Tracing the origin and evolutionary fate of recent gene retrocopies in natural populations of the house mouse

Wenyu Zhang* and Diethard Tautz

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, D-24306 Plön, Germany

*Correspondence to
Wenyu Zhang
Department of Evolutionary Genetics
Max Planck Institute for Evolutionary Biology
August-Thienemann-Str. 2, D-24306 Plön, Germany
Phone: +49 4522-763-372
Fax: +49 4522-763-281
Email: wyzhang@evolbio.mpg.de

Running title: Population dynamics of gene retrocopies

Key words: Gene retroposition, house mouse, natural populations, adaptation, population divergence

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract

While the contribution of retrogens to the evolution of genes and genomes has long been recognized, the evolutionary patterns of very recently derived retrocopies that are still polymorphic within natural populations have not been much studied so far. We use here a set of 2,025 such retrocopies in nine house mouse populations from three subspecies (*Mus musculus domesticus*, *Mus musculus musculus*, and *Mus musculus castaneus*) to trace their origin and evolutionary fate. We find that ancient house-keeping genes are significantly more likely to generate retrocopies than younger genes and that the propensity to generate a retrocopy depends on its level of expression in the germline. While most retrocopies are detrimental and quickly purged, we focus here on the subset that appears to be neutral or even adaptive. We show that retrocopies from X-chromosomal parental genes have a higher likelihood to reach elevated frequencies in the populations, confirming the notion of adaptive effects for “out-of-X” retrogens. Also, retrocopies in intergenic regions are more likely to reach higher population frequencies than those in introns of genes, implying a more detrimental effect when they land within transcribed regions. For a small subset of retrocopies, we find signatures of positive selection, indicating they were involved in a recent adaptation process. We show that the population-specific distribution pattern of retrocopies is phylogenetically informative and can be used to infer population history with a better resolution than with SNP markers.

Introduction

Gene duplication is a major mechanism to generate adaptive evolutionary novelties in organisms (Ohno 1970; Long, et al. 2003). Gene retroposition (or RNA-based gene duplication) is a particular type of gene duplication in which a gene’s transcript is used as a template to generate new gene retrocopies, and this has a variety of evolutionary implications (Kaessmann, et al. 2009; Casola and Betran 2017). Despite of inherent lack of the ancestral regulatory sequences, abundant previous between-species comparison studies have found a substantial number of new gene retrocopies acting as functional retrogens that encode full-length proteins across organisms (Bai, et al. 2007; Baertsch, et al. 2008; Kaessmann, et al. 2009; Casola and Betran 2017), via recruiting regulatory elements from the vicinity or evolving them de novo. Also, given that the majority of genomic regions are accessible to transcription (Neme and Tautz 2016), many of the inactivated
gene retrocopies with truncation (i.e., processed pseudogenes) could also be transcribed, and thus their transcripts could influence the expression and function of the parental genes (Tam, et al. 2008; Marques, et al. 2011). In addition, the distinct patterns on the extensive gene traffic on the X chromosome indicate that the fixation of most of the functional retrogenes is driven by positive selection (Betran, et al. 2002; Emerson, et al. 2004). While these between-species comparison studies have provided important insights into the functional evolution of gene retrocopies at a long-term evolutionary time scale, the evolutionary dynamic and trajectory of newly arrived gene retrocopies, especially at the very early stage when they are still polymorphic within natural populations, remain largely unexplored.

The accumulating population genomic sequencing data allowed to study polymorphic gene retrocopies (retroCNV alleles or retroCNVs) at individual genome level within natural populations. Through analyzing the genomic sequencing dataset of natural samples from the 1,000 Genomes Project Consortium (Genomes Project, et al. 2012), several studies have expanded the gene retrocopy repository in human populations, and revealed substantial polymorphisms of retroCNVs among human individuals (Abyzov, et al. 2013; Ewing, et al. 2013; Schrider, et al. 2013; Kabza, et al. 2015; Zhang, et al. 2017). However, studies that explore the retrocopy dynamics in the context of natural populations from different subspecies are still lacking. We analyze here the evolutionary patterns of gene retrocopies at individual genome level in house mouse populations, based on a set of well-defined natural populations from different lineages and subspecies where evolutionary processes and retroposition rates can be analyzed at a much broader comparative level than it is possible for humans.

Owing to its well-defined evolutionary history, as well as a set of largely homogeneous genomic resources from wild mice collected from nine world-wide natural populations and two out-group species (Harr, et al. 2016), the house mouse (Mus musculus) represents a unique model system for the study of evolutionary dynamics of recently originated gene retrocopies. The inclusion of the genomic sequencing data from out-group species allows the detection of newly arrived species-specific gene retrocopies, while comparable population data of out-groups (e.g., Neandertals or Denisovans) is not available in human studies. Using this unique genomic dataset, we identified previously more than 2,000 house mouse lineage-specific gene retrocopies, and showed that the primary rate of gene retroposition in house mouse natural populations is orders of magnitude higher than the previously estimated long-term rate (Zhang, et al. 2021). Our analysis
showed that the majority of them have deleterious effects, not only through disrupting functional
genomic regions, but also by virtue of their expression that can interfere with their parental genes,
especially when they are transcribed in antisense direction. Given the advantageous nature of the
sequencing read features (e.g., sequencing coverage, read length, and insert size), our previous
study revealed a much higher detection rate of insertion sites of gene retrocopies than was
reported in human studies (Zhang, et al. 2021). Therefore, this made it possible to use this dataset
to directly trace the origin and evolutionary dynamics of very recently derived gene retrocopies
and their potential contribution to the adaptation process in natural populations.

