The mutational load in natural populations is significantly affected by high primary rates of retroposition

Wenyu Zhang[1], Chen Xie[1], Kristian Ullrich[1], Yong E. Zhang[2] and Diethard Tautz[1,#]

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

[2]Key Laboratory of Zoological Systematics and Evolution and State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

E-mail address for all authors:

Wenyu Zhang: wyzhang@evolbio.mpg.de; Chen Xie: xie@evolbio.mpg.de

Kristian Ullrich: Ullrich@evolbio.mpg.de; Yong E. Zhang: zhangyong@ioz.ac.cn

#corresponding author:

Email: tautz@evolbio.mpg.de

Phone: +49 4522763390

Running title:

RetroCNVs in house mouse natural populations

Key words:

Gene retroposition rate, house mouse, natural population, selection, genetic load

Abstract

Gene retroposition is known to contribute to patterns of gene evolution and adaptations. However, possible negative effects of gene retroposition remain largely unexplored, since most previous studies have focussed on between-species comparisons where negatively selected copies are mostly not observed, as they are quickly lost from the populations. Here, we show for natural house mouse populations that the primary rate of retroposition is orders of magnitude higher than previously thought. Comparisons with SNP distribution patterns in the same populations show that most retroposition events are deleterious. Transcriptomic profiling analysis shows that new retroposed copies become easily subject to transcription and have an influence on the expression level of their parental genes, especially when transcribed in the antisense direction. Our results imply that the impact of retroposition on the mutational load in natural populations has been highly underestimated, which has also major implications for strategies of disease allele detection in humans.

Introduction

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 copies (retrocopies), and this has a variety of evolutionary implications (1–3). The intronless retrocopies have initially been viewed as evolutionary dead-ends with little biological effects (4, 5), mainly due to the assumed lack of regulatory elements and promotors. However, this hypothesis has become of less relevance after it has become clear that a large portion (> 80%) of the mammalian genome is transcribed (6, 7) and that there is a fast evolutionary turnover of these transcribed regions, implying that essentially every part of the genome is accessible to transcription (8). In addition, retrocopies can recruit their own regulatory elements through a number of mechanisms (2, 3). Hence, retrocopies can act as functional retrogenes that encode full-length proteins, and it has, therefore, been proposed that they contribute to the evolution of new biological functions through neofunctionalization or sub-functionalization (2, 3, 9–11). However, the possibility that

retroposition events could also be deleterious has been much less considered so far. Deleterious effects could be due to insertions into functional sites, and this has indeed been detected in a retrogene population analysis in humans (12). However, even if they land in non-functional intergenic regions, they could be transcribed and their transcripts could interfere with the function of the parental genes (13–15). In SNP based association studies, this would become apparent as a trans-effect on the parental gene, but the true reason for the trans-effect would remain unnoticed when the retrocopy is not included in the respective genomic reference sequence. Hence, if transposition rates are high and if the transposed copies are frequently transcribed, they could have a substantial impact on the mutational landscape of genomes.

Retroposition mechanisms were initially studied in between-species comparisons with single genomes per species (*e.g.* (16)), but these will miss all cases of retropositions with deleterious effects. Accumulating population genomics data are now providing the opportunity to detect novel retroposed gene copy number variants (retroCNVs) that are still polymorphic in populations (3), but a broad comparative dataset from related evolutionary lineages is required to obtain a deeper insight. A population analysis representing natural samples is available in humans, based on the 1,000 Genomes Project Consortium data (12, 17–20). However, the power of the discovery of retroCNVs in these studies has been limited due to the heterozygous and relatively low coverage sequencing datasets. Moreover, in humans it is not possible to compare the data with very closely related other lineages, since these are extinct (*e.g.*, Neandertals or Denisovans). Hence, a comprehensive analysis on the evolutionary dynamics of retroCNVs at comparable individual genome level, especially based on a set of well-defined natural populations from different lineages where evolutionary processes and transposition rates can be studied, is still missing.

The house mouse (*Mus musculus*) is a particularly suitable model system for comparative genomic analyses in natural populations, owing to its well-studied evolutionary history (21, 22). Currently, three major lineages of *Mus musculus* are distinguished, classified as subspecies which diverged roughly 0.5 million year ago: the western house mouse *Mus musculus domesticus*, the eastern house mouse *Mus musculus musculus* and the southeast-Asian house mouse *Mus musculus castaneus*. Previously, we have

generated a unique genomic resource using wild mice collected from multiple geographic regions (each representing one natural population) covering the three major house mouse sub-species, with a carefully designed sampling procedure to maximize the possibility of capturing the genetic diversity from each population (23). This was complemented by a well-controlled experimental set-up to generate largely homogeneous genomic/transcriptomic sequencing datasets at relatively high coverage for the same individuals (24), which makes it possible to trace directly the effects of new retroposed copies on the expression of their parental genes.

Here we show that the gain and loss rates of retroCNVs are many-fold higher than previously expected and the frequency spectra of retroCNV alleles in populations in comparison to SNP allele frequency spectra implies mostly deleterious effects. Transcriptome data show that the new retroCNVs are usually transcribed and have indeed an effect on the parental gene transcripts. A new strand-specific RNA-Seq dataset for one of the populations shows that antisense transcribed retroCNVs are highly underrepresented compared to sense transcripts, implying strong selection against them. We conclude that deleterious effects of newly transposed retrocopies of genes have been largely underestimated so far. We also discuss the implications for human disease allele detection.

Results

Full genome resequencing data of 96 individuals derived from nine natural populations corresponding to the three major subspecies (*M. m. domesticus, M. m. musculus and M. m. castaneus*) of the house mouse (*Mus musculus*), as well as nine individuals from two outgroup species (*M. spicilegus* and *M. spretus*), were used to assess gene retroposition events (Figure 1 and Supplementary Tables S1 and S2A). By adapting an exon-exon junction and exon-intron-exon junction mapping based approach for short read genomic sequencing data (18, 19, 25), we refined a computational pipeline to identify retroCNV genes (*i.e.*, retroposed parental genes) (Supplementary Material Text S1 and Figure S1) including a power

analysis for optimizing mapping conditions. A retroCNV gene is called, on condition that both the intron loss event and the presence of parental gene can be observed in the same individual sequencing dataset (25).

Due to the need to detect at least one exon-exon junction, only protein coding genes with ≥ 2 exons (~ 92.4% of all coding genes annotated in Ensembl v87) were assayed as potential source of gene retroposition. To compensate for the variance in the sequencing (read length, coverage, and etc.) and individual intrinsic features (*i.e.*, sequence divergence from the mm10 reference genome), we optimized the parameters (*i.e.*, alignment identity, spanning read length, and number of supporting reads) of the retroCNV gene discovery pipeline for each individual genome (Supplementary Material Text S1), and the resultant computational pipeline gave a low false positive discovery rate < 3% (Supplementary Material Figure S2) and a high recall rate of > 95% (Supplementary Material Figures S3-S5) for all the individual genomes tested.

A subset of the retroCNV alleles that were identified as newly arisen in one of our populations is also present in the mm10 reference genome. For these ones, we directly called their presence based on the alignment data of individual sequencing datasets to the reference genome. The insertion sites in the genome of retroCNV alleles that are absent in the mm10 reference genome, were estimated based on discordant aligned paired-end reads when these could be uniquely mapped (see Methods).

High numbers of retroCNVs in natural populations

Applying the above defined pipeline, we screened for retroCNV genes (*i.e.*, the parental genes from which retrocopies are derived) and retroCNVs (*i.e.*, alleles of the inserted retrocopies, or insertion sites in the genome in case that the retrocopies are not present in the reference genome) in the mouse individual genome sequencing datasets. To study turnover rates (*i.e.*, gains and losses), we focused on the recently

originated gene retroposition events in the house mouse lineage, *i.e.*, retroCNV genes and retroCNVs occurring in the *Mus musculus* subspecies but absent in the outgroup species.

In total, we identified 21,160 house mouse (*Mus musculus*) specific retroposition events within the 96 individuals derived from 1,663 unique retroCNV genes (Supplementary Table S3). Only 80 (4.8%) of these retroCNV genes have annotated recently originated retrocopies in the mm10 reference genome (≥ 95% alignment identity to their parental gene), while the other 1,583 retroCNV genes represent undescribed gene retroposition events in house mouse wild individuals. Approximately 3.9% of them show more than one retroCNV allele for the same retroCNV gene in the same individual genome (Supplementary Material Figure S8).

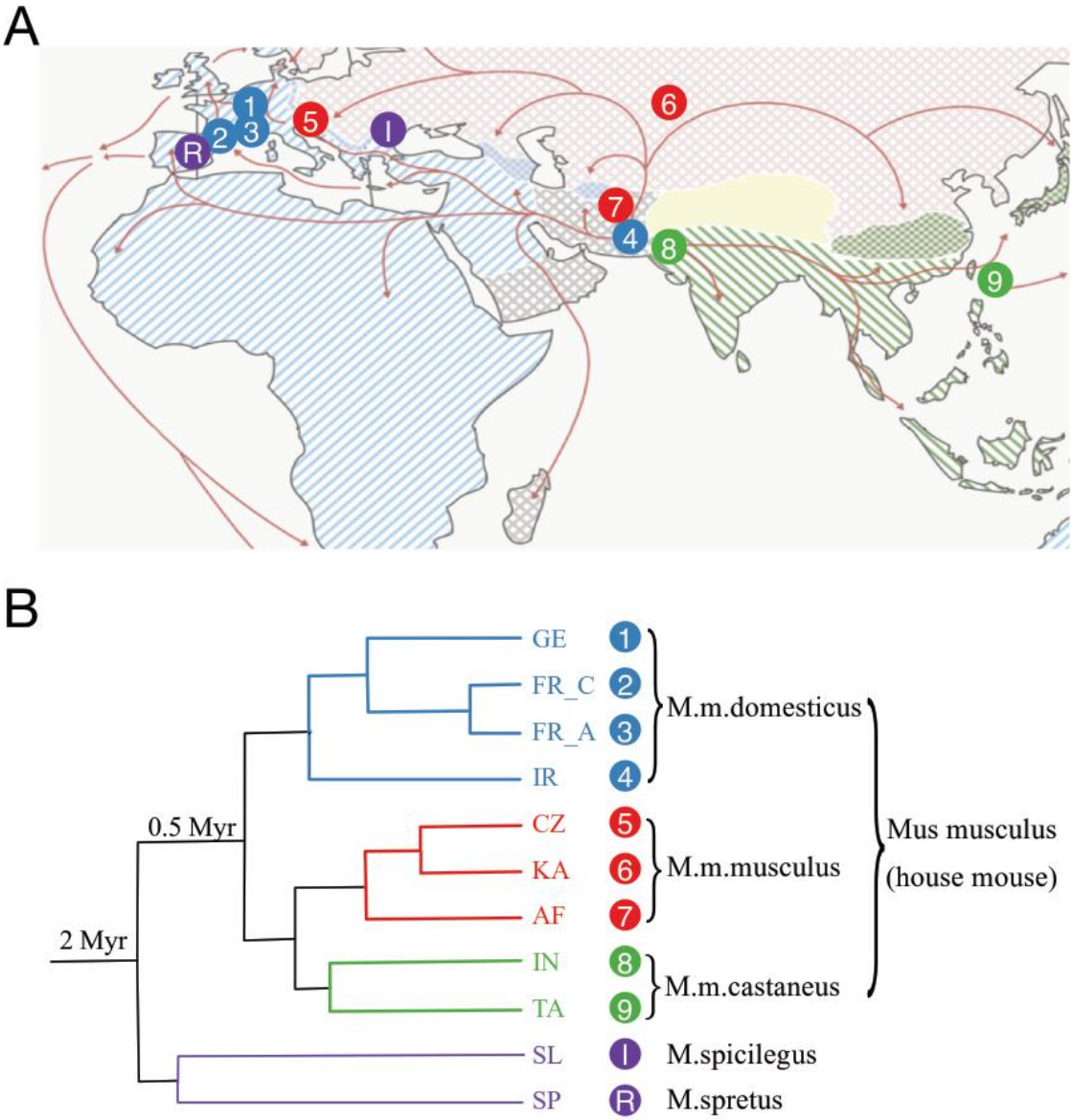

Figure 1: Geographic locations and phylogenetic relationships of house mouse subspecies and two out-group species samples. (A) Geographic location information on the sampled mouse individuals. This map is modified from (24). Territory areas for each house mouse subspecies: *M. m. domesticus* (blue); *M. m. musculus* (red); *M. m. castaneus* (green). Red arrows indicate possible migration routes, mostly during the spread of agriculture and trading. Geographic locations: 1, Cologne-Bonn/Germany (GE); 2, Massif Cenral/France (FR_C); 3, Auvergne-Rhône-Alpes/France (FR_A); 4, Ahvaz/Iran (IR); 5, Studenec/Czech Republic (CZ); 6, Almaty/Kazakhstan (KA); 7, Mazar/Afghanistan (AF); 8, Himachal Pradesh/India (IN); 9, Taiwan (TA); I, Sása/Slovakia (SL); R, Madrid/Spain (SP). (B) Phylogenetic relationships and split time estimates (branches not shown to scale) among the house mouse populations and two out-group species in the study.

