molecular Biology and Evolution*, 31, 1646–1648. <https://doi.org/10.1093/molbev/msu139>
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research*, 20, 1313–1326. <https://doi.org/10.1101/gr.101386.109>
- Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H., Zody, M. C., Anderson, N., ... Lindblad-Toh, K. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics*, 39, 1321–1328. <https://doi.org/10.1038/ng.2007.10>
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... Marra, M. A. (2009). Circo: An information aesthetic for comparative genomics. *Genome Research*, 19, 1639–1645. <https://doi.org/10.1101/gr.092759.109>
- Kukekova, A. V., Johnson, J. L., Xiang, X., Feng, S., Liu, S., Rando, H. M., ... Zhang, G. (2018). Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. *Nature Ecology & Evolution*, 2, 1479–1491. <https://doi.org/10.1038/s41559-018-0611-6>
- Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., ... Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438, 803–819. <https://doi.org/10.1038/nature04338>
- Long, M., Betran, E., Thornton, K., & Wang, W. (2003). The origin of new genes: Glimpses from the young and old. *Nature Reviews Genetics*, 4, 865–875. <https://doi.org/10.1038/nrg1204>
- Long, M., & Langley, C. H. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science*, 260, 91–95.
- Long, M. Y., VanKuren, N. W., Chen, S. D., & Vibranovski, M. D. (2013). New gene evolution: Little did we know. *Annual Review of Genetics*, 47(47), 307–333. <https://doi.org/10.1146/annurev-genet-111212-133301>
- Malmstrom, H., Vila, C., Gilbert, M. T., Stora, J., Willerslev, E., Holmlund, G., & Gotherstrom, A. (2008). Barking up the wrong tree: Modern northern European dogs fail to explain their origin. *BMC Evolutionary Biology*, 8, 71. <https://doi.org/10.1186/1471-2148-8-71>
- Mamedov, I. Z., Arzumanyan, E. S., Amosova, A. L., Lebedev, Y. B., & Sverdlov, E. D. (2005). Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Research*, 33, e16. <https://doi.org/10.1093/nar/gni018>
- Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., & Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. *PLoS Biology*, 3, e357. <https://doi.org/10.1371/journal.pbio.0030357>
- Moss, T., Langlois, F., Gagnon-Kugler, T., & Stefanovsky, V. (2007). A housekeeper with power of attorney: The rRNA genes in ribosome biogenesis. *Cellular and Molecular Life Sciences*, 64, 29–49. <https://doi.org/10.1007/s00018-006-6278-1>
- Okamura, K., Feuk, L., Marques-Bonet, T., Navarro, A., & Scherer, S. W. (2006). Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics*, 88, 690–697. <https://doi.org/10.1016/j.ygeno.2006.06.009>
- Pan, D., & Zhang, L. Q. (2009). Burst of young retrogenes and independent retrogene formation in mammals. *PLoS ONE*, 4, e5040. <https://doi.org/10.1371/journal.pone.0005040>
- Parker, H. G., VonHoldt, B. M., Quignon, P., Margulies, E. H., Shao, S., Mosher, D. S., ... Ostrander, E. A. (2009). An expressed *Fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science*, 325, 995–998. <https://doi.org/10.1126/science.1173275>
- Percharde, M., Lin, C. J., Yin, Y., Guan, J., Peixoto, G. A., Bulut-Karslioglu, A., ... Ramalho-Santos, M. (2018). A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell*, 174, 391–405.e19. <https://doi.org/10.1016/j.cell.2018.05.043>
- Podlaha, O., & Zhang, J. Z. (2004). Nonneutral evolution of the transcribed pseudogene *Makorin1-p1* in mice. *Molecular Biology and Evolution*, 21, 2202–2209. <https://doi.org/10.1093/molbev/msh230>
- Pompella, A., Visvikis, A., Paolicchi, A., De Tata, V., & Casini, A. F. (2003). The changing faces of glutathione, a cellular protagonist. *Biochemical Pharmacology*, 66, 1499–1503. [https://doi.org/10.1016/S0006-2952\(03\)00504-5](https://doi.org/10.1016/S0006-2952(03)00504-5)
- Rosso, L., Marques, A. C., Weier, M., Lambert, N., Lambot, M. A., Vanderhaeghen, P., & Kaessmann, H. (2008). Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. *PLoS Biology*, 6, e140. <https://doi.org/10.1371/journal.pbio.0060140>
- Schmid, C. W. (1998). Does SINE evolution preclude Alu function? *Nucleic Acids Research*, 26, 4541–4550. <https://doi.org/10.1093/nar/26.20.4541>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2498–2504. <https://doi.org/10.1101/gr.1239303>



- Sies, H. (1999). Glutathione and its role in cellular functions. *Free Radical Biology and Medicine*, 27, 916–921. [https://doi.org/10.1016/S0891-5849\(99\)00177-X](https://doi.org/10.1016/S0891-5849(99)00177-X)
- Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalen, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, 25, 1515–1519. <https://doi.org/10.1016/j.cub.2015.04.019>
- Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., ... Ostrander, E. A. (2007). A single IGF1 allele is a major determinant of small size in dogs. *Science*, 316, 112–115. <https://doi.org/10.1126/science.1137045>
- Takeuchi, Y., Kaneko, F., Hashizume, C., Masuda, K., Ogata, N., Maki, T., ... Mori, Y. (2009). Association analysis between canine behavioural traits and genetic polymorphisms in the Shiba Inu breed. *Animal Genetics*, 40, 616–622. <https://doi.org/10.1111/j.1365-2052.2009.01888.x>
- Vibrantovski, M. D., Zhang, Y., & Long, M. (2009). General gene movement off the X chromosome in the *Drosophila* genus. *Genome Research*, 19, 897–903. <https://doi.org/10.1101/gr.088609.108>
- Vinckenbosch, N., Dupanloup, I., & Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 3220–3225. <https://doi.org/10.1073/pnas.0511307103>
- Wang, W., & Kirkness, E. F. (2005). Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research*, 15, 1798–1808. <https://doi.org/10.1101/gr.3765505>
- Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., ... Wang, J. (2006). High rate of chimeric gene origination by retroposition in plant genomes. *The Plant Cell*, 18, 1791–1802. <https://doi.org/10.1105/tpc.106.041905>
- Wu, C. I., & Xu, E. Y. (2003). Sexual antagonism and X inactivation - the SAXI hypothesis. *Trends in Genetics*, 19, 243–247. [https://doi.org/10.1016/S0168-9525\(03\)00058-1](https://doi.org/10.1016/S0168-9525(03)00058-1)
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19, 908–917. <https://doi.org/10.1093/oxfordjournals.molbev.a004148>
- Zhang, J., Yang, H., Long, M., Li, L., & Dean, A. M. (2010). Evolution of enzymatic activities of testis-specific short-chain dehydrogenase/reductase in *Drosophila*. *Journal of Molecular Evolution*, 71, 241–249. <https://doi.org/10.1007/s00239-010-9384-5>
- Zhang, Z., Harrison, P., & Gerstein, M. (2002). Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Research*, 12, 1466–1482. <https://doi.org/10.1101/gr.331902>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Gao X, Li Y, Adetula AA, Wu Y, Chen H. Analysis of new retrogenes provides insight into dog adaptive evolution. *Ecol Evol*. 2019;9:11185–11197. <https://doi.org/10.1002/ece3.5620>