We mainly focus on the possible adaptive effects of new gene retrocopies in the present report.
We show that evolutionarily old house-keeping genes are the major source of gene retroposition
to generate new gene retrocopies in house mouse wild individuals. Chromosomal movement
analysis reveals a step-wise emergence of the “out-of-X” pattern of gene retroposition, suggesting
that the action of positive selection on a subset of new gene retrocopies can result in an increase
of their frequencies in populations. Positive selection signal detection analysis further shows the
potentially direct action of positive selection on a subset of gene retrocopies that are private in
single subspecies, hinting these gene retrocopies might play a role in the adaptation process of the
respective populations. Given that the results show that retrocopies are strongly influenced by
positive and negative selection in natural populations, we ask whether they are actually better
suited for analyzing population differentiation processes than mostly neutral SNPs. We show that
this is indeed the case, thus confirming the largely non-neutral evolutionary dynamics of newly
derived retrocopies.

Results

We made use of 2,025 house mouse lineage-specific retroCNVs with mapped insertion site that
were identified in our previous study (Zhang, et al. 2021). All of them have not reached fixation
among the 96 wild caught mice individuals derived from nine natural populations, covering all
the three major subspecies of the house mouse (M. m. domesticus, M. m. musculus, and M. m.
castaneus). These retroCNVs are derived from 1,372 unique parental genes (Supplementary
Table S1). Only 81 (4%) of these retroCNVs are annotated as recently originated retrocopies in
the mm10 reference genome based on RetrogeneDB v2 (Rosikiewicz, et al. 2017) (≥95%
alignment identity to their parental gene, see (Zhang, et al. 2021)) (Supplementary Figure S1A).

Most of the retroCNV parental genes (71%) have only one retroCNV allele, while the rest are recurrently retroposed to generate multiple gene retrocopies with up to eight distinct insertion sites across individual genomes (Supplementary Figure S1B).

House-keeping genes as the major source of gene retroposition

Through analyzing the transcriptomic sequencing data from thirteen tissues of the Mus musculus C57BL/6 inbred line (Lin, et al. 2014), we found that almost all the retroCNV parental genes (98.9%) showed detectable expression in germline tissues of testis and ovary (Figure 1A), consistent with the prediction of a ubiquitous germline expression of parental genes (Kaessmann, et al. 2009). The remainder may be expressed during germline development to enable gene retroposition events to become inherited. We found both a significant higher fraction of genes with expression (FPKM>0, Figure 1A) and a significant higher expression level for retroCNV parental genes in germline tissues (Figure 1B), compared with all the annotated protein-coding genes with at least 2 exons as a control (only protein-coding genes with at least 2 exons were used for the detection of gene retrocopies, see (Zhang, et al. 2021)). Additionally, in-depth gene expression pattern analysis revealed that retroCNV parental genes being recurrently retroposed to generate multiple retrocopies had a significant higher expression level in germline tissues, in comparison with the ones that were retroposed once only (Supplementary Figure S2).

Interestingly, a similar pattern on a higher expression level of retroCNV parental genes was also observed for all the tested somatic tissues (Figure 1A and 1B, and Supplementary Figure 2).

Furthermore, gene ontology (GO) enrichment analysis on these retroCNV parental genes revealed their involvement in a variety of biological processes for the maintenance of cellular function (Supplementary Table S2), a typical characteristic of house-keeping genes (Eisenberg and Levanon 2003), suggesting these retroCNV parental genes tend to be house-keeping genes.
Figure 1 Characteristics of retroCNV parental genes. Comparison of fraction of genes with detectable expression (A) and expression level (B) between retroCNV parental genes and annotated protein-coding genes with at least two exons. The detectable expression is defined as FPKM > 0. (C) shows the fraction of excess of observed retroCNV parental genes in comparison with null expectation, on the basis of the distribution of annotated protein-coding genes of at least two exons within each evolutionary phylostratum (PS). The PS assignment data for mouse protein-coding genes were retrieved from (Neme and Tautz 2013). The greater values succeeding the prefix “PS” indicate the younger gene groups. Error bars in panel (A) show the standard errors of average fractions of expressed genes for annotated protein-coding genes with at least two exons (N=20,500), from which 100 times random sampling processes of the same gene number as that of retroCNV parental genes (N=1,372) were performed. All the statistical analyses were performed by using Wilcoxon rank-sum tests. ***: p-value ≤ 0.001.

Given that the evolutionary old genes usually harbor broader expression and higher expression levels in multiple tissues (Zhang, et al. 2015), one can speculate that retroCNV parental genes should be enriched in evolutionary older gene groups. To test this hypothesis, we calculated the fraction of excess of observed retroCNV parental genes with null expectation, on the basis of the distribution of annotated protein-coding genes with at least two exons (same as above), for which the gene age assignment data of mouse protein-coding genes was retrieved from (Neme and...
Tautz 2013). Our results indeed showed a positive excess of retroCNV parental genes that originated before Metazoan, while a consistent negative excess can be observed for all the evolutionary younger gene groups after that split (Figure 1C). The analysis based on another independent gene age assignment data from (Shao, et al. 2019) revealed the same pattern on the enrichment of retroCNV parental genes on evolutionary older gene groups (Supplementary Figure S3). These findings further support the conclusion of house-keeping genes as the major source of gene retroposition.

Indicators for positively selected gene retrocopies

Most newly originated gene retrocopies are deleterious and are quickly selectively purged (Zhang, et al. 2021), but a fraction of them show higher frequencies in natural populations, either because they are neutral and have accumulated through drift effects, or they have a beneficial role and rise in frequency due to positive selection.

To compare these two possibilities, we have categorized retroCNVs into three distinct groups (Figure 2A), with singletons (N=977, individual private retroCNVs, i.e., those found only in a single individual) forming the low-frequency group, retroCNVs present in 2-5 individuals forming the intermediate-frequency group (N=552) and present in ≥6 individuals forming the high-frequency group (N=496, note that the latter includes retrocopies that are fixed in given populations).