Random resampling analysis of subsamples of individuals showed that the number of detectable retroCNV genes has not reached saturation with the given number of sampled individuals in our dataset (Figure 2A), implying that many more retroCNV transposition events should be found when more individuals would be

analyzed. Importantly, as suggested by (26), we can also show that previous CNV detection pipelines will usually not detect retroCNVs. In a direct comparison with data from genic CNV detection (27) for the same dataset, we found that most of the retroCNVs were not captured. Only <1%, on average, of the retroCNV genes detected in our analysis overlaps with genic CNVs according to this pipeline (Supplementary Material Figure S9).

On average, in each tested individual, there are 212 retroCNV genes, but the populations differ somewhat in the number of average retroCNV genes per individual (Figure 2B). Slightly higher numbers were found in the ancestral populations (*i.e.*, Iran population for *M. m. domesticus*, Afghanistan population for *M. m. musculus*, and India population for *M. m. castaneus*), presumably since they have higher effective population sizes where more neutral or semi-neutral retroCNVs could segregate. The majority of retroCNV genes (91% - 95%) in the wild-derived laboratory inbred strains representing the three subspecies (*M. m. domesticus*: WSB_EiJ; *M. m. musculus*: PWK_PhJ; *M. m. castaneus*: CAST_Eij) can also be discovered in house mouse wild individuals (Figure 2C-E). Conversely, the majority of retroCNV genes (73% - 87%) in wild-derived house mouse individuals are not present in the inbred mouse strains, since these represent essentially only single haplotypes from the wild diversity.

Around 38% - 78% of retroCNV alleles (mostly based on the estimates of insertion sites) for wild house mouse individuals could be identified (Supplementary Material Figure S10), depending on the nature of the sequencing read data features of each individual, *e.g.*, sequencing coverage, read length and insert size. The detection rate of insertion sites at the individual genome level presented here is much higher than the one that was reported from pooled human population genomes when the same criteria to define reliable insertion sites where applied (30% in (12)). Following the "gold standard" for calling novel retrocopies (*i.e.*, with detectable genomic insertions, (20)), unless stated separately, all the following analysis were conducted on the basis of retroCNVs (retrocopies), rather than retroposed parental genes (*i.e.*, retroCNV genes). Correspondingly, we included 2,025 house mouse specific retroCNVs (Supplementary Table S4) for further analysis.

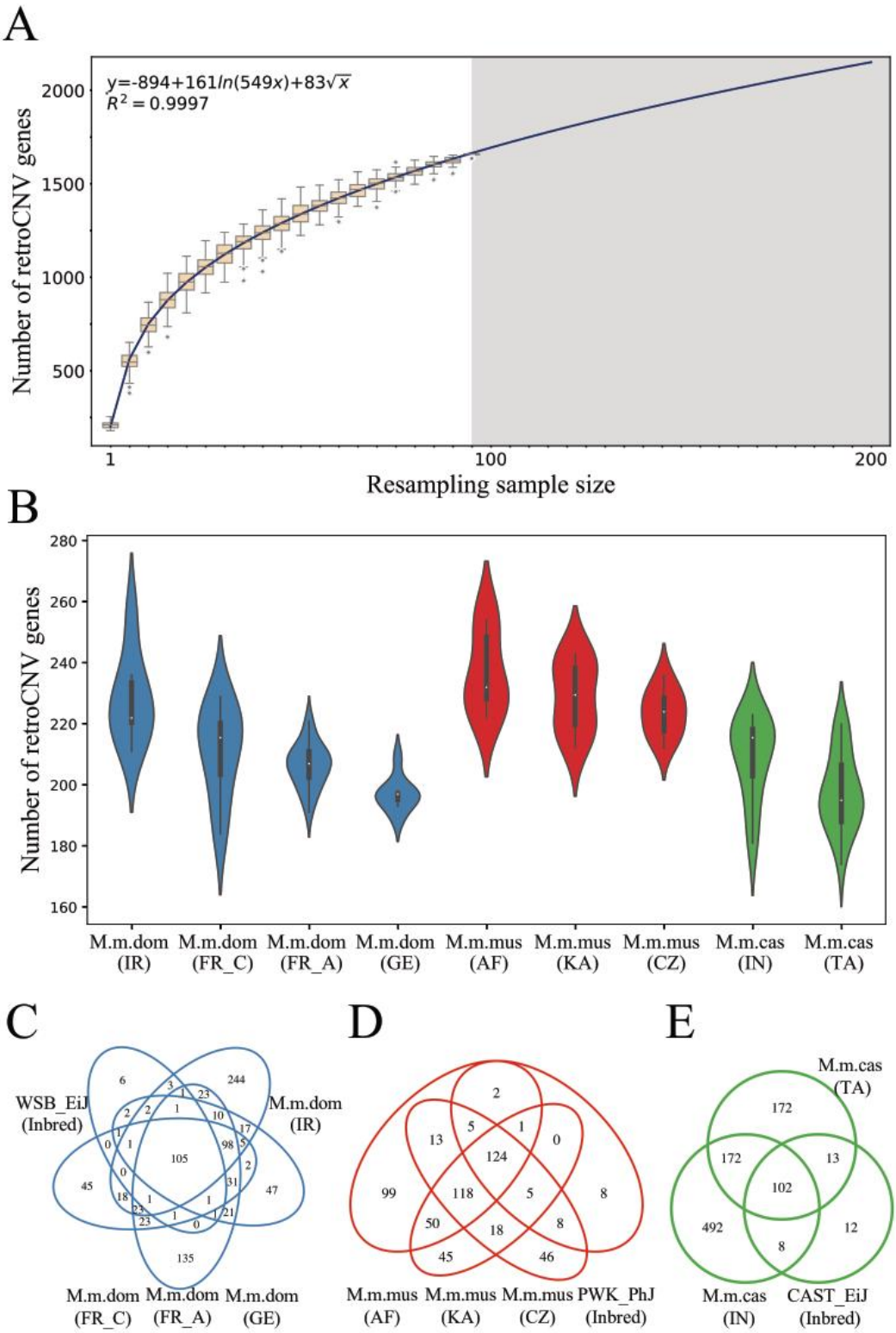

Figure 2: Distribution of the number of detected retroCNV genes across house mouse populations. Only *Mus musculus* specific retroCNV genes are included in this analysis. (A) Number of detected retroCNV genes with various random resampling sample sizes. The resampling subsample sizes were selected from 1 to 95, with step size of 5. Data points represent the average number of detected retroCNV genes of 100 replicates for each subsample, and

whiskers the standard variance of the mean deviation. The gray area shows the prediction after doubling the number of current sampling house mouse individuals. (B) Distribution of the number of detected retroCNV genes within each house mouse natural population (see Supplementary Material Figure S10 for a corresponding depiction of retroCNVs). (C) – (E) Depiction of the overlap of detected retroCNV genes between house mouse natural populations and inbred mouse lines derived from each of the three house mouse subspecies, respectively. Inbred mouse strains for three subspecies: WSB_EiJ (*M .m. domesticus*); PWK_PhJ (*M. m. musculus*); CAST_EiJ (*M. m. castaneus*). Abbreviations for geographic regions: *IR*, Iran; *FR_C*, France (Massif Central); *FR_A*, France (Auvergne-Rhône-Alpes); *GE*, Germany; *AF*, Afghanistan; *KA*, Kazakhstan; *CZ*, Czech Republic; IN, India; *TA*, Taiwan (see also Figure 1A for geographic representation).

Fast turnover of retroCNVs

The high-resolution retroCNV landscape in the house mouse wild-derived individuals gave us the opportunity to extensively explore the retroCNV variation at different levels. Among the entire 2,025 house mouse specific retroCNV alleles (Supplementary Table S4), only 71 (~3.5%) are found in all three house mouse subspecies (Figure 3A), of which 23 (~1.1%) are found segregating in all nine natural populations (Figure 3B). 1,797 (89%) retroCNVs are private to single subspecies (Figure 3A), around 70% can only be found in a single population (Figure 3B), and around 48% of them only in a single individual (Figure 3C). In fact, since the distribution of these private retroCNVs follows the known population history they can be used as population differentiation markers (Supplementary Material Text S2).

The Western European *M. m. domesticus* populations are derived from Iranian (IR) populations and have invaded Western Europe about 3,000 years ago where they quickly radiated. The split from the Iranian population would have occurred no more than 10,000 years ago (28, 29). This provides a time line to estimate retroCNV emergence rates by comparing the population and lineage-specific retroCNVs. We used the populations FR_C, GE and IR for this, since they are represented by the same number of individuals and were sequenced in a similar way. We found 60 and 57 private retroCNVs in FR_C and GE, respectively (Supplementary Table S4). Assuming these populations have split soon after their arrival, this would suggest in the order of 200 new retroCNV events in 10,000 years. In the IR population, we find 284 private retroCNVs (Supplementary Table S4), *i.e.*, assuming a separation of 10,000 years, this would be of the same order. These numbers are two orders of magnitude higher than those found previously in primate genome comparisons (21 to 160/Myr) (16). Intriguingly, even between the most recently separated

populations (FR_C and GE), we found only three retroCNVs shared, implying that the retroCNVs from the joint history of these two populations were quickly lost again and many lineage specific ones were created anew. This proofs that newly generated retroCNVs are usually deleterious and that their quick disappearance leads to a strong underestimation of primary retroposition rates. Hence, if one compares only single genomes from different species, as it was done for example for primates in (16), one would miss most of the events that can be identified in population samples. However, we note that if we would count only the retroCNVs occurred in all the nine house mouse natural populations, we would arrive at similar estimates as for the primate data (~12/Myr, Supplementary Table S4). Hence, our analysis shows that the turnover dynamics of retroposition events is orders of magnitude higher than previously assumed.

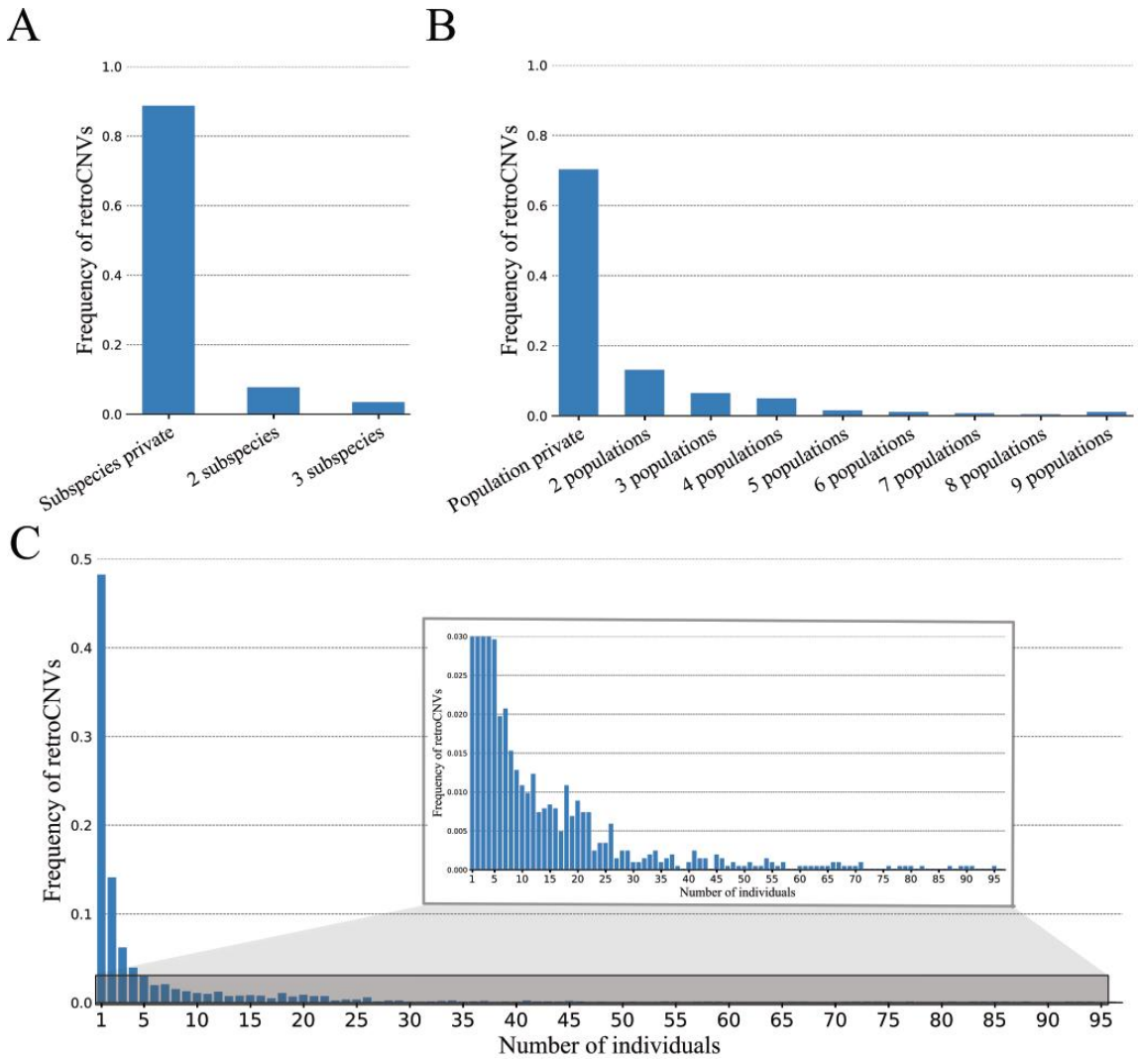

Figure 3: Distribution of the frequency of detected retroCNVs across different house mouse subspecies (A), populations (B) and individuals (C). Enlarged in the inset box (focused on frequencies below 0.03) is to show the frequencies of retroCNVs present in larger number of individuals.

