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

Wenyu Zhang[a], Chen Xie[a], Kristian Ullrich[a], Yong E. Zhang[b,c], and Diethard Tautz[a,1]

[a]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, D-24306 Plön, Germany; [b]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, 100101 Beijing, China; and [c]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, 100101 Beijing, China

Edited by Michael Lynch, Arizona State University, Tempe, AZ, and approved December 18, 2020 (received for review June 26, 2020)

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 focused on between-species comparisons where negatively selected copies are mostly not observed, as they are quickly lost from populations. Here, we show for natural house mouse populations that the primary rate of retroposition is orders of magnitude higher than the long-term rate. Comparisons with single-nucleotide polymorphism 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 levels of their parental genes, especially when transcribed in the antisense direction. Our results imply that the impact of retroposition on the mutational load has been highly underestimated in natural populations. This has additional implications for strategies of disease allele detection in humans.

gene retroposition rate | house mouse | natural population | selection | genetic load

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). This has a variety of evolutionary implications (1–3). The intronless retrocopies have initially been viewed as evolutionary dead ends with little biological effect (4, 5), mainly due to the assumed lack of regulatory elements and promoters. However, this hypothesis has become less relevant as it has become clear that a large portion of the mammalian genome (>80%) is transcribed (6, 7) and that there is a fast evolutionary turnover of these transcribed regions. This implies 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. Therefore, it has been proposed that they contribute to the evolution of new biological functions through neofunctionalization or subfunctionalization (2, 3, 9–11).

As of yet, the possibility that retroposition events could be deleterious has not been considered as thoroughly. Deleterious effects could be due to insertions into functional sites, which have indeed been detected in a retrogene population analysis in humans (12). Even if these retrocopies land in nonfunctional intergenic regions, they could still be transcribed, and their transcripts could interfere with the function of the parental genes (13–15). In single-nucleotide polymorphism (SNP)-based association studies, this would become apparent as a transeffect on the parental gene, but the true reason for the transeffect would remain unnoticed when the retrocopy is not included in the respective genomic reference sequence. Hence, if retroposition rates are high, and if the retroposed 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., ref. 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 heterozygous and relatively low-coverage sequencing datasets. Moreover, in humans it is not possible to compare the data with very closely related lineages since they represent extinct species (e.g., Neandertals or Denisovans). As such, a comprehensive analysis is still missing 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 retroposition rates can be studied.

The house mouse (*Mus musculus*) is a particularly suitable model system for comparative genomic analyses in natural populations, as a result of its well-studied evolutionary history (21, 22). Currently, three major lineages of *M. musculus* are distinguished, classified as subspecies that diverged roughly 0.5 Mya: the Western European house mouse *Mus musculus domesticus*, the Eastern

Significance

The phenomenon of retroposition (the reintegration of reverse-transcribed RNA into the genome) has been well studied in comparisons between species and has been identified as a source of evolutionary innovation. However, less attention has been paid to possible negative effects of retroposition. To trace the evolutionary dynamics of these negative effects, our study uses a unique genomic dataset of house mouse populations. It reveals that the initial retroposition rate is very high and that most of these newly transposed retrocopies have a deleterious impact, apparently through modifying the expression of their parental genes. In humans, this effect is expected to cause disease alleles, and we propose that genetic screening should include the search for newly transposed retrocopies.

Author contributions: W.Z. and D.T. designed research; W.Z. and C.X. performed research; W.Z., C.X., K.U., and Y.E.Z. analyzed data; and W.Z., Y.E.Z., and D.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: tautz@evolbio.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2013043118/-/DCSupplemental.

Published February 1, 2021.

European house mouse *Mus musculus musculus*, and the Southeast Asian house mouse *Mus musculus castaneus*. Previously, we generated a unique genomic resource using wild mice collected from multiple geographic regions covering these three major house mouse subspecies (with each represented by natural populations) using 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 setup to generate largely homogeneous genomic/transcriptomic sequencing datasets at relatively high coverage for the same individuals (24). This made it possible to directly trace the effects of new retroposed copies on the expression of their parental genes.