First, we studied the chromosomal location of parental genes for retroCNVs in the three classes, since it has been suggested that retroCNVs from the X chromosome should have a higher likelihood to have an advantageous effect (Emerson, et al. 2004; Schrider, et al. 2013), presumably triggered by the compensation of meiotic sex chromosome inactivation. The retroCNVs in the singleton group are expected to represent the most recent integration events, thus they can be taken as the baseline of gene retroposition pattern. We found that they show no bias towards coming from genes on the X chromosome, while the high-frequency group shows a strong bias and the intermediate frequency group lying between (Figure 2A). When a direct comparison between low-frequency retroCNVs and high-frequency retroCNVs was performed, we found a significant excess of gene retroposition coming from the X chromosome (p-value =
This pattern suggests that positive selection may play a role on the retention and increase in frequency of some retroCNVs derived from the X chromosome (see Discussion).

**Figure 2** Features of retroCNVs from different frequency groups. (A) Fraction excess of observed retroCNVs with each frequency group generated from the X chromosome. The observed and expected number of retroCNVs generated from the X chromosome for each frequency group is provided within the respective bar. The expectation of generating retroCNVs from each chromosome is on the basis of the number of protein-coding genes (with ≥2 exons) within the chromosome. (B) Fraction of retroCNVs from each frequency group with insertion site landing in different types of genomic elements. The annotation data of genomic elements in mouse mm10 reference genome is retrieved from Ensembl version 87 (Cunningham, et al. 2019). The statistical significances (p-values) on the excess of null expectation between X chromosome and autosomes, and the comparisons between pairs of frequency groups, were computed using Fisher’s exact tests. **: p-value ≤ 0.01; *: p-value ≤ 0.05; ns: p-value > 0.05.
Next, we explored the insertion site pattern of retroCNVs from the three frequency groups, regarding the genomic context of their placement (Figure 2B). As expected, most of the retroCNVs are enriched in the intergenic regions, since the non-coding intergenic regions make up the majority of mouse genome sequences. However, the intergenic regions are somewhat less represented for retroCNVs than random expectation, given that they take up >95% of mouse genome sequences (Cunningham, et al. 2019). This bias can be primarily explained by the enrichment of low-complexity repetitive sequences in the intergenic regions (Mouse Genome Sequencing, et al. 2002), which reduces the power to detect retroCNVs in our computational discovery pipeline, since it requires one read from given paired-end reads to map to a unique genomic region flanking the retroCNV insertion site (Zhang, et al. 2021). It should be noted that such a pattern should not affect the conclusions below, as we were analyzing the propensity of inserting in each type of genetic element for retroCNVs from three frequency groups independently.

Interestingly, we found that high-frequency retroCNVs are more likely to occur in intergenic regions (high-frequency retroCNVs versus low-frequency retroCNVs: p-value=0.008, Fisher’s exact test). This suggests that they are enriched in regions where they do not disrupt other genes, i.e., they have a higher likelihood to be neutral or even advantageous. Besides, we observed that intron regions are depleted for high-frequency retroCNVs (high-frequency retroCNVs versus low-frequency retroCNVs: p-value = 0.03, Fisher’s exact test), which might be ascribed to the avoidance of retrocopy insertions to disrupt mRNA splicing process (Zhang, et al. 2017). With respect to the insertions into CDS (coding DNA sequence) regions, we saw no differences among these three groups, in line with the notion that gene retroposition can be an important mechanism to generate chimeric retrogenes (Casola and Betran 2017).

**Signatures for positively selected gene retrocopies**

To detect direct signatures of positive selection connected to the increases of retroCNV frequency, we focused on retroCNVs of high-frequency (≥50%) in all representative populations of a focal subspecies (Supplementary Table S3), since the positive selection signals of high-frequency retroCNVs are more likely to be stronger and easier for detection. We performed this line of
analysis for each subspecies independently (*i.e.*, focusing on subspecies-specific retroCNVs), so as to minimize the influence of confounding factors, such as demographic effects.

Three types of population genetic signals have been reported as associated signatures of positive selection acting on retroCNVs with increased frequency in populations: i) reduction of average nucleotide diversity (Pi) among nucleotide sequences near the retroCNV insertion site (*i.e.*, selective sweep) for retroCNV carriers in comparison with non-carriers (Schrider, et al. 2013); ii) reduced Tajima’s D value among nucleotide sequences near the retroCNV insertion site for retroCNV carriers in comparison to background distribution (Llopart, et al. 2002); iii) elevated linkage disequilibrium (LD, or r^2) between SNPs near the retroCNV insertion site for retroCNV carriers in comparison to background distribution (Cardoso-Moreira, et al. 2016). Herein, we performed rigorous statistical tests on the above three statistics for each focal retroCNV against genomic background distributions (Cardoso-Moreira, et al. 2016), which were generated through randomly assigning “pseudo” insertion site of focal retroCNV into genomic regions (see Materials and Methods). We computed the statistical significance as how likely a same or stronger positive selection signature can emerge from the random genomic background distributions, and defined the positively selected retroCNVs as having statistical significances (p-value<0.05) for all three statistics.