To assess whether the gains and losses of retroCNVs are compatible with a neutral model, or whether they are influenced by positive or negative selection, we compared the site frequency spectra of the retroCNVs (Figure 4) with those of SNP allele categories for the same population samples. Based on the functions of these SNPs, we further categorized them into four distinct groups (30): 1) High effect SNPs that change the coding gene structure (stop codons or splice sites); 2) Moderate effect SNPs that change amino acid sites; 3) Low effect SNPs with synonymous changes; 4) Modifier effect SNPs that locate in non-coding regions. We found significantly more retroCNVs in the private category, *i.e.*, occurring only in a single animal for each of the categories (Fisher's exact test, retroCNV vs. high effect SNPs: p-value = 1.7×10^{-18}; retroCNV vs. moderate effect SNPs: p-value = 2.6×10^{-18}; retroCNV vs. low effect SNPs: p-value = 3.5×10^{-67}; retroCNV vs. modifier effect SNPs: p-value = 1.3×10^{-64}).

To test for similarity of the distributions, we used the two sided Kolmogorov-Smirnov tests and found more similar distributions between retroCNVs and the more constrained SNP categories (Kolmogorov's D statistic for retroCNV vs. high effect SNPs: D=0.14, retroCNV vs. moderate effect SNPs: D=0.13, retroCNV vs. low effect SNPs: D=0.21, retroCNV vs. modifier effect SNPs: D=0.21). Hence, we conclude that most new retroCNVs are under negative selection, *i.e.*, would not only be lost by drift, but are effectively "disease" alleles that are lost by selective purging in natural populations.

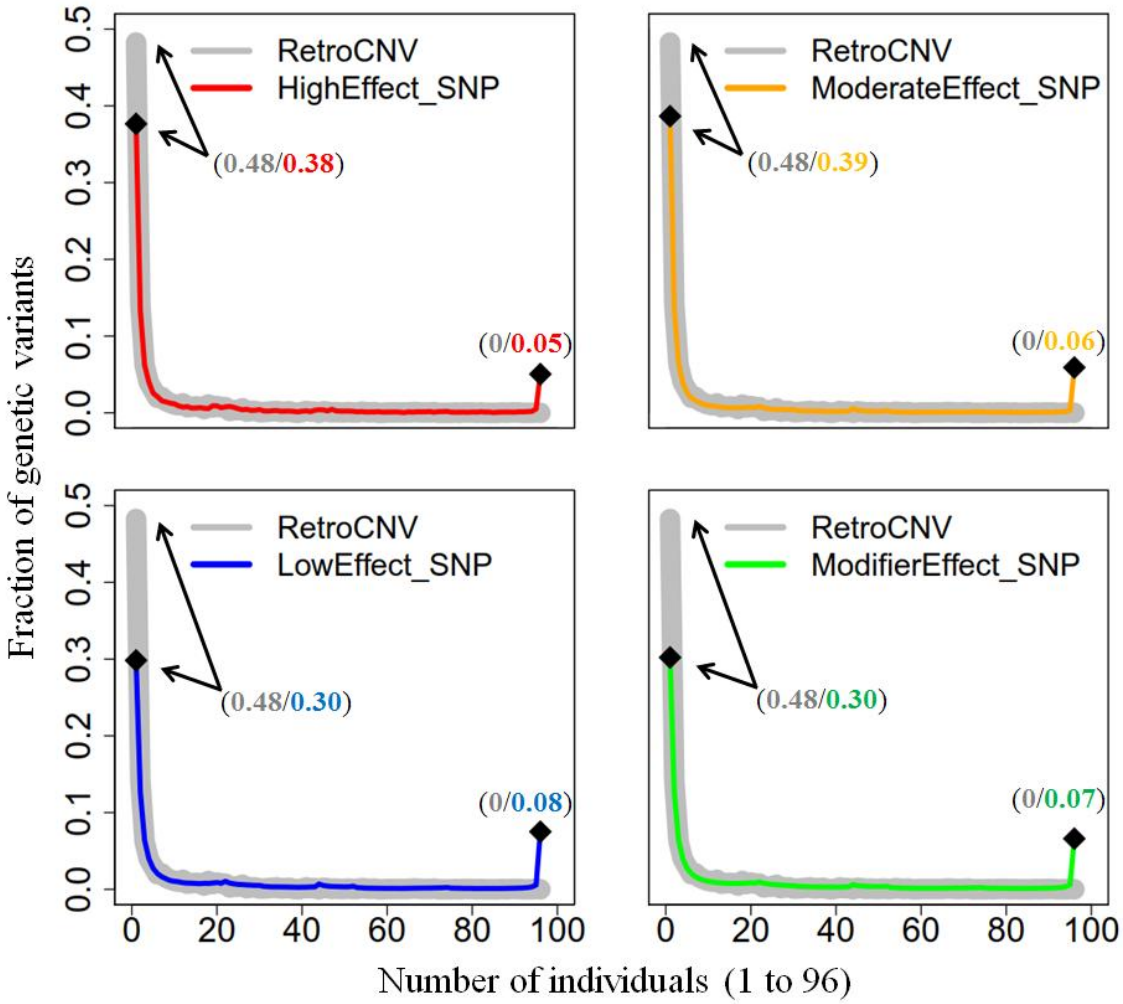

Figure 4 Comparison of the frequency spectrum of retroCNVs with the site frequency spectra of SNPs. High effect SNPs: the ones causing the gain/loss of start/stop codon or change the splicing acceptor/donor sites; Moderate effect SNPs: the ones resulting in a different amino acid sequence; Low effect SNPs: the ones occurring within the general region of the splice site, changing the final codon of an incompletely annotated transcript, changing the bases of start/stop codon (while start/terminator remains), or where there is no resulting change to the encoded amino acid; Modifier effect SNPs: the ones occurring around the coding regions of the genes (UTR, intron, up/downstream), non-coding gene regions, or intergenic regions. The numbers within the parentheses indicate the fractions of retroCNVs (in grey) and SNPs (colors corresponding to SNP categories) that are individual private or reach fixation in all 96 tested house mouse individuals, respectively.

RetroCNV expression

It has previously been shown that retrocopies can recruit proto-promotors from the vicinity of their insertion site (31). To assess expression of the newly inserted retroCNVs in the mouse populations, we used the transcriptomic dataset that was generated from the same individuals of the three natural *M. m.*

domesticus populations from Germany, Massif Central of France and Iran (GE, FR_C and IR) for which also the genome sequences were obtained that we used for the retroCNV detetion (Supplementary Table S2B). To combine this information, we focused on the recently originated retrocopies present in the mm10 reference genome, as annotated in RetrogeneDB version 2 (32), since full length information for the inserted fragment is available for them. As the newly originated retrocopies are usually highly similar to their parental genes (25), we implemented an effective length (a proxy to the divergence to the parental gene) based approach to calculate their specific expression, on the basis of reads perfectly and uniquely mapped to the new retrocopies (see Methods). Fifty-nine retrocopies with non-zero effective lengths were available for this analysis across the three *M. m. domesticus* populations. We found that most of them (55 out of 59) are expressed in at least one tissue or at least one population (Supplementary Table S5, summarized in Supplementary Table S6). Most are expressed in multiple tissues, whereby the expression levels usually differ between the populations. This implies that the majority of retroCNV copies became transcribed, although they responded differently to the regulatory context in their respective cell types and populations.

RetroCNV effect on parental gene expression

Given that we have the expression data from the same animals for which we have the genome sequences, it was possible to ask whether the presence of a new retrocopy in a given individual would affect the expression of the parental gene in the same individual. We restricted this analysis to the animals with singleton retroCNVs in the FR_C and GE populations, *i.e.*, the cases where only one individual of a given population carried the retroCNV. This allowed us to use the remainder of the seven individuals from the same populations to calculate an average parental gene expression plus its variance. We then used a Wilcoxon rank sum test to ask whether the presence of a retroCNV led to a significant expression change in the respective individual. We focused this analysis only on the loss of expression, since a gain could be due to the additional reads from the retroCNV itself (note that in contrast to the analysis in the above

section, we did not attempt to distinguish these reads for this extended dataset). We found that 22% (GE) and 31% (FR_C) of the singleton retroCNVs have in at least one tissue a significant negative effect (after multiple testing correction, FDR ≤ 0.05) on the expression of their parental gene (Table 1- Supplementary Table S7).

Table 1: Singleton retroCNVs with negative effects on their parental genes' expression in their population

Population	Total # of singleton retroCNVs	# of singleton retroCNVs with significant (FDR ≤ 0.05) negative effect on parental gene expression	Average # of tissues affected per singleton retroCNV
M. m. domesticus (GE)	74	16	1.81 ± 0.98 SD
M. m. domesticus (FR_C)	71	22	2.14 ± 0.99 SD

Strand-specific expression of retroCNVs

To assess whether the deleterious effects retrocopies could be due to silencing effects from antisense transcribed copies, we generated a strand specific RNA-Seq dataset that allowed sense and antisense transcripts to be distinguished. For this we used five tissues from 10 males from the outbred stock of *M. m. domesticus* FR_C population. Note that these are different individuals than the ones used in Harr et al. (24), but from the same breeding stock of outbred animals. Hence, we could use the same reference genome set of retroCNVs as above (50 retroCNVs occurred in FR_C population), for which parental and retroCNV transcripts can be distinguished. We found that 42 of these 50 retroCNVs are transcribed in at least one tissue, but with an extreme bias towards sense-transcripts (Table 2 - Supplementary Table S8). This applies not only to the number of transcribed retroCNVs per tissue, but also to the level of transcription (Supplementary Table S8). Since there is no reason to assume that the direction of integration into the chromosomes could be biased, we interpreted this finding as a strong selection against

retroCNV copies that showed antisense transcription, implying that they are affecting their parental genes via dsRNA silencing.

Table 2: RetroCNV expression patterns in the strand-specific RNA-Seq dataset

	Testis	Brain	Kidney	Liver	Heart
# of expressed retroCNVs (FPKM > 0, Sense strand)	35	25	31	20	23
[1]Average expression level in FPKM (Sense strand)	2.2 (SEM:0.9)	1.7 (SEM:0.7)	1.9 (SEM:0.9)	2.2 (SEM:0.9)	2.8 (SEM:1.7)
# of expressed retroCNVs (FPKM > 0, Antisense strand)	16	10	8	6	7
[1]Average expression level in FPKM (Antisense strand)	0.10 (SEM:0.05)	0.03 (SEM:0.01)	0.06 (SEM:0.04)	0.03 (SEM:0.02)	0.01 (SEM:0.01)

[1] Only the retroCNVs with non-zero expression were included;

SEM: standard error of mean

Discussion

Our population based retroCNV analysis allowed a much deeper insight into the retrogene formation dynamics than was previously possible. Most importantly, we found that the primary origination rate of retroCNVs must be orders of magnitudes higher than previously assumed. At the same time, the data showed that many newly retroposed copies influence the expression of their parental genes and are mostly subject to negative selection, *i.e.*, can be considered as "disease" alleles. Further, we showed that retroposed copies are not readily detected by previously established CNV detection procedures, *i.e.*, their impact on generating deleterious mutations has been highly underestimated.

The comparison between very recently separated mouse populations provided the unique possibility to estimate primary retroposition rates, *i.e.*, get an insight into the events that disappear over time from the populations due to negative selection. Such a disappearance of negatively selected variants is well known

for SNPs and it has been shown to lead to a time dependence effect on measuring primary mutation rates. It was found that rates are much higher when very recent time horizons are studied, since the negative mutations can still segregate for some time in the populations (33). We have previously shown that this effect can also be traced in mitochondrial mutation patterns of mice after island colonization (34) and we observed it here for the comparisons of gene retroposition events between the most recently diverged populations.

Our finding that most new retroCNVs are transcribed explains why they are mostly deleterious, especially since their effects would be genetically dominant through their transcripts. Accordingly, our estimate of about 20,000 events per Myr is still a minimal estimate. It is mostly based on tracing singleton events in individuals and is thus dependent on the number of individuals sequenced in a given population. For example, when we compare FR_C (N = 8) with FR_A (N = 20) we found 60 versus 136 population-specific retroCNVs (Supplementary Table S4). Hence, the number of primary retroposition events could even be much higher, which explains also why we do not reach saturation even in our extensive samples (Figure 2A).

Several of our analyses support the notion of strong negative selection acting on most new retroCNVs. In the comparison with the mutational spectra of SNP categories, we found that retroCNVs are even more deleterious than the category of the most deleterious SNPs, presumably because of their dominant effects.