Here, we show that the turnover (gain and loss) rates of retroCNVs are manyfold higher than previously estimated from comparisons between species and that the frequency spectra of retroCNV alleles in populations in comparison with SNP allele frequency spectra imply mostly deleterious effects. Transcriptome data show that the new retroCNVs are usually transcribed and have indeed an effect on the parental gene transcripts. A strand-specific RNA-Seq (RNA sequencing) dataset for one of the populations shows that antisense transcribed retroCNVs are highly underrepresented compared with sense transcripts, implying strong selection against them. We conclude that deleterious effects of newly retroposed copies 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 house mouse (*M. musculus*) individuals derived from nine natural populations, corresponding to the three major subspecies (*M. m. domesticus*, *M. m. musculus*, and *M. m. castaneus*), as well as nine individuals from two out-group species (*Mus spicilegus* and *Mus spretus*) were used to assess gene retroposition events (Fig. 1, *SI Appendix*, Table S1, and Dataset S1*A*). 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 retroposition events (Fig. 1*C*), including a power analysis for optimizing mapping conditions (*SI Appendix*, *Materials and Methods*). A retroposition event is identified on the condition that both the intron loss and the presence of a parental gene can be observed in the same individual sequencing dataset (25).

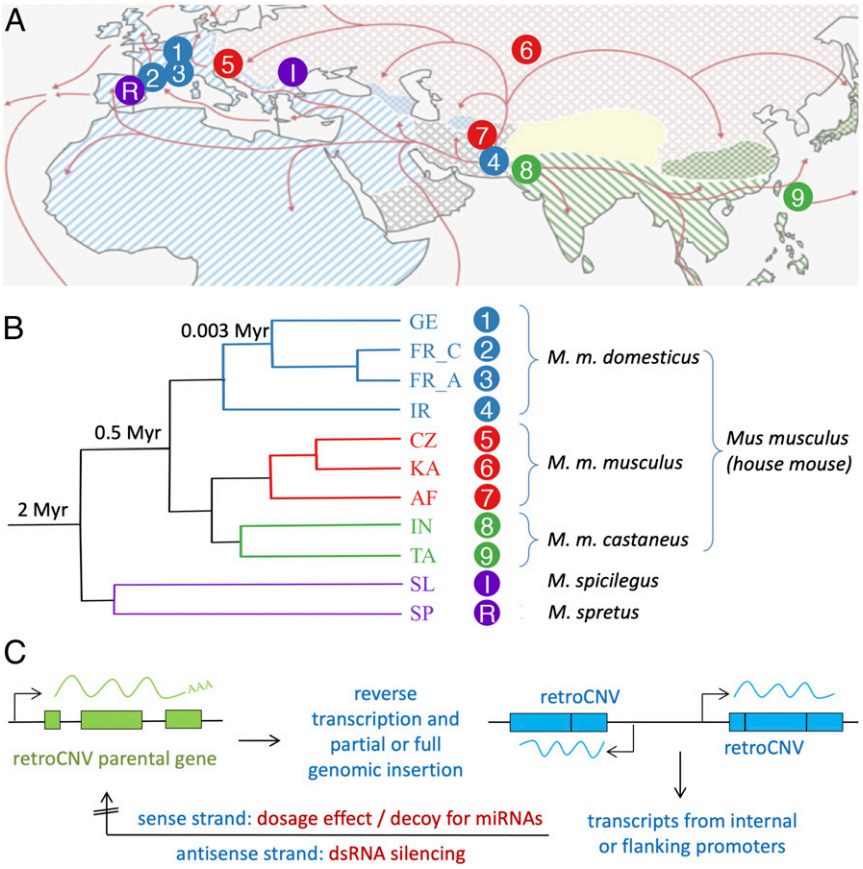

Fig. 1. Depiction of the study system. (*A*) Geographic location information on the sampled mouse individuals. Territory areas for each house mouse subspecies: *M. m. domesticus* (blue), *M. m. musculus* (red), and *M. m. castaneus* (green). Red arrows indicate possible migration routes, mostly during the spread of agriculture and trading. Geographic locations: 1, Cologne–Bonn/GE; 2, FR_C; 3, FR_A; 4, Ahvaz/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); and R, Madrid/Spain (SP). Modified from ref. 24, which is licensed under CC BY 4.0. (*B*) Phylogenetic relationships and split time estimates among the house mouse populations and two out-group species in the study (branches not shown to scale). (*C*) Depiction of the retroposition process and inferred inhibition on the parental gene. Throughout the text, genes that give rise to a processed retrocopy are called retroCNV parental genes, and the insertions of these retrocopies into the genome are called retroCNVs. Note that one retroCNV parental gene can give rise to more than one retroCNV which may be represented by different length variants. While some retroCNVs may lead to a direct disruption of other genes, most will become integrated in intergenic regions. However, given that most intergenic DNA is known to be transcribed (8), also newly integrated retroCNVs are transcribed, and these transcripts can interfere with the RNA of their parental genes. This makes most retroCNVs deleterious, such that they contribute significantly to the mutational load in the genome (the text has further details). miRNA: microRNA; dsRNA: double-stranded RNA.