Among 18 retroCNVs eligible for such tests (Supplementary Table S3), we found three with significant positive selection signatures based on the above criteria (Figure 3). Given the pre-selection of high-frequency retroCNVs that may show strong positive selection signals, these results suggest that most high-frequency retroCNVs could be neutral or have only moderate positive selection effects on their populations.
Figure 3 Subspecies-specific high-frequency retroCNVs that show significant positive selection signatures. Positive selection signature for each retroCNV was tested based on three distinct statistics: i) reduction of average nucleotide diversity ($\Pi_i$) among nucleotide sequences near the retroCNV insertion site (i.e., selective sweep) for retroCNV carriers in comparison with non-carriers (Schrider, et al. 2013); ii) reduced Tajima’s D value among nucleotide sequences near the retroCNV insertion site for retroCNV carriers (Llopart, et al. 2002); iii) elevated linkage disequilibrium (LD, or $r^2$) between SNPs near the retroCNV insertion site for retroCNV carriers (Cardoso-Moreira, et al. 2016). Only three retroCNVs showing statistical significance (p-value<0.05) for all the three statistics compared to random genomic background distributions are presented in this figure. The results on all the tested retroCNVs can be found in Supplementary Figure S4 and Table S3. For each retroCNV, the IGV snapshot in the upper panel shows the genomic context (i.e., annotated genes in the mm10 reference genome) along the 50kb upstream and downstream flanking region of the retroCNV insertion site. The solid lines in the lower panel represent nucleotide diversity ($\Pi_i$) within the same flanking region examined in a sliding window of 10kb and a step size of 5 kb. The dashed grey lines mark the insertion site of retroCNVs. Blue line: *M. m. domesticus* individuals of retroCNV carriers; Green line: *M. m. castaneus* individuals of retroCNV carriers; Grey line: retroCNV non-carriers in corresponding focal subspecies.

RetroCNVs as population differentiation markers

The patterns of retroCNVs in natural populations are shaped by a combination of purifying selection (Zhang, et al. 2021) and positive selection forces, as shown above. Hence retroCNVs that show elevated population-specific frequencies can be predicted to be more effective markers
for recent population differentiation than neutral SNPs (Lee, et al. 2014). We set out to test this prediction in our data. Firstly, pair-wise individual comparison on the sharing of retroCNVs showed indeed a more similar retroCNV landscape for individuals from the same subspecies/population (Supplementary Figure S5), indicating the potential of retroCNVs to infer population relationships. Secondly, in the overall PCA analysis, we observed clear clusters of individuals from each mouse population, including the closely related ones when analyzed by subspecies (Figure 4A). Similarly, heat map plotting based on the pair-wise comparison of individual retroCNV landscapes resolves the populations in defined clusters according to their origin (Figure 4B). Finally, on the basis of the presence fraction spectrum of retroCNVs within each population, it is possible to resolve a phylogenetic tree of house mouse populations (Figure 4C), for which most of the split nodes are of high confidence (bootstrap support value >80%). The only one exception is the split node between the two French populations, since they are geographically adjacent to each other. The phylogenetic tree is mostly consistent with the one built with the same computational protocol based on the SNP variants that were called from the same genomic sequencing dataset (Figure 4D), but the nodes of the retroCNV tree are better resolved as predicted. This is most evident for the relationship between the three M. m. musculus populations, where the SNP data would suggest a more basic position of the CZ population and a sister group relationship between the KA and AF populations, but with only a poor resolution. The retroCNV data, on the other hand, resolve this clearly and place the AF population more basic, which is also in line with previous conclusions of AF being ancestral and CZ being derived (Hardouin, et al. 2015). We conclude that retroCNVs can serve as effective genomic markers to differentiate natural populations better than SNP markers.

Given their effectiveness as population markers, we then performed a retroCNV frequency based approach to detect the retroCNVs as population differentiating markers (see Materials and Methods). Each retroCNV was assigned a population differentiation index (PDI), to indicate its power to differentiate populations. After applying a false discovery rate (FDR) threshold of 0.05, we detected in total 1,800 retroCNVs (88.6% of all retroCNVs) that could differentiate house mouse natural populations (with significant larger PDI values compared with null distribution from 1000 times random shuffling; see Supplementary Table S1).
Figure 4 Correlation between the retroCNV landscape and population differentiation. (A) Projection of the top two principal components (PCs) of retroCNV variation in house mouse individuals. Enlarged insets represent the analysis results focusing on populations from the subspecies of *M. m. domesticus* and *M. m. musculus* separately, since they cannot be well distinguished in the main figure. (B) Heat map on the distance between pair-wise individual comparison on retroCNV landscape of all sampled house mouse individuals. The distance between pair-wise individuals was defined as the fraction of different retroCNVs of the two individuals to be compared. (C) and (D) show phylogenetic trees built on the basis of the presence fraction spectrum of retroCNVs within each population and SNP variants that were called from the same genomic dataset, respectively. Split nodes marked in yellow are the ones with bootstrap support value >80%. Abbreviations for geographic regions: IR, Iran; FR_C, France (Central Massif); FR_A, France (Auvergne-Rhône-Alpes); GE, Germany; AF, Afghanistan; KA, Kazakhstan; CZ, Czech Republic; IN, India; TA, Taiwan.

Discussion

Gene copy number variation represents a major source of genetic divergence, and has been shown to play a role in the population differentiation in natural populations of the house mouse (Scavetta and Tautz 2010; Pezer, et al. 2015). However, as a particular type of gene copy number variation, gene retrocopies have been missed in previous studies of copy number variation.
Consequently, their contributions to the adaptation and population differentiation remain largely unknown. Our study represents such a direct analysis to address this question in natural populations of the house mouse. Here we present a comprehensive study of retrocopy dynamics in natural population samples encompassing several lineages and subspecies. We had previously shown that many new retrocopies have deleterious effects in these populations, thus contributing significantly to the overall mutational load (Zhang, et al. 2021). Here we explore the retrocopies that are not quickly purged, and we find that they contribute to new adaptations in the respective populations.

**Source of gene retroposition**

In order to generate heritable retrocopies, gene retroposition events need to occur in the germline, thus it was suggested that genes with higher expression in the germline tend more likely to be the source of gene retroposition (Kaessmann, et al. 2009). In support of this prediction, our analysis indeed showed a ubiquitous expression of retroCNV parental genes in the testis and ovary. This is further supported by the observation that retroCNV parental genes being recurrently retroposed to generate multiple retrocopies have a significant higher expression level than the ones that are retroposed once only. These findings therefore suggest that the propensity of a gene to serve for a retrocopy may be a simple function of the amount of mRNA present in the germline.