Most intriguingly, our data allowed to directly showing the impact of new retroCNVs on the transcription of their parental genes. Especially the result on the strong transcriptional asymmetry bias among retroCNVs segregating in populations provides a direct clue why they may often be deleterious. Antisense RNA transcripts of retroCNVs would directly interfere with the function of the parental genes via RNA interference. While this may in a few cases have beneficial effects (13, 14) one can expect that it would mostly be deleterious. This would lead to a strong selection against highly expressed antisense retroCNVs and can thus explain why they are rare among segregating retroCNVs, or at least very poorly expressed. In our analysis of singleton retroCNV effects in populations (Table 1), we found between 22-31% having a

negative effect on the expression of their parental genes. If one assumes that the primary integration of a retroCNV copy is random with respect to the orientation of transcription, half of the singletons could be in antisense direction, *i.e.*, if the above percentage of negative effects is mostly due to antisense transcription, more than half of them are deleterious. Moreover, we have to assume that the most strongly deleterious ones are not represented in the sample, since they would be most quickly purged.

However, even sense copies could be deleterious due to dosage effects, or functional interference with their parental genes when truncated versions of the protein are produced, or through sponging regulatory micoRNAs (14). Hence, while many previous reviews on retrogenes have focused of the evolutionary potential generated by retrogenes, the apparently strongly deleterious effects have been overlooked.

Implications for human genetic disease studies

Our analysis thus suggests that the generation of retroCNV copies is a major contributor to the mutational load in natural populations. Mammalian genomes are estimated to carry up to about 1,000 deleterious SNP mutations per genome, mostly recessive ones (35), while we found around 200 retroCNV genes per mouse genome, whereby this is an underestimate as discussed above. A comparable retroCNV study in human populations (12) revealed also a very high rate of new retroCNVs, although about 3 times less (1663 retroCNV genes in house mouse populations versus 503 in human populations). However, the sequencing depth on the mouse samples is higher and our detection pipeline was further optimized. It is therefore reasonable to assume that the actual rate of retrocopy generation could be similar in humans and mice. Given their mostly dominant effect, this means that retrocopies are not less likely to cause a genetic disease than new SNP mutations. GWA studies of complex genetic diseases often find SNP associations in intergenic regions which are interpreted as regulatory variants. It may well be that such SNPs are in close linkage to an undetected retroCNV exerting a trans-regulatory influence on its parental gene and thus cause a disturbance of a genetic network. We note, however, that the variety of methods that are now

available for SNP detection or structural variation detection do not yet include specific pipelines for retroCNV analysis (36). Although there are a few known cases where retroCNVs have caused a genetic disease (3, 37), a much more systematic approach to trace such events seems warranted, given the high rate of primary retroposition events detected in our study. Hence, the focus should shift from exome sequencing studies to full-genome sequencing studies to identify the molecular underpinning of disease alleles.

Materials and Methods

Glossary and definitions

Retroposed parental gene: A gene whose processed transcript is reverse-transcribed and re-inserted into the genome as a copy lacking the introns.

Retrocopy: A genomic DNA fragment that is generated from a parental gene through retroposition (RNA-based duplication).

RetroCNV gene: A retroposed parental gene for which the gene retroposition event is polymorphic in the target species.

RetroCNV (allele): The new retrocopy generated from a retroCNV gene which is polymorphic in the target species. On condition that the new retrocopy is present in the reference genome, it is denoted as the full length retrocopy; otherwise as the insertion site of the retrocopy in the genome when full length retrocopy is absent in the reference genome. In case that multiple retrocopies are generated from the same retroCNV gene, each retrocopy is considered as a distinct retroCNV.

Reference genome sequences and gene annotation sources

We obtained the mouse reference genome sequence (mm10/GRCm38) and gene annotation data from Ensembl version 87 (38). This reference genome was built mainly based on the C57BL/6 mouse strain,

which was derived from an inbred strain of the subspecies *Mus musculus domesticus*, but also included a small fraction of genome regions admixed from other house mouse subspecies, either due to inadvertent crosses in the initial breeding phase (39), or as remnants of natural introgression patterns (40).

We also retrieved the genome assembly sequence data for two out-group sister species (SPRET_EiJ_v1: *Mus spretus*; GCA_003336285.1: *Mus spicilegus*) from the NCBI Genbank database (41). Except *Mus spicilegus* (currently only at scaffold level), all other reference genomes were assembled at almost-complete chromosomal levels (42).

Identification of retroposed parental genes in the mouse inbred lines

We directly retrieved the datasets of retroCNV genes in the inbred strains of three house mouse subspecies (*Mus mus domesticus*: WSB_EiJ_v1; *Mus mus musculus*: PWK_PhJ_v1; *Mus mus castaneus*: CAST_EiJ_v1) provided in (18), which were generated based on short read genomic sequencing dataset by using a similar computational pipeline as shown below.

Additionally, we also identified recently retroposed parental genes found in the reference genome sequences of two out-group species, *i.e*, *Mus spicilegus* and *Mus spretus*. By refining previous strategies (25, 43), we applied a computational pipeline that searches mm10 reference genome exon-exon junction libraries (as shown below) against above out-group species reference genomes by using BLAT v36 with default parameters (≥90% alignment identity) (44), and only retained uniquely mapped regions. In case at least one exon-exon junction read (≥30 bp on each side) was uniquely mapped in the out-group species genome sequences, the corresponding gene was taken as a retroposed parental gene.

Individual genomic sequencing datasets and short read alignment

We downloaded the whole-genome sequencing data from 96 wild individuals from 9 natural populations of 3 house mouse subspecies (listed in Supplementary Table S2A) from either our previously generated dataset (ENA accession: PRJEB9450, PRJEB11742, PRJEB14167 and PRJEB2176) (24), or a publicly available dataset in the European Nucleotide Archive (ENA accession: PRJEB15873). In addition, we also included the whole genome sequencing data for 9 wild individuals from 2 out-group species (ENA accession: PRJEB11513 and PRJEB11742; 1 individual from *Mus spicilegus* and 8 individuals from *Mus spretus*) for comparison (8, 24). Detailed description on these genomic sequencing data, including read length and sequencing coverage, can be found in Supplementary Table S2A.

We aligned short sequencing reads from each individual genome to the mm10 reference genome sequence (Ensembl v87) in paired-end mode by using BWA mem (v0.7.15-r1140) (45), with default parameter settings, except the penalty for a mismatch (option "-B") setting as 1, in order to compensate the sequence divergences of individuals from various populations and species. We only kept the alignment results to linear complete chromosomes. We further sorted alignment bam data by using the samtools sort function (v1.3.1) (46), and filtered PCR duplicates by using PICARD (v2.8.0) (http://broadinstitute.github.io/picard). The resulting alignment bam files were used for the further analysis. The sequence divergence to mm10 reference genome for each sequencing dataset was calculated from the above alignment bam files (mappings with mapping quality ≥20) by using samtools stats function (v1.3.1) (46).

Identification of polymorphic retroCNV genes

Based on previous approaches (18, 19, 25), we developed a refined computational pipeline for the discovery of retroCNV genes based on the short read sequencing datasets from individual genomes (Supplementary Figure S1). This pipeline combines both exon-exon and exon-intron-exon junction read mapping strategies to identify gene retroposition events, and the discovery process is independent of the

presence of newly generated retrocopies in the reference genome. The reliability of this approach has been experimentally verified with PCR techniques in a couple of previous studies (19, 25).

We firstly generated an exon-exon junction database, by extracting all possible consecutive exon-exon junction sequences (100bp from each side, or shorter for small exons, but ≥50bp) for all 20,378 protein coding genes with ≥2 exons annotated in the mm10 reference genome based on Ensembl v87 (38). Similar to the mapping procedure to the reference genome, short sequencing reads from each individual genome were aligned to the exon-exon junction database by using BWA mem (v0.7.15-r1140) (45), with default parameter settings except the penalty for a mismatch (option "-B") setting as 1. Given the rather short length (≤ 200bp) of exon-exon junction sequences, the alignment to the exon-exon database was done in single-end mode. We defined uniquely mapped reads (for both exon-exon junction dataset mapping and reference genome mapping) as the ones satisfying both of the two criteria: 1) mapping quality ≥20; 2) the difference between the best and second best alignment scores ≥5 (25).

A retroposed parental gene is called, in case that both the intron loss events (*i.e.*, exon-exon junction mapping) and the presence of a parental gene (*i.e.*, exon-intron and intron-exon junction mapping) can be observed in the same individual sequencing dataset (25). On the basis of our simulation results in Supplementary Text S1, for the calling of both events, we required at least 2 distinct supporting reads (to be adjusted according to individual genomic sequencing dataset as shown below) from the same individual genome, that span at least 30bp on each side of exon-exon/exon-intron-exon junction.

Given the distinct features (*i.e.*, read length, sequencing coverage, and divergence to the mm10 reference genome) of the short read sequencing dataset for the individuals from each population (Supplementary Table S1), we further conducted a comprehensive simulation analysis to tailor the discovery pipeline for each population individuals with different parameter settings (Supplementary Text S1): 1) alignment identity; 2) minimum spanning length on each side of junction; 3) minimum number of supporting reads. The optimized parameter setting for the discovery of retroposed parental genes in each individual genome is provided in Supplementary Table S2A.

In this study, we only focused on house mouse specific retroCNV genes, which were defined as retroposed parental genes detected in at least one house mouse individual, while absent in both the out-group species (*Mus spicilegus* and *Mus spretus*), neither in wild individual genome sequencing data nor in the inbred reference genomes.

Detection of retroCNV alleles

Based on the above detected house mouse specific retroCNV genes, we performed detection of retroCNV alleles at individual genome level. For the retrocopies present in the mm10 reference genome (with sequence identity >=95% to their parental gene), we used the ones annotated in RetrogeneDB version 2 (32). We searched for proper paired-end alignments (with both correct orientation and expected mapping distance) from each house mouse individual sequencing dataset, that have one read uniquely mapped to the flanking region of the annotated retrocopy and the other read mapped within the focal retrocopy region (unique mapping not required). Two criteria were required to define the unique alignments to the flanking region: 1) mapping quality ≥20; 2) the difference between the best and second-best alignment scores ≥5. A presence of retrocopy is called, if there are at least two supporting reads on both flanking sides or at least four supporting reads on either flanking side.

In order to detect the insertion sites for the retroCNVs that are absent in the mm10 reference genome (*i.e.*, full length retrocopy is not available), we applied a discordant alignment based approach by using the above paired-end read alignments from each individual genome sequencing data (12, 17). We searched for paired-end alignments in proper orientation, and with one read uniquely mapped (*i.e.*, anchor read) within exonic sequences of the parental gene and the other read uniquely mapped to a distinct genomic region, *i.e.*, on a different chromosome or on the same chromosome but with an unexpected mapping distance (> 2 x average insert size of the paired-end library). Similarly, a unique alignment was required to meet both criteria: 1) mapping quality ≥20; 2) the difference between the best and second-best alignment score ≥5. The procedure of the clustering of the above discordant alignments was performed by following (12, 17),

with 500 bp as the cut-off for average linkage distance to stop clustering. The insertion site was taken as the middle point of the cluster. An insertion site would be considered to be valid, if there are at least two supporting reads on both sides (*i.e.*, strands), or at least four supporting reads on either side (12).

For retroCNVs of which the insertion sites inferred from different individuals that were from the same retroCNV parental gene, we used 1kb as the clustering distance threshold (also required on the same chromosomal strand) to define the same gene retroposition events. Accordingly, the multiple retrocopies (with distinct insertion sites) from the same retroCNV gene were taken as distinct retroCNVs.

Transcriptional profiling of retroCNVs

We used two different sets of transcriptomic sequencing data for the transcriptional profiling of retroCNVs: 1) one non-strand-specific RNA-Seq dataset from our previously published data (24); 2) one strand-specific RNA-Seq dataset newly generated in the present study. The detailed description about these two datasets can be found in Supplementary Table S2B and S2C.

The non-strand-specific RNA-Seq dataset used the same individuals which are included in the above whole genome sequencing data of *M. m. domesticus* (24). Four individuals in Iran population were excluded for analysis due to imperfect match between genomic/transcriptomic datasets. We downloaded the transcriptomic sequencing data of 10 tissue samples (*i.e.*, Brain, Gut, Heart, Kidney, Liver, Lung, Muscle, Testis, Spleen, Thyroid) from 20 mice individuals of 3 natural populations (Germany population (GE): 8 individuals; France Massif Central population (FR_C): 8 individuals; Iran population (IR): 4 individuals). With a few exceptions, most of these individuals have expression profiling data from all these 10 tissues (Supplementary Table S2B).

To study the orientation-dependent transcription of the retroCNVs, we generated a strand-specific RNA-Seq dataset from 10 male individuals of the France Massif Central population (FR_C). The (whole) brains, hearts, livers (right medial lobes), (right) kidneys, and (right) testes from ten 24-26 weeks old males were

carefully collected and immediately frozen with liquid nitrogen. Total RNAs were purified using RNeasy 96 Universal Tissue Kit (Catalog no. 74881), and sent to Competence Centre for Genomic Analysis in Kiel for stranded mRNA library preparation and sequencing on Illumina NovaSeq 6000 (2 × 150 bp). For both datasets, fastq files were trimmed with Trimmomatic (0.38) (47), and only paired-end reads passed filtering process were used for further analyses.