Due to the need to detect at least one exon–exon junction, only protein coding genes with two or more exons (~92.4% of all coding genes annotated in Ensembl v87) were assayed as a potential source of gene retroposition. To compensate for the variance in sequencing (read length, coverage, 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 retroposition event discovery pipeline for each individual genome (*SI Appendix*, *Materials and Methods*). The resultant computational pipeline gave a low false-positive discovery rate of <3% (*SI Appendix*, Fig. S2) and a high recall rate of >95% (*SI Appendix*, Fig. S5) for all the tested individual genomes. This optimization ensures that the calling probability for retroposition events is in the same order as that for SNP calling based on GATK (GenomeAnalysisToolkit) (26) (i.e., retroCNV and SNP frequency data become comparable).

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. We directly called these alleles based on the alignment data of individual sequencing datasets to the reference genome. For those retroCNV alleles that are absent in the mm10 reference genome, we inferred their insertion sites in the genome by using discordant aligned paired-end reads when these could be uniquely mapped (*SI Appendix*, *Materials and Methods*). Additionally, a detailed discussion on the possible technical issues of retroCNV discovery can be found in *SI Appendix*, *Materials and Methods*.

High Numbers of retroCNVs in Natural Populations. Applying the above pipeline, we screened for retroCNV parental 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 the 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 parental genes and retroCNVs occurring in the *M. musculus* subspecies but absent in the out-group species).

In total, we identified 21,160 house mouse-specific retroposition events across all 96 surveyed individuals (*SI Appendix*, Fig. S6); this number also includes those detected in more than one individual, as well as 8,483 for which no insertion site could be mapped (note that we omitted these from the more detailed analysis below). These 21,160 retroposition events are derived from 1,663 unique retroCNV parental genes (Dataset S2). Only 80 (4.8%) of these retroCNV parental genes have annotated recently originated retrocopies in the mm10 reference genome based on RetrogeneDB v2 (≥95% alignment identity to their parental genes) (27), while the other 1,583 retroCNV parental genes represent newly detected gene retroposition events in house mouse wild individuals. Approximately 3.9% of these events show more than one retroCNV allele for the same retroCNV parental gene in the same individual genome (*SI Appendix*, Fig. S8).

Random resampling analysis of individuals' subsamples showed that the number of detectable retroCNV parental genes has not reached saturation with the given number of sampled individuals in our dataset (Fig. 2*A*). This implies that many more retroCNV retroposition events should be found when more individuals would be analyzed. Importantly, as suggested by ref. 28, we also found that CNV detection pipelines that do not specifically consider retroCNVs underestimate their prevalence. In a direct comparison with data from genic CNV detection (29), only <1%, on average, of the retroCNV parental genes detected in our analysis overlap with genic CNVs according to this pipeline (*SI Appendix*, Fig. S9).

On average, in each tested individual there are 212 retroCNV parental genes. However, the populations differ somewhat in these numbers (Fig. 2*B*). 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 nearly neutral retroCNVs could segregate. The majority of retroCNV parental genes (91 to 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 (Fig. 2 *C*–*E*). Conversely, the majority of retroCNV parental genes (73 to 87%) in wild-derived house mouse individuals are not present in the inbred mouse strains since these essentially represent only single haplotypes from the wild diversity.

Among the above detected retroposition events for wild house mouse individuals, between 38 and 78% of their insertion sites in the genome could be identified (*SI Appendix*, Fig. 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 were applied (30% in ref. 12). Following the "gold standard" for calling novel retrocopies [i.e., with detectable genomic insertions (20)] and unless stated separately, all of the following analyses were conducted on the basis of retroCNVs (corresponding to 12,677 retroposition events with detected insertion site), rather than retroCNV parental genes