Intriguingly, these retroCNV parental genes were also found to be highly expressed in all tested somatic tissues. In fact, these genes are likely house-keeping genes involved in a variety of biological processes for the maintenance of the cellular function and thus are probably subject to general transcription in all cells, which is in accordance with the observation of the source of parental genes that are retroposed to generate processed pseudogenes detected in the mouse reference genome (Zhang, et al. 2004). This conclusion is strengthened by the finding that retroCNV parental genes are enriched in evolutionary older gene groups, which tend more likely to be broadly expressed house-keeping genes (Zhang, et al. 2015).

**Positive selection acting on new gene retrocopies**
An intriguing pattern on the chromosomal traffic of gene retroposition has been observed with respect to the X chromosome, which has generated a disproportionately high number of fixed functional retrogenes (called “out-of-X” pattern) in mammalian genomes (Emerson, et al. 2004; Carelli, et al. 2016). It was proposed that this could be related to meiotic sex chromosome inactivation (MSCI), where genes on the X and Y chromosomes are transcriptionally silenced in male germ cells. But retrocopies of house-keeping genes from the X chromosome are expressed in this phase and may therefore partially compensate the silencing of the X-chromosomal genes in the post-meiotic phase of sperm development (Kaessmann, et al. 2009). Hence, the observed “out-of-X” pattern would therefore not be due to a higher propensity of X-chromosomal genes to become retrogenes, but due to a higher potential for adaptive retention of these genes (Emerson, et al. 2004).

Our data confirm this assumption, as we did not find evidence for a strong initial bias in the generation of new gene retrocopies by parental genes from the X chromosome. However, there is a tendency among the ones occurring at higher frequency to have derived from the X chromosome, supporting the notion that more positively selected ones are among them.

The action of positive selection on some retrocopies is further supported by identifying direct positive selection signatures for three subspecies-specific high-frequency retroCNVs, making them candidates for a particularly strong adaptive process in the respective populations. The agreement between the elevated LD in the flanking regions and the high frequency of these retroCNVs suggests that these retroCNVs are under direct positive selection (Cardoso-Moreira, et al. 2016). Moreover, we found no SNPs or CNVs (copy number variants) within the 50kp up/downstream flanking regions that show complete LD ($r^2=1$) to these retroCNVs. This finding further supports these retroCNV as the direct target of positive selection. Still, it should be noted one cannot completely rule out the possibility from the effects of partial linkage of these variants or other undetected polymorphisms. Hence, we only interpret these findings as suggestive evidence to show the adaptive signatures of these retroCNVs. All three of the parental genes for these gene retrocopies are located on autosomes. They code for a transcription factor interacting protein (Atf7ip2), an enzyme (Apobec2) and a vesicle protein (Vta1), none of which was so far implicated in specific evolutionary processes.
Our results also show that the evolutionary fate of gene retrocopies is somewhat associated with the genomic context of their placement. The retroCNVs landing in the intergenic regions have a higher probability to increase their frequencies in populations, while those ones occurring in the genic regions are restricted to lower frequency. Even though a low fraction of the latter ones may evolve adaptive function such as generating chimeric retrogenes through landing in the exon regions of existing genes (Casola and Betran 2017), the majority of them can be expected to interfere the function of host genes (Schrider, et al. 2013). The retroCNVs occurring in the intergenic regions, on the other hand, tend more likely to be neutral or even advantageous, and consequently have a higher likelihood to reach fixation than those ones in the genic regions. All together, these findings suggest that positive selection may drive the spread of retroCNVs in populations, and these retroCNVs might play a role in the adaptation process of natural populations of the house mouse.

**Contribution of gene retrocopies on population divergence**

Given that the patterns of retroCNVs in natural populations are shaped by a combination of purifying selection (Zhang, et al. 2021) and positive selection forces as shown in this study, it is reasonable to assume that the retroCNVs could serve as more effective markers to trace the evolutionary history of natural populations in better phylogenetic resolution than the widely used mostly neutral SNPs (Lee, et al. 2014). Even though a few attempts have been made to explore gene retrocopy variation in natural populations, so far the analyses of the contribution of gene retroposition on population differentiation were only performed at the level of retroCNV parental genes due to the low resolution to detect gene retrocopy alleles (Abyzov, et al. 2013; Zhang, et al. 2017). However, in case of inserting into different genomic regions, distinct gene retrocopies can be generated from gene retroposition events from the same parental genes. Owing to the sufficient resolution to call insertion sites of retroCNVs with the unique population genomic dataset (Zhang, et al. 2021), our study allows a direct analysis to explore the contribution of gene retrocopies to population divergence in the house mouse.

Indeed, we found that gene retrocopies could be efficient markers to infer population relationships, and to resolve the phylogenetic tree of house mouse natural populations, even with better performance than the commonly used SNP variants. We further found that most of the
retroCNVs could differentiate house mouse natural populations. Overall, we conclude that the retroCNVs per se can be applied as effective genomic markers to trace the divergence and differentiation history of natural populations.

**Materials and Methods**

**RetroCNV and SNP variant datasets**

The dataset of house mouse lineage-specific retroCNVs was retrieved from our previous study (Zhang, et al. 2021). In brief, an optimized computational pipeline combining both exon-exon and exon-intron-exon junction read mapping strategies was exploited to identify gene retroposition events occurring in 96 surveyed mice individuals from nine natural populations of the house mouse (*Mus musculus*), while absent in both outgroup species (*Mus spretus* and *Mus spicilegus*). Following the “gold standard” for calling novel retrocopies (*i.e.*, with detectable genomic insertions (Richardson, et al. 2014)), we only focused on the house mouse lineage-specific retroposition events with mapped insertion site (Table “All data” of Dataset S3 in (Zhang, et al. 2021)), corresponding to 2,025 unique house mouse lineage-specific retroCNV alleles that are derived from 1,372 unique parental genes (Supplementary Table S1) (Zhang, et al. 2021).