In order to accurately quantify expression levels, we focused on the recent retrocopies present in the mm10 reference genome (originated from house mouse specific retroCNV genes, and with sequence identity ≥ 95% compared with their parental genes), for which the information was directly inferred from RetrogeneDB version 2 (32). As the recently originated retrocopies are usually highly similar to their parental genes, we implemented an effective-length based approach to calculate their expression values (25). Firstly, we simulated 5000X 100bp paired-end sequencing reads on the basis of retrocopies' sequences (including 500bp up/downstream flanking regions) with Art_illumina (v 2.5.8) (48), and then re-mapped them to the combined reference sequences (mm10 reference genome + transcript sequences of the parental genes of these retrocopies) using BWA mem (v0.7.15-r1140). In order to distinguish from parental genes, we applied a high penalty for the mismatch (or insertion/deletion, -B 10, - O [10,10]) for the mapping process. Only the uniquely aligned reads (a. mapping quality ≥ 20; b. perfectly match in the retrocopy regions, but at least one site mismatch for the second-best alignment to the parental gene, AS-XS ≥ 11) were included. We calculated the effective length of each retrocopy as the number of uniquely mapping locations in the retrocopy region, and only kept 59 retrocopies (50 of them are present in FR_C population, thus were used for strand-specific RNA-Seq dataset analysis) with non-zero effective length for further analysis. Secondly, we pooled the transcriptomic sequencing data of the retroCNV carriers of the same population individuals for each tissue, and mapped them to the same combined reference sequences with the same BWA mem pipeline as mentioned above. For the strand-specific RNA-Seq dataset, we counted the reads of all those 10 tested individuals from the two strands (sense and antisense strands relative to their parental genes) separately. The normalized FPKM value for each retrocopy from

each strand within each tissue was calculated on the basis of the above computed effective length of each retrocopy.

The impact on parental gene expression from singleton retroCNVs

Given the limit number of mice individuals (N=4) with matched genomic/transcriptomic sequencing dataset in Iran population (Supplementary Table S2B), we restricted this analysis to the animals with singleton retroCNVs in the FR_C (71 singleton retroCNV genes) and GE (74 singleton retroCNV genes) populations (N=8 for both popuations) only, *i.e.*, the cases where only one individual of a given population carried the retroCNVs. In case that one singleton retroCNV gene has multiple retroCNV alleles (*i.e*, ≥2 detected insertion sites), the effects of all these alleles were combined (Supplementary Table S7), since it is unlikely to separate their own effects. Moreover, the singleton retroCNV genes that have no detectable insertion site (likely landing in the repetitive genomic region) were also included for analysis here.

To profile the tissue expression pattern of retroCNV parental genes, we mapped the non-strand-specific RNA-Seq reads from each individual/tissue from FR_C and GE population individuals to mm10/GRCm38 reference genome sequence with HISAT2 (2.1.0) (49), 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 (1.6.3) (50). Finally, we calculated the expression level in FPKM (Fragments Per Kilobase of transcript per Million mapped reads) for each annotated coding gene within each individual/tissue on the basis of gene's transcript length and the mapped sequencing dataset size from each individual/tissue.

To explore the impact on parental gene expression, we compared the expression level of the focal parental gene in the mice individual with the singleton retroCNV with those in the reminder of seven retroCNV non-carriers from the same population. We used Wilcoxon rank sum test to calculate the significance (p-

value) of the parental gene expression change between singleton retroCNV carrier and non-carriers, and the significant expression changes after multiple testing corrections (FDR ≤ 0.05) were also included. As the upregulation of parental gene expression could be due to the additional reads from the retroCNV itself, we focused this analysis on the downregulation of the parental gene expression only.

Genic copy number variation (CNV) calling

We predicted CNV calls for each individual based on the above PCR-duplicates filtered alignment bam files by using the sequencing read depth based approach implemented in the program CNVnator v0.3.3 (51), as the reliability of this approach has been experimentally confirmed (27). We chose the optimal bin size for each individual, such that the ratio of the average read-depth signal to its standard deviation was between 4 and 5 (51). Bin size ranged from 100-1,500bp and was inversely proportional to genome coverage. Following the convention in (27), we discarded the CNV calls below 1kb in length and intersecting annotated gaps in the reference genome, and defined "genic CNVs" as the genes for which at least one whole transcription unit is completely contained within CNV calls. Similarly, house mouse specific "genic CNVs" were defined as the ones (duplications or deletions) can be found in at least one house mouse individuals, but not in any individual from outgroup species.

SNP variants calling based on genomic sequencing reads

We followed the general GATK version 3 Best Practices (52) to call variants. Specifically, we realigned the above PCR-duplicates filtered alignment bam files around the indels (with flag "PASS") detected from the Mouse Genome Consortium (v5) (39) with GATK (v3.7), and recalibrated base quality scores with by using SNP variants (with flag "PASS") founded in the Mouse Genome Consortium (version 5) (39) to get analysis-ready reads.

In the first, we called raw genetic variants for each individual using the HaplotypeCaller function in GATK (v3.7), and then jointly genotyped genetic variants for all the individuals using GenotypeGVCFs function in GATK (v3.7). We only retained bi-allelic SNP variants that passed the hard filter "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 3.0", and dropped the variants with missing calling value in any individual. We only kept the SNP variants with unambiguous ancestral states in out-group species (*i.e.*, same homozygous genotype for all 9 tested individuals from 2 out-group species), while with alternative allele in house mouse individuals for further analysis.

We predicted the functional effects of each SNP by using Ensembl VEP v98.2 (30), based on the gene annotation data from Ensembl version 87 (38). Consistent with Ensembl variation annotation (38), we categorized these SNPs into four groups given their predicted impacts: 1) High effect - SNPs causing the gain/loss of start/stop codon or change of the splicing acceptor/donor sites; 2) Moderate effect - SNPs resulting in a different amino acid sequence; 3) Low effect - SNPs occurring within the region of the splice site, changing the final codon of an incompletely annotated transcript, changing the bases of start/stop codon (while start/terminator remains), or where there is no resulting change to the encoded amino acid; 4) Modifier effect - SNPs occurring within the genes' non-coding regions (including UTR, intron, up/downstream), or intergenic regions.

The frequency spectrum of house mouse retroCNVs was compared with the site frequency spectrum of SNPs from the four above defined categories. We quantified the distances between spectrum distributions by using two sided Kolmogorov-Smirnov tests, and calculated the statistical significances of the fraction of individual private variants by using Fisher's exact tests.

Correlation between retroCNV pattern and population differentiation

We checked the association between retroCNVs pattern and population differentiation by conducting the following four lines of analyses: 1) Calculation and comparison on the fractions of shared retroCNVs

between all possible pair-wise combinations from all individuals, individuals from different subspecies, different populations but same subspecies, and from the same populations; 2) Principal component analysis on the individual retroCNV landscape with R package "ggfortify" v0.4.8 (https://cran.r-project.org/web/packages/ggfortify/index.html); 3) Heat map analysis on the distance between pair-wise individual comparison on retroCNV landscape with python package "matplotlib" v3.0.1 (https://matplotlib.org/). The distance between pair-wise individuals was defined as the fraction of different retroCNVs of these 2 individuals to be compared.

We exploited a retroCNV frequency based approach to detect population differentiating retroCNVs, similar to that was used for retroCNV parental genes (12). Each retroCNV was assigned a population differentiation index (PDI), a measure equivalent to the fixation index (53), with the following equations:

$$\mathrm{PDI} = \frac{\mathrm{P}(1-\mathrm{P}) - \sum_{i=1}^{N} \mathrm{F}i * \mathrm{P}i * (1-\mathrm{P}i)}{P(1-P)} \quad (1)$$

$$\mathrm{F}i = \mathrm{N}i \,/\, \sum_{i=1}^{N} \mathrm{N}i \quad (2)$$

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, and significant retroCNVs after multiple testing corrections (FDR $\leq$ 0.05) were taken as candidate population differentiation markers.

Ethics statement

This study did not include in vivo experiments on animals and the species used are not protected species. All the mouse genomic sequencing data and most of the transcriptomic sequencing data were directly retrieved from public available datasets, which were mostly previously generated from our laboratory. The strand-specific transcriptomic sequencing data generated in this study was from the progeny of breeding of wild-derived individuals (from which a subset of genomic/transcriptomic sequencing data was generated previously) from Massif Central in France. The 10 males used in this study were taken from the breeding stocks kept in our mouse house and they were humanely killed via CO_2 asphyxiation before organs were prepared. Maintenance and handling of the animals of the breeding stock were conducted in accordance with German animal welfare law (Tierschutzgesetz) and FELASA guidelines. Permits for keeping mice were obtained from the local veterinary office "Veterinäramt Kreis Plön" (permit number: 1401-144/PLÖ-004697).

Data access

The raw strand-specific RNA-Seq data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under study accession number PRJEB36991.

Acknowledgements

We appreciate Peter Keightley and Guy Reeves for reading through the manuscript and providing helpful comments. We are grateful to Julien Dutheil for valuable suggestion on statistical analysis. We thank Yuanxiao Gao for valuable suggestions on data presentation and visualization. We appreciate the laboratory members for helpful discussions and suggestions. Computing was supported by the Wallace HPC cluster of the Max-Planck Institute for Evolutionary Biology. This work was supported by institutional funding through the Max Planck Society to DT.

Author contributions: W.Z. and D.T. designed the study, interpreted the data and wrote the paper. W.Z., C.X., K.U. and Y.Z. analyzed the data and performed the statistical analysis. C.X. collected the animals and generated the new sequencing dataset. All authors read and approved the final manuscript.

Competing interests

The authors declare that no competing interests exist.

References

1. M. Long, E. Betrán, K. Thornton, W. Wang, The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).

2. H. Kaessmann, N. Vinckenbosch, M. Long, RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **10**, 19–31 (2009).

3. C. Casola, E. Betrán, The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? *Genome Biol. Evol.* **9**, 1351–1373 (2017).

4. P. Jeffs, M. Ashburner, Processed pseudogenes in Drosophila. *Proc. R. Soc. B Biol. Sci.* **244**, 151–159 (1991).

5. Z. Zhang, N. Carriero, M. Gerstein, Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* **20**, 62–67 (2004).

6. S. Djebali, *et al.*, Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

7. M. J. Hangauer, I. W. Vaughn, M. T. McManus, Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet.* **9**, e1003569 (2013).

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

9. M. V. Han, J. P. Demuth, C. L. McGrath, C. Casola, M. W. Hahn, Adaptive evolution of young gene duplicates in mammals. *Genome Res.* **19**, 859–867 (2009).

10. H. Innan, F. Kondrashov, The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).

11. H. Kaessmann, Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).

12. Y. Zhang, S. Li, A. Abyzov, M. B. Gerstein, Landscape and variation of novel retroduplications in 26 human populations. *PLoS Comput. Biol.* **13**, e1005567 (2017).

13. O. H. Tam, *et al.*, Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).

14. T. Watanabe, *et al.*, Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539–543 (2008).

15. A. C. Marques, J. Tan, C. P. Ponting, Wrangling for microRNAs provokes much crosstalk. *Genome Biol.* **12**, 132 (2011).

16. F. C. P. Navarro, P. A. F. Galante, A genome-wide landscape of retrocopies in primate genomes. *Genome Biol. Evol.* **7**, 2265–2275 (2015).

17. A. Abyzov, *et al.*, Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res.* **23**, 2042–2052 (2013).

18. A. D. Ewing, *et al.*, Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **14**, R22 (2013).

19. D. R. Schrider, *et al.*, Gene Copy-Number Polymorphism Caused by Retrotransposition in Humans. *PLoS Genet.* **9**, e1003242 (2013).

20. S. R. Richardson, C. Salvador-Palomeque, G. J. Faulkner, Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population. *BioEssays* **36**, 475–481 (2014).

21. J. L. Guénet, F. Bonhomme, Wild mice: An ever-increasing contribution to a popular mammalian model. *Trends Genet.* **19**, 24–31 (2003).

22. M. Phifer-Rixey, M. W. Nachman, Insights into mammalian biology from the wild house mouse Mus musculus. *Elife* **4**, e05959 (2015).

23. S. Ihle, I. Ravaoarimanana, M. Thomas, D. Tautz, An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol. Biol. Evol.* **23**, 790–797 (2006).

24. B. Harr, *et al.*, Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Sci. Data* **3**, 160075 (2016).

25. S. Tan, *et al.*, LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res.* **26**, 1663–1675 (2016).

26. D. R. Schrider, K. Stevens, C. M. Cardeño, C. H. Langley, M. W. Hahn, Genome-wide analysis of retrogene polymorphisms in Drosophila melanogaster. *Genome Res.* **21**, 2087–2095 (2011).

27. Ž. Pezer, B. Harr, M. Teschke, H. Babiker, D. Tautz, 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 (2015).

28. E. A. Hardouin, *et al.*, Eurasian house mouse (Mus musculus L.) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. *BMC Evol. Biol.* **15** (2015).

29. T. Cucchi, *et al.*, Tracking the Near Eastern origins and European dispersal of the western house mouse. *Sci. Rep.* **10**, 8276 (2020).

30. W. McLaren, *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

31. F. N. Carelli, *et al.*, The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**, 301–314 (2016).

32. W. Rosikiewicz, *et al.*, RetrogeneDB-a database of plant and animal retrocopies. *Database (Oxford).* **2017**, bax038 (2017).

33. S. Y. W. Ho, *et al.*, Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**, 3087–3101 (2011).

34. E. A. Hardouin, D. Tautz, Increased mitochondrial mutation frequency after an island colonization: Positive selection or accumulation of slightly deleterious mutations? *Biol. Lett.* **9**, 20121123 (2013).

35. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).

36. S. S. Ho, A. E. Urban, R. E. Mills, Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).

37. J. Ciomborowska, W. Rosikiewicz, D. Szklarczyk, W. Makałowski, I. Makałowska, "Orphan" retrogenes in the human genome. *Mol. Biol. Evol.* **30**, 384–396 (2013).

38. F. Cunningham, *et al.*, Ensembl 2019. *Nucleic Acids Res.* **47 (D1)**, D745–D751 (2019).

39. T. M. Keane, *et al.*, Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).

40. F. Staubach, *et al.*, Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). *PLoS Genet.* **8**, e1002891 (2012).

41. D. A. Benson, *et al.*, GeneBank. *Nucleic Acids Res.* **41**, D36-42 (2013).

42. J. Lilue, *et al.*, Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**, 1574–1583 (2018).

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

44. W. J. Kent, BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

45. L. Heng, D. Richard, Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760 (2009).

46. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

47. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

48. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: A next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

49. D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

50. Y. Liao, G. K. Smyth, W. Shi, FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

51. A. Abyzov, A. E. Urban, M. Snyder, M. Gerstein, CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).

52. G. A. Van der Auwera, *et al.*, From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).

53. K. E. Holsinger, B. S. Weir, Genetics in geographically structured populations: Defining, estimating and interpreting FST. *Nat. Rev. Genet.* **10**, 639–650 (2009).

Supplementary material for:

The mutational load in natural populations is significantly affected by high primary rates of retroposition

Supplementary Text:

Text S1 Parameter optimization for the computational discovery of retroposed parental genes

Text S2 RetroCNVs acting as house mouse population differentiation markers

Supplementary Figures:

Figure S1 Computational pipeline for the discovery of retroposed parental genes

Figure S2 False positive discovery rates of retroposed parental genes based on the different spanning sizes cutoffs

Figure S3 A flow chart to show the parameter optimization of the retroposed parental gene detection pipeline

Figure S4 Recall rates of the discovery of retroposed parental genes with uniform setting of parameters

Figure S5 Recall rates of the discovery of retroposed parental genes with optimized parameter settings

Figure S6 Fraction of shared retroCNVs between pair-wise individuals comparison

Figure S7 Correlation between the house mouse specific retroCNV landscape and population differentiation

Figure S8 Distribution of the number of detected retroCNV alleles for retroCNV genes

Figure S9 Fraction of overlap between retroCNV genes and genic CNVs across house mouse populations

Figure S10 Distribution of gene retroposition events across house mouse natural populations

Supplementary Tables (each in separate files):

Table S1 Summary statistics of genomic short read sequencing datasets

Table S2 List of genomic and transcriptomic sequencing datasets used in this study

Table S3 List of gene retroposition events detected in all house mouse individuals

Table S4 Population differentiation information of all house mouse retroCNVs

Table S5 RetroCNV expression based on the non-strand specific RNA-Seq dataset

Table S6 Summary of RetroCNV expression based on the non-strand specific RNA-Seq dataset

Table S7 The impact on parental gene expression for singleton retroCNVs

Table S8 RetroCNV expression on the sense/antisense strand based on strand-specific RNA-Seq dataset

Text S1 Parameter optimization for the computational discovery of retroposed parental genes

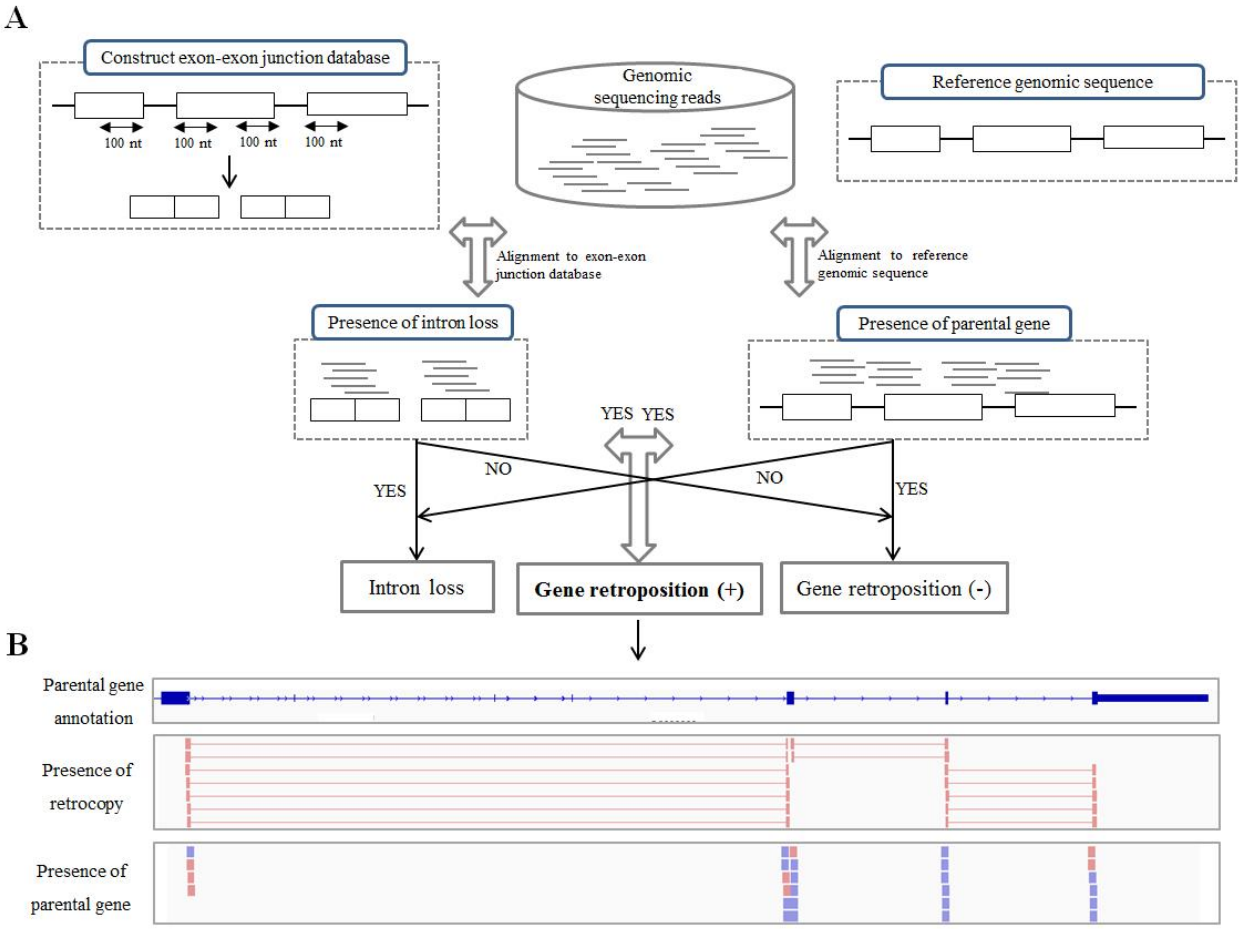

Figure S1 Computational pipeline for the discovery of retroposed parental genes. (A) shows a simplified flow chart of our computational calling pipeline. (B) shows an example of gene retroposition event in the view of an IGV snapshot.

The genomic sequencing datasets from wild mouse individuals varied in read length, sequencing coverage and sequence divergence to the mm10 reference genome (Supplementary Table S1 and Table S2A). These distinct features needed to be considered for the parameter optimization of the discovery of retroposed parental genes. Several parameters were included in the discovery pipeline: 1) criteria to define unique read mapping; 2) spanning read length (both exon and intron side) to define reliable supporting evidence; 3) number of supporting evidences to define "true" gene retroposition event.

In order to detect recent retroposed parental genes in each individual genome, only mapped reads with alignment identity ≥95% and mapping quality MapQ ≥20 were included, and the alignment identity parameter further adjusted based on individual genome sequence divergence to mm10 reference genome (see below text). Furthermore, a customized Perl script was implemented to compare the alignment scores of multiple-hit reads and retained only the uniquely-mapping reads if the difference between the best and second best alignment score is ≥5 (1).

To determine the optimal spanning read length, we searched the built exon-exon junction database against mm10 reference genome by using BLAT v36 with default parameters (2), and only kept the returning searching hits with ≥95% alignment identity. Furthermore, we selected alignment results for those spanning at least 10bp, 20bp, 30bp, 40bp and 50bp on both sides of consecutive exons. The corresponding genes of these aligned exon-exon junctions were putative retroposed parental genes. In case that one retroposed parental gene was not included in any of those three published resources: RetrogeneDB (v2) (3), UCSC (RetroV6) (4) and GENCODE (v20) (5), it was considered

as a false positive prediction. As shown in Supplementary Material Figure S2, the false discovery rate (FDR) stabilized at spanning size of 30bp (~ 2.5%), with no major reduction of FDR as the spanning size increased. Therefore, this size of 30bp was set as the basis of threshold selection of spanning read length for both exon and intron side. The relative longer read length of our sequencing datasets permitted spanning size threshold that was much larger than previous reports (15bp in (1), 5bp in (6)), and this helped to define a more reliable retroposed parental gene dataset. This parameter was further adjusted according to the features of individual sequencing datasets (see text below).

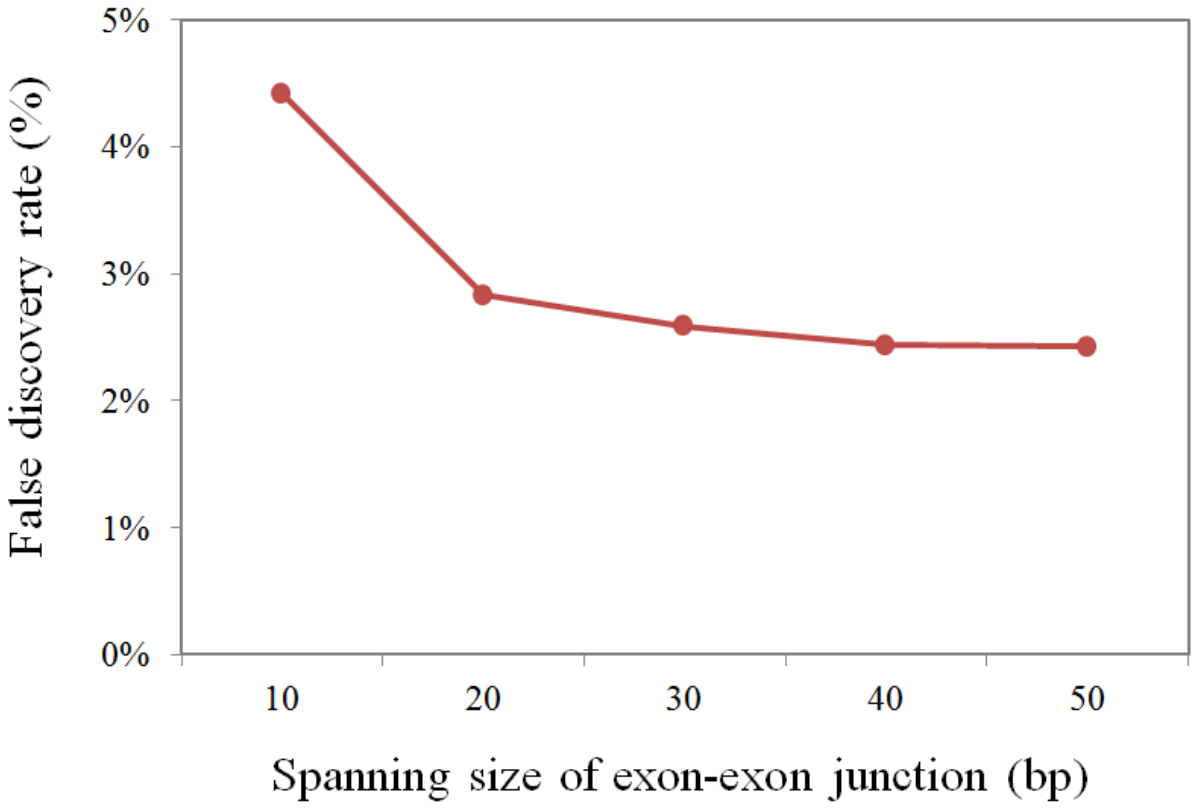

Figure S2 False positive discovery rates of retroposed parental genes based on the different spanning sizes cutoffs. False discovery rate was calculated as the number of predicted retroposed genes that were not included in any of those three previous annotation datasets (RetrogeneDB v2, UCSC Retro V6, and GENCODE v20), divided by the total number of predicted retroposed parental genes.