Since the full-length sequences of most gene retrocopies are not available (*i.e.*, those are absent in the mm10 reference genome), we did not separate the ones with intact ORFs (open reading frames) from those not.

We obtained the dataset of house mouse lineage-specific SNP variants generated in the same study (Zhang, et al. 2021), for which the SNP variants were called by analyzing the same genomic sequencing dataset for the detection of gene retroposition events, based on the general GATK version 3 Best Practices (Van der Auwera, et al. 2013). Similarly, we only kept the SNP variants with unambiguous ancestral states in out-group species (*i.e.*, same homozygous genotype for all tested individuals from 2 out-group species), while with alternative allele in house mouse individuals for analysis in this study.

**Feature characterization of retroCNV parental genes**
We characterized the features of retroCNV parental genes from three different aspects: a) gene expression pattern across multiple tissues; b) GO functional term enrichment; c) gene age distribution pattern.

Firstly, we downloaded the raw Illumina paired-end RNA-Seq fastq data for *Mus musculus* C57BL/6 inbred line generated by the lab of Michael Snyder in Stanford (Lin, et al. 2014), from the ENCODE portal (Davis, et al. 2018) (https://www.encodeproject.org/). This dataset contains strand-specific transcriptomic sequencing data for 13 tissues dissected from 10-week old mice (Supplementary Table S4), including 2 germline tissues (ovary and testis), and 11 somatic tissues (adipose tissue, adrenal gland, brain, heart, kidney, liver, lung, pancreas, sigmoid colon, small intestine and spleen). We trimmed these fastq files by using Trimmomatic (v0.38) (Bolger, et al. 2014), and only kept paired-end reads passed filtering process for further analyses. We mapped the trimmed RNA-Seq reads to mm10/GRCm38 reference genome sequences with HISAT2 (v2.1.0) (Kim, et al. 2015), taking advantage of the mouse gene annotation in Ensembl v87 by using the --ss and --exon options of the hisat2-build. Then we counted the fragments mapped to the annotated genes with featureCounts (v1.6.3) (Liao, et al. 2014). Finally, we calculated the expression level in FPKM (Fragments Per Kilobase of transcript per Million mapped fragments) for each annotated coding gene within each tissue on the basis of gene’s transcript length and the number of mapped fragments in the dataset from each tissue. The expression patterns of retroCNV parental genes were compared with those of all protein-coding genes (with ≥2 exons) annotated in Ensembl v87 as a control. Given the sample size imbalance between retroCNV parental genes (N=1,372) and all protein-coding genes with multiple exons (N=20,500), we performed a bootstrapping analysis for the latter gene group, by randomly sampling the same number of genes as that in the former group for 100 times. The average fractions of genes with expression and standard errors were then computed for all protein-coding genes with multiple exons based on the random sampling datasets. Additionally, we also compared the expression levels between retroCNV parental genes that were retroposed once only and those ones retroposed recurringly (i.e., with ≥2 insertion alleles in the dataset). We utilized Wilcoxon rank-sum tests to compute the statistical significances on the differences of the fraction of genes with expression (FPKM>0) and gene expression levels.

Secondly, we performed gene ontology (GO) enrichment analysis on retroCNV parental genes with DAVID [Database for Annotation, Visualization and Integrated Discovery (DAVID)
Bioinformatics Resources 6.8] (Jiao, et al. 2012). Only significant GO BP (biological process) terms after multiple testing corrections (FDR < 0.05) were kept for presentation.

Lastly, we analyzed the evolutionary gene age pattern of these retroCNV parental genes. Two independent mouse protein-coding gene age assignment dataset were downloaded from GenTree (http://gentree.ioz.ac.cn/download.php) (Shao, et al. 2019) and (Neme and Tautz 2013), respectively. In brief, each mouse protein-coding gene was dated and given phylolstratum/branch assignment by inferring the absence and presence of orthologs along the phylogenetic tree, while using different computational approaches in two datasets. The expected number of retroCNV parental genes in each age group was calculated on the basis of the proportion of protein-coding with at least 2 exons in this group, since single-exon genes were excluded for the detection of retroCNV alleles in our previous study (Zhang, et al. 2021). The fraction of excess of observed retroCNV parental genes for each age group was computed as the deviation between observation value and expectation value divided by the value of expectation within the group. This analysis was performed for each gene age group separately.

### Chromosomal pattern analysis

All these 2,025 retroCNVs were classified into three roughly similar-sized groups based on their frequencies in all sampled house mouse individuals: low-frequency group (frequency=1; N=977), intermediate-frequency group (frequency between 2 to 5; N=552), and high-frequency group (frequency≥6; N=496).

We computed the fraction of excess of observed retroCNVs generated from X chromosome and autosomes for each frequency group. The expected number of generated retroCNVs from each chromosome was calculated based on the number of protein-coding genes (with ≥2 exons) in the focal chromosome. The fraction of excess of observed retroCNVs for each frequency group was computed as the deviation between observation value and expectation value divided by the value of expectation within the focal frequency group, and the statistical significance (p-value) on the excess within each frequency group was calculated via Fisher’s exact tests. Additionally, we also calculated the statistical significances on the excess of retroCNVs generated from X chromosome between pairs of frequency groups by using Fisher’s exact tests.
Insertion site pattern analysis

We analyzed and compared the overlap pattern between insertion sites of retroCNVs from the above frequency groups and genomic elements, including intergenic region, protein-coding gene region, intron region, exon region, CDS region and UTR region annotated in Ensembl version 87 (Cunningham, et al. 2019). The “Protein coding gene” category included the whole gene regions, i.e., the exon and intron regions. The “Intergenic region” category included the rest of the genomic regions not covered by these protein-coding genes. We required that the “Intron”, “Exon”, “CDS”, and “UTR” categories only come from protein coding gene regions.