In order to assess the potential influence of sequencing dataset features on the discovery of retroposed parental genes, we conducted a systematic simulation analysis by incorporating various aspects of sequencing datasets: 1) sequence divergence to mm10 reference genome; 2) sequencing read length; 3) sequencing depth (Supplementary Material Figure S3). We randomly selected 1000 consecutive exon-exon junction sequences (corresponding to 1000 "putative" retroposed parental genes) and then randomly inserted them into mm10 reference genome sequences. On the basis of real individual sequencing data (Supplementary Table S1), four types of random mutation rate (u=0.00, 0.01, 0.02, 0.03) at single nucleotide sites, two types of sequencing read length (75bp/100bp) and three types of sequencing depth (10X, 20X and 30X) were simulated to generate in total 24 distinct sequencing datasets. The simulation of genomic sequencing datasets was performed by using Art_illumina (v 2.5.8) (7), with pair-end read simulation mode (Insert size for 75 bp reads: 150bp; Insert size for 100 bp reads: 200bp). The above computational pipeline (Supplementary Material Figure S1) was applied to identify retroposed parental genes from these 24 simulated genomic sequencing datasets.

As expected, in case that uniform setting of parameters (Alignment identity ≥95%; Spanning read length ≥30bp; Number of supporting evidences ≥3) were applied, genomic sequencing datasets with longer read length and higher sequencing depth tended to have larger recall rate (sensitivity) (Supplementary Material Figure S4), while mutation rates (in relation to sequence divergence) only played minor roles after increasing to 0.03. Therefore, the discovery parameters were adjusted with the rule of lowering strictness for datasets with shorter read length, larger mutation rate and lower sequencing depth (Supplementary Material Figure S3). After the optimization of parameters, relative constant recall rates (≥95%) were observed for all 24 simulated sequencing datasets (Supplementary Material Figure S5). Accordingly, the detection parameters for the real sequencing datasets were also adjusted based on the same rules (Supplementary Table S2A). It was less likely that the adjustment of detection parameters could introduce more false positive discoveries, with the fact that most real genomic sequencing datasets have either relatively long read length or high sequencing coverage (Supplementary Table S2A).

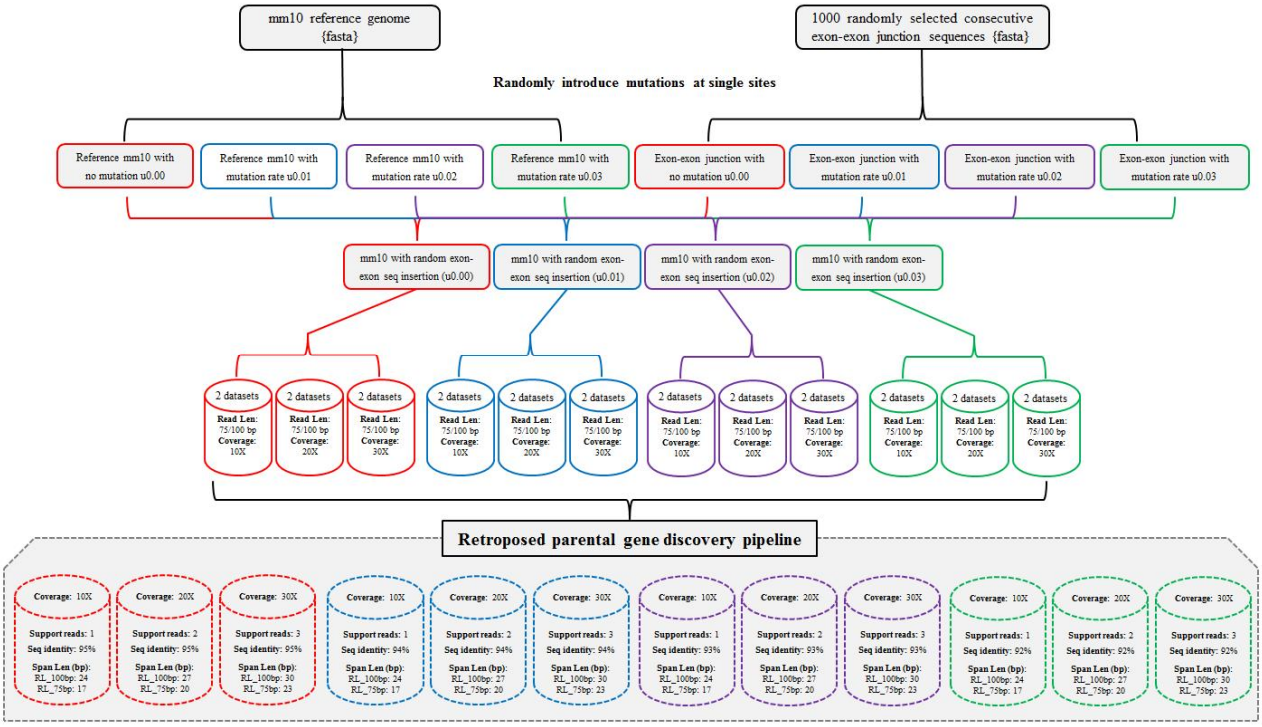

Figure S3 A flow chart to show the parameter optimization of the retroposed parental gene detection pipeline. In total, 24 genomic read sequencing datasets were simulated on the basis of the features of real individual sequencing datasets. The bottom panel shows the optimized parameters for the discovery of retroposed parental genes for different genomic sequencing datasets.

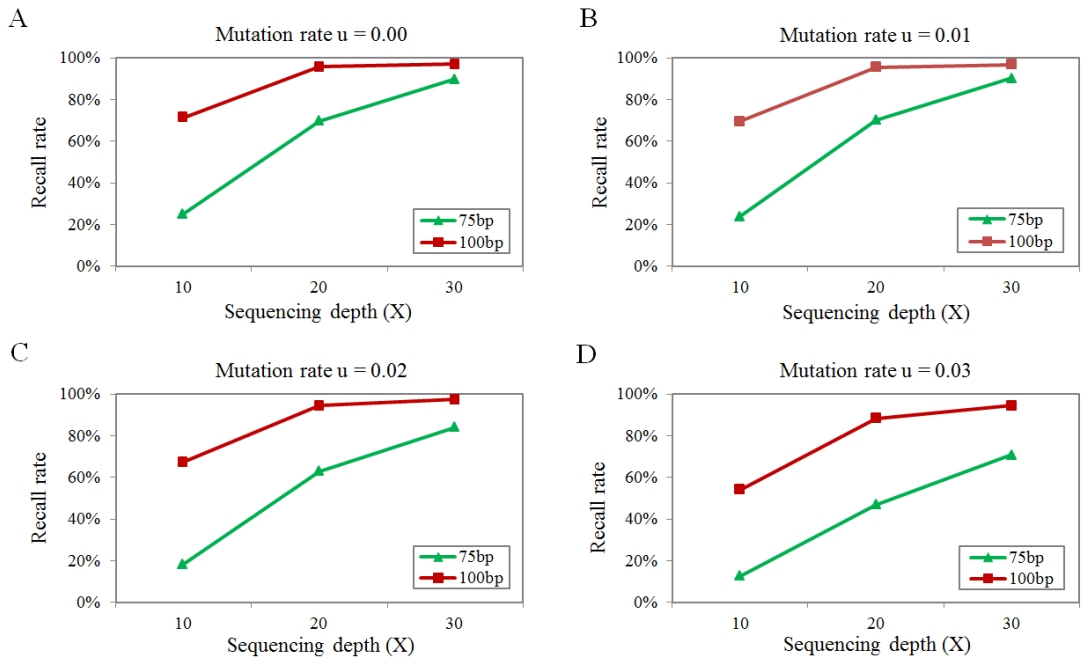

Figure S4 Recall rates of the discovery of retroposed parental genes with uniform setting of parameters. For all simulated genomic sequencing datasets, the same settings of parameters were applied: Alignment identity ≥95%; Spanning read length ≥30bp; Number of supporting evidences ≥3. The recall rate for each simulated genomic sequencing dataset was calculated as the fraction of identified 'true' retroposed parental genes discovered from 1,000 random simulations.

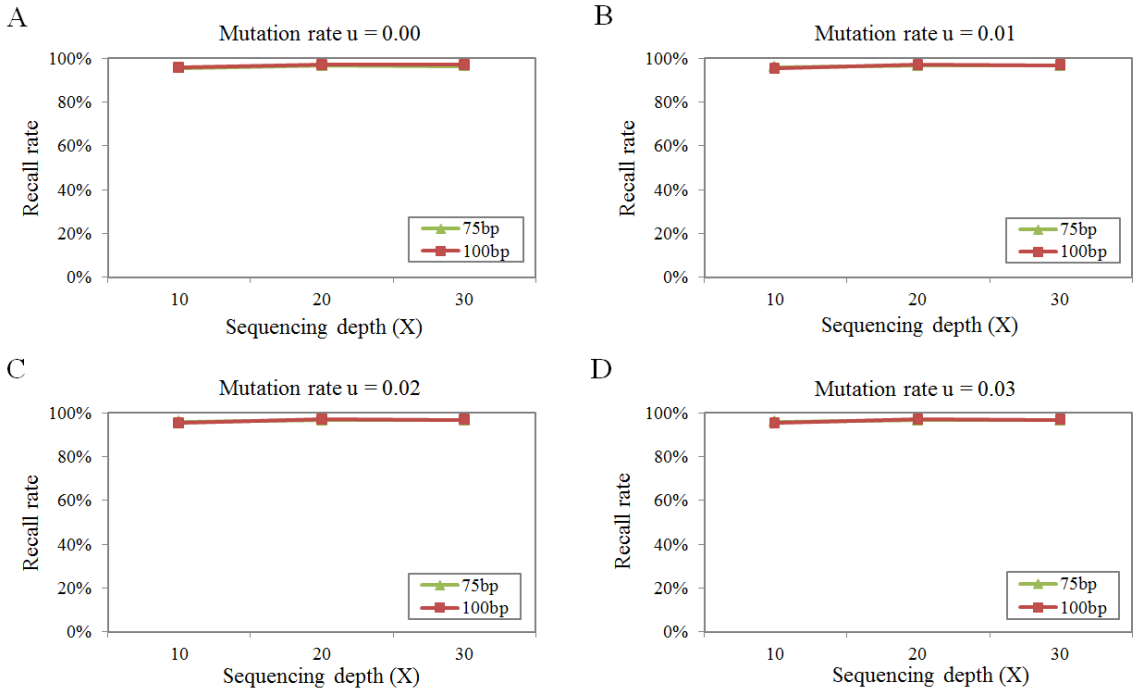

Figure S5 Recall rates of the discovery of retroposed parental genes with optimized parameter settings. The optimized discovery parameter setting for each simulated genomic sequencing dataset is provided in the bottom panel of Supplementary Material Figure S3. The recall rate for each simulated genomic sequencing dataset was calculated as the fraction of identified 'true' retroposed parental genes from 1,000 random simulations.

Text S2 RetroCNVs acting as house mouse population differentiation markers

Given their fast turnover patterns, we asked whether the retroCNVs could also be effective markers to differentiate house mouse natural populations. Pair-wise individual comparison on the sharing of retroCNVs showed indeed a more similar retroCNV landscape for individuals from the same subspecies/population (Supplementary Material Figure S6), indicating the potential to infer population relationships. Based on the plotting on the top two principle axes from retroCNVs, we observed clear clusters of the individuals among the three different sub-species, as well as some well-diverged populations from the same sub-species (Supplementary Material Figure S7A). Heat map plotting based on the pair-wise comparison of individual retroCNVs provided a finer resolution to separate all house mouse individuals from different natural populations (Supplementary Material Figure S7B).

Next, we performed a population retroCNV frequency based approach (8) to detect the retroCNV polymorphisms as differential population markers (see Methods). After applying a false discovery rate (FDR) threshold of 0.05, we detected in total 1,828 retroCNVs (~90% of all house mouse specific retroCNVs) that could potentially differentiate house mouse natural populations (Supplementary Table S4).

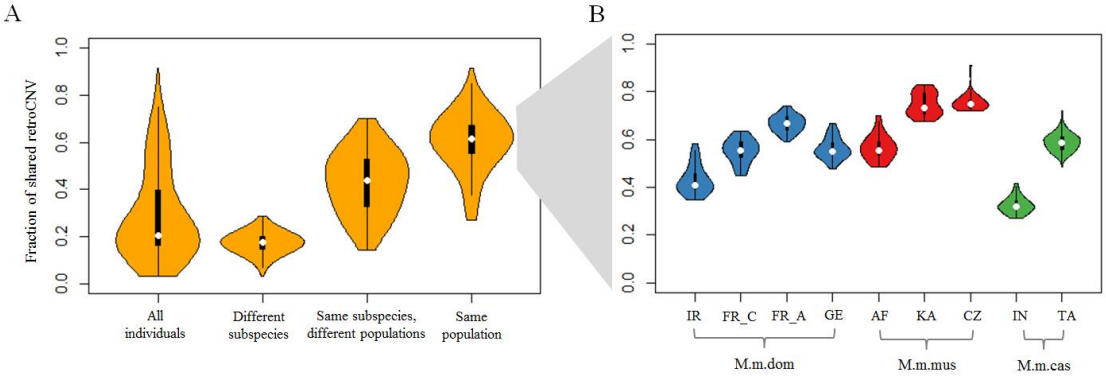

Figure S6 Fraction of shared retroCNVs between pair-wise individuals comparison. The fraction of shared retroCNVs between each pair of individuals was defined as the number of overlapping genes divided by the average number of detected retroCNVs of the two individuals. The lower fraction of sharing of retroCNVs for the Indian population is attributed to a more geographically diverged sampling scheme. 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.

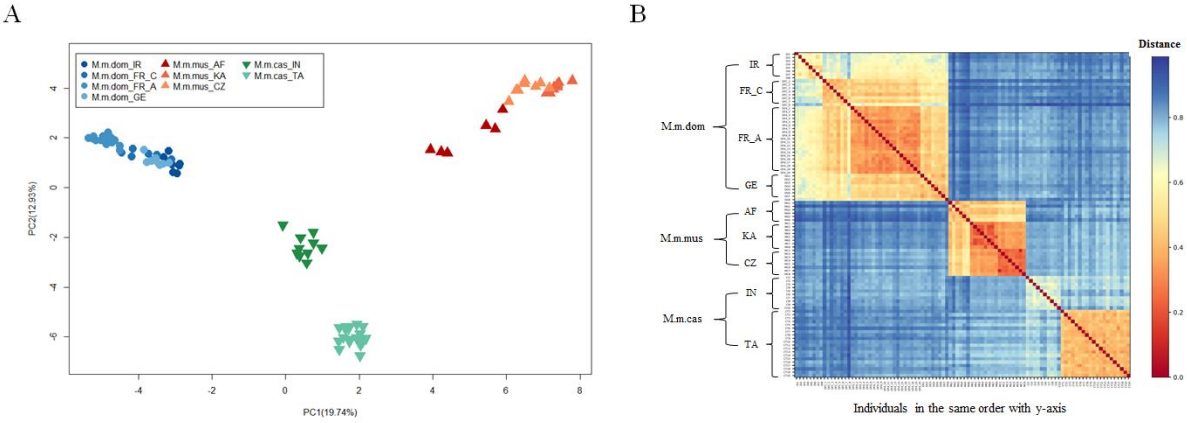

Figure S7: Correlation between the house mouse specific retroCNV landscape and population differentiation. (A) Projection of the first two principal components (PCs) of retroCNV variation in house mouse wild-derived individuals. (B) Heat map on the distance between pair-wise individual comparison on retroCNV landscape of all sampled house mouse wild-derived individuals. The distance between pair-wise individuals was defined as the fraction of different retroCNVs of these 2 individuals to be compared. The abbreviations for geographic regions follow Figure 1 / Supplementary Material Figure S6.

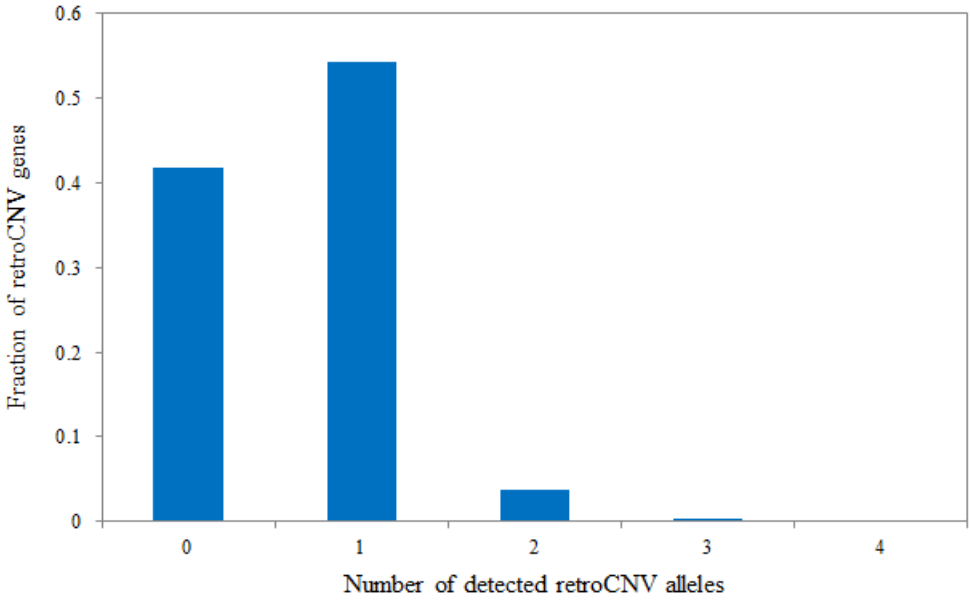

Figure S8 Distribution of the number of detected retroCNV alleles for retroCNV genes. The calculation on the number of retroCNV alleles for each retroCNV gene was based on the data provided in Supplementary Table S3. It was based on the number of retrocopies (present in the mm10 reference genome) or insertion sites (full-length retrocopies absent in the mm10 reference genome) found in the genome of a given animal. 0 means that the insertion site for a given retroCNV gene could not be uniquely localized likely due to having landed in a repetitive genome region.

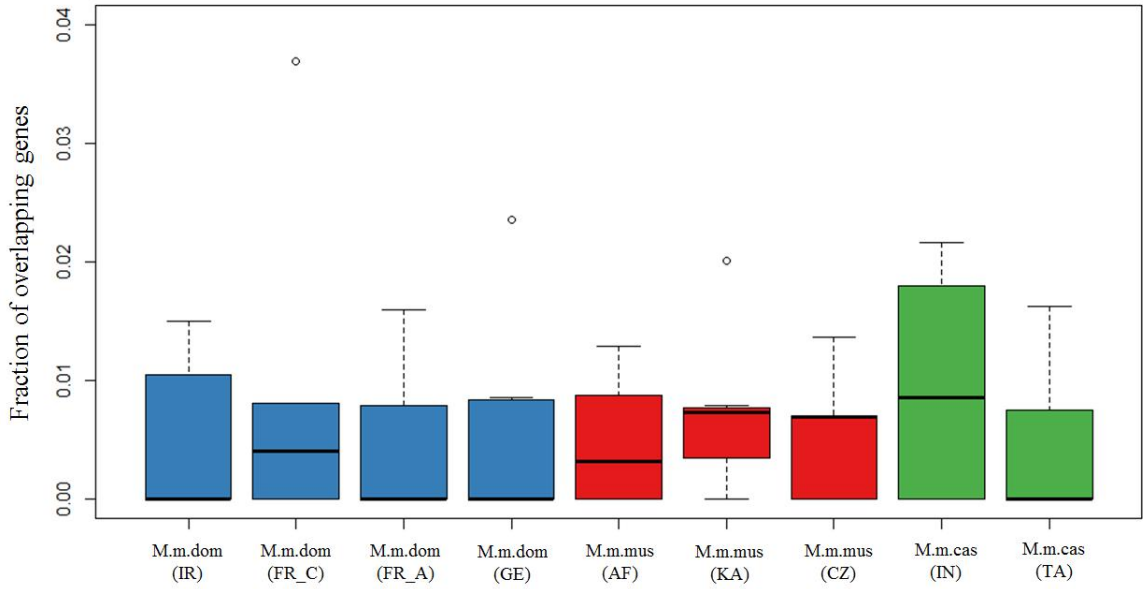

Figure S9 Fraction of overlap between retroCNV genes and genic CNVs across house mouse populations. The fraction of overlap for each individual was defined as the number of overlapping genes divided by the average number of detected retroCNV genes and genic CNVs. The abbreviations for geographic regions follow Figure 1 / Supplementary Material Figure S6.

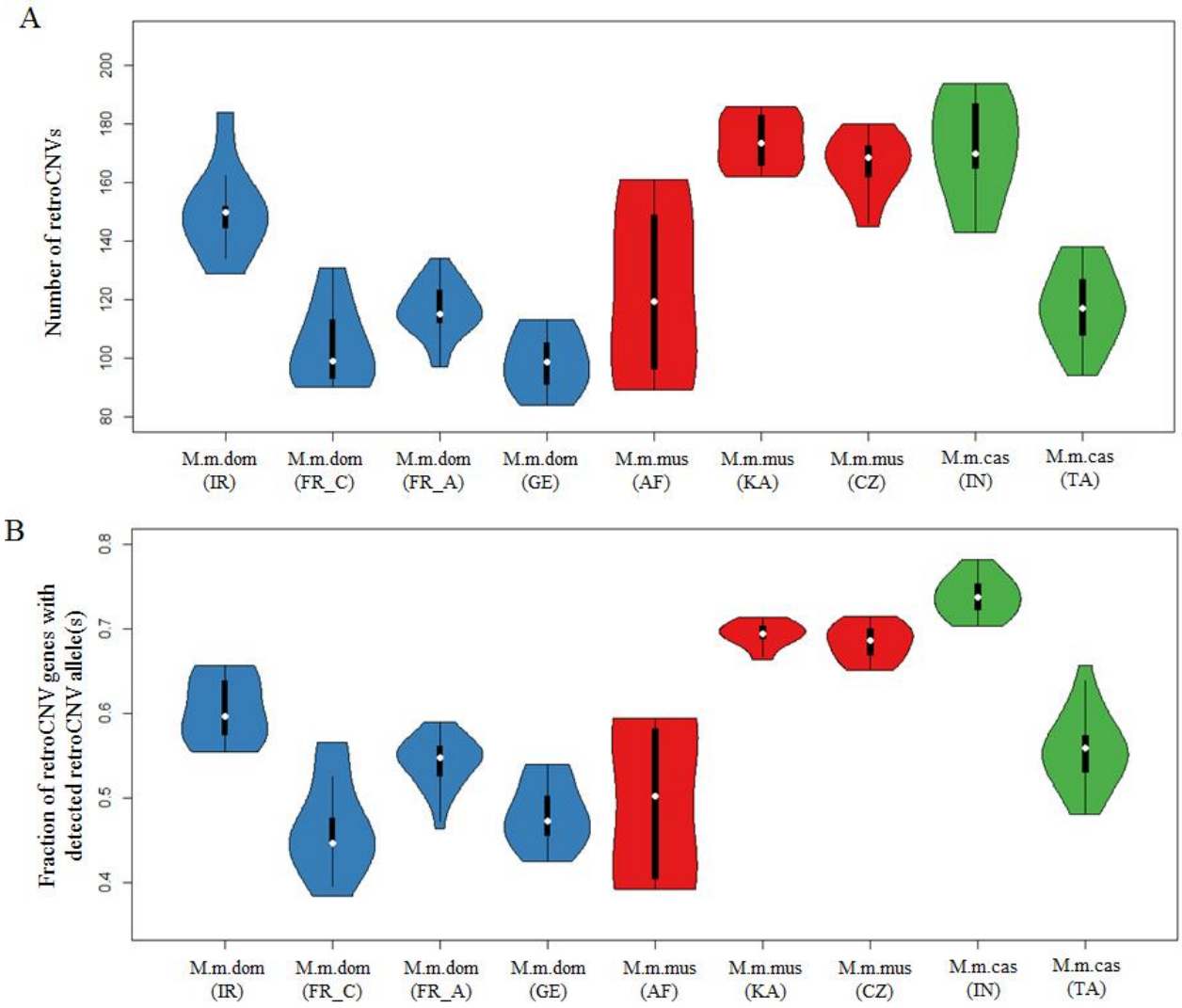

Figure S10 Distribution of gene retroposition events across house mouse natural populations. (A) Distribution of number of retroCNVs among different house mouse natural populations. In case of multiple insertion sites (or retrocopies present in the mm10 reference genome) detected for one retroCNV gene, each insertion site was taken as one independent gene retroposition event (*i.e.*, retroCNV). Note that Figure 2b in the main text shows the corresponding numbers for retroCNV genes. The overall pattern is similar, but with a relatively lower discovery rate in the Afghanistan population, which likely attributes to the shorter read length and insert size of the sequencing data for individuals from this population (Supplementary Table S1). (B) shows the fraction of retroCNV genes with detectable retroCNV allele(s) of individuals across nine house mouse natural populations. The calculation on the fraction of detectable retroCNV allele(s) was based on the data provided in Supplementary Table S3. The abbreviations for geographic regions follow Figure 1 / Supplementary Material Figure S6.

Reference

1. S. Tan, *et al.*, LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res.* **26**, 1663–1675 (2016).

2. W. J. Kent, BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

3. W. Rosikiewicz, *et al.*, RetrogeneDB-a database of plant and animal retrocopies. *Database (Oxford).*, bax038 (2017).

4. R. M. Kuhn, D. Haussler, W. James Kent, The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

5. A. Frankish, *et al.*, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

6. D. R. Schrider, K. Stevens, C. M. Cardeño, C. H. Langley, M. W. Hahn, Genome-wide analysis of retrogene polymorphisms in Drosophila melanogaster. *Genome Res.* **21**, 2087–2095 (2011).

7. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: A next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

8. Y. Zhang, S. Li, A. Abyzov, M. B. Gerstein, Landscape and variation of novel retroduplications in 26 human populations. *PLoS Comput. Biol.* **13**, e1005567 (2017).