Following the convention in (Zhang, et al. 2017), we enlarged the insertion intervals to 500bp, i.e., 250bp upstream and downstream from the middle point of the estimated insertion site range. We counted the number of genomic elements by using BEDTools v2.26.0 (Quinlan and Hall 2010), if there was at least 1bp overlap between the insertion intervals and the genomic elements. We computed the statistical significances on the fractions of retroCNVs landing in each type of genomic element between pairs of frequency groups by using Fisher’s exact tests.

Positive selection signal detection analysis

We refined a subset of 27 subspecies-specific high-frequency retroCNVs, by keeping those ones present in at least 50% of mice individuals for all respective populations of the focal subspecies (Supplementary Table S3). This dataset includes 9 retroCNVs in M. m. domesticus, 13 retroCNVs in M. m. musculus, and 5 retroCNVs in M. m. castaneus. Only retroCNVs that are both present and absent in at least 3 mice individuals for respective subspecies were included, and retroCNVs without SNP variants detected in the flanking regions of the insertion site were excluded for analysis. As a result, eighteen retroCNVs were kept for the following analysis (Supplementary Table S3). All SNPs located within 50kb upstream and downstream of the insertion sites (middle point) for the mice individuals with and without presence of the retroCNVs in the focal subspecies were extracted separately for further analysis by using VCFtools v0.1.14 (Danecek, et al. 2011), based on the above SNP dataset (Zhang, et al. 2021).
We firstly calculated three statistics for each retroCNV by using VCFtools v0.1.14 (Danecek, et al. 2011): i) the ratio of average nucleotide diversity (Pi) along 50kb up/downstream flanking regions between retroCNV carriers and non-carriers (Schrider, et al. 2013); ii) average Tajima's D along 50kb up/downstream flanking region for retroCNV carriers (Llopart, et al. 2002); iii) average LD ($r^2$) along the 5kb up/downstream flanking region for retroCNV carriers (Cardoso-Moreira, et al. 2016). Then we generated the genomic background distribution by assigning “pseudo” insertion site of focal retroCNV into random genomic regions (avoiding sequencing gaps annotated in the mm10 reference genome) for 100 times, and keeping the same setup of retroCNV “carrier” and “non-carrier” individuals. The random insertions were assigned to autosomes only, since all the tested retroCNVs are landing into autosomes (Supplementary Table S3). We further calculated the three above statistics for each “pseudo” retroCNV by following the same procedure above. The statistical significance (p-value) for each statistic was computed by using Wilcoxon rank-sum test, through checking how likely to get the same or stronger positive selection signatures (i.e., identical or smaller Pi ratio; identical or smaller Tajima's D; identical or greater LD) from these "pseudo" retroCNV insertions as that from true retroCNV insertion. As a result, we found three retroCNVs of positive selection signals with statistical significance (p-value<0.05) for all three statistics. For visualization, we calculated and plotted the average nucleotide diversity levels within the 50kb up/downstream (in a sliding window size of 10 kb and a step size of 5 kb) of the insertion site of each retroCNV for carriers and non-carriers separately (Figure 3 and Supplementary Figure S4).

Additionally, with the SNPs and CNVs (copy number variants) called from the same set of individual genomes (Zhang, et al. 2021), we searched for the variants that are within 50kb up/downstream flanking regions of these three retroCNVs and show exactly the same distribution pattern as that of the focal retroCNV (i.e., complete linkage: $r^2$=1). We did not find such SNPs or CNVs for any of these three positively selected retroCNVs.

**Correlation analysis between retroCNV pattern and population divergence**

We explored the association between retroCNV pattern and population differentiation by conducting three lines of analyses as following: a) Calculation and comparison on the fractions of shared retroCNVs between all possible pair-wise combinations from all individuals, individuals...
from different subspecies, from different populations but same subspecies, and from the same populations. b) Principal component analysis (PCA) on the individual retroCNV landscape with R package “ggfortify” v0.4.8 (https://cran.r-project.org/web/packages/ggfortify/index.html). Due to the close relationships among individuals from natural population in *M. m. domesticus* and *M. m. musculus*, we ran additional PCAs for these two subspecies separately. c) Calculation of the distance between pair-wise individual comparison on retroCNV landscape, on the basis of the fraction of different retroCNVs of two individuals to be compared. The heat map plotting of the distance measurement was conducted by using python package “matplotlib” v3.0.1 (https://matplotlib.org/).

**Building phylogenetic trees of house mouse natural populations**

We constructed the phylogenetic tree for nine natural populations in the house mouse by using neighbor-joining tree estimation function implemented in the R package “ape” v5.3 (Paradis and Schliep 2019), on the basis of the fraction of retroCNV presence spectrum in each population. We used Euclidean distance as the distance measure between each pair of populations. To assess the certainty of the above built phylogenetic tree, we performed 1,000 bootstrap replications (Efron, et al. 1996) with boot.phylo function implemented in R package “ape” v5.3. Population split nodes with at least 80% bootstrap support value generated from multiscale bootstrap resampling were taken as ones of high confidence.

To assess the performance of retroCNVs to trace evolutionary history of house mouse natural populations, we built the phylogenetic tree with the aforementioned SNP variants that were called from the same genomic sequencing dataset for the detection of gene retroposition events (Zhang, et al. 2021). To reduce computation complexity, we performed LD (linkage disequilibrium) pruning on the SNP dataset by using PLINK v1.90b4.6 (Purcell, et al. 2007), removing one of a pair of SNPs with LD ≥0.2 in sliding window of 500 SNPs and step wise of 100 SNPs. For each of the remaining SNPs, the fraction of individuals with derived SNP variant (without distinguishing homozygous or heterozygous status) was calculated for each population. We then followed the same procedure to construct the phylogenetic tree for the filtered SNP variants, as it was done for the analysis of retroCNVs.
Detection of population differentiation retroCNV markers

We exploited a retroCNV frequency based approach to detect population differentiating retroCNVs, similar to that was used for analysis of retroCNV parental genes (Zhang, et al. 2017). Each retroCNV was assigned a population differentiation index (PDI), a measure equivalent to the fixation index (Holsinger and Weir 2009), with the following equations:

\[
PDI = \frac{P(1-P) - \sum_{i=1}^{N} F_i \cdot p_i \cdot (1-p_i)}{p(1-P)} \tag{1}
\]

\[
F_i = \frac{N_i}{\sum_{i=1}^{N} N_i} \tag{2}
\]

Here, N represents the total number of natural populations (N=9), P is the frequency of a given retroCNV in the individuals from all populations, and \( p_i \) is the frequency in the \( i \)th population of the same retroCNV. Equivalent to the relative population size, \( F_i \) is calculated as the number of individuals of the \( i \)th population (\( N_i \)) divided by the total number of individuals from all populations. The significance of PDI for each retroCNV was tested by a randomization test in which we randomly shuffled individual labels for 1,000 times, while keeping the same population sizes. The p-value was calculated by comparing the observed PDI for each retroCNV against the PDI values from the above randomized null distribution. In total 1,800 retroCNVs after multiple testing corrections (FDR \( \leq 0.05 \)) were taken as population differentiation markers (Supplementary Table S1).

Data Availability:

There are no data to be archived. Computer codes, parameters and related datasets are available at GitLab: https://gitlab.gwdg.de/wenyu.zhang/mouse_retrocnv/.

Author Contributions:

W.Z. and D.T. designed the study, interpreted the data and wrote the paper. W.Z. analyzed the data and performed the statistical analysis. Both authors read and approved the final manuscript.
Acknowledgements

We appreciate the laboratory members for their helpful discussions. We thank Yong E. Zhang for his insightful suggestions. Computing was supported by the Wallace high-performance computing cluster of Max Planck Institute for Evolutionary Biology. This work was supported by institutional funding through the Max Planck Society (to D.T.).

References

Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, Genomes Project C, Lee C, Gerstein M. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Res 23:2042-2052.

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. BMC Genomics 9:466.

Bai Y, Casola C, Feschotte C, Betran E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol 8:R11.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in Drosophila. Genome Res 12:1854-1859.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Cardoso-Moreira M, Arguello JR, Gottipati S, Harshman LG, Grenier JK, Clark AG. 2016. Evidence for the fixation of gene duplications by positive selection in Drosophila. Genome Res 26:787-798.

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res 26:301-314.

Casola C, Betran E. 2017. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? Genome Biol Evol 9:1351-1373.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. 2019. Ensembl 2019. Nucleic Acids Res 47:D745-D751.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46:D794-D801.

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A 93:13429-13434.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet 19:362-365.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science 303:537-540.

Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome S, Analysis P, Platform, Harris CC, Ding L, Wilson RK, Haussler D. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol 14:R22.
Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Hardouin EA, Orth A, Teschke M, Darvish J, Tautz D, Bonhomme F. 2015. Eurasian house mouse (Mus musculus L.) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. BMC Evol Biol 15:26.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Z, Babiker H, Linnenbrink M, Montero I, Scavetta R, et al. 2016. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci Data 3:160075.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10:639-650.

Jiao X, Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2012. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics 28:1805-1806.

Kabza M, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polanski A, Makalowska I. 2015. Interpopulation Differences in Retrogene Loss and Expression in Humans. PLoS Genet 11:e1005579.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10:19-31.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357-360.

Lee TH, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923-930.

Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, et al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A 111:17224-17229.

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M. 2002. Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. Proc Natl Acad Sci U S A 99:8121-8126.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865-875.

Marques AC, Tan J, Ponting CP. 2011. Wrangling for microRNAs provokes much crosstalk. Genome Biol 12:132.

Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 14:117.

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife 5:e09977.

Ohno S. 1970. Evolution by gene duplication: Springer, Berlin, Heidelberg.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526-528.

Pezer Z, Harr B, Teschke M, Babiker H, Tautz D. 2015. Divergence patterns of genic copy number variation in natural populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major population-specific expansions. Genome Res 25:1114-1124.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Richardson SR, Salvador-Palomeque C, Faulkner GJ. 2014. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. Bioessays 36:475-481.
Rosikiewicz W, Kabza M, Kosinski JG, Ciomborowska-Basheer J, Kubiak MR, Makalowska I. 2017. RetrogeneDB-a database of plant and animal retrocopies. Database (Oxford) 2017.

Scavetta RJ, Tautz D. 2010. Copy number changes of CNV regions in intersubspecific crosses of the house mouse. Mol Biol Evol 27:1845-1856.

Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet 9:e1003242.

Schrider DR, Stevens K, Cardeno CM, Langley CH, Hahn MW. 2011. Genome-wide analysis of retrogene polymorphisms in Drosophila melanogaster. Genome Res 21:2087-2095.

Shao Y, Chen C, Shen H, He BZ, Yu D, Jiang S, Zhao S, Gao Z, Zhu Z, Chen X, et al. 2019. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. Genome Res 29:682-696.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 453:534-538.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11 10 11-11 10 33.

Zhang W, Landback P, Gschwend AR, Shen B, Long M. 2015. New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol 16:202.

Zhang W, Xie C, Ullrich K, Zhang YE, Tautz D. 2021. The mutational load in natural populations is significantly affected by high primary rates of retroposition. Proc Natl Acad Sci U S A 118.

Zhang Y, Li S, Abyzov A, Gerstein MB. 2017. Landscape and variation of novel retroduplications in 26 human populations. PLoS Comput Biol 13:e1005567.

Zhang Z, Carriero N, Gerstein M. 2004. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet 20:62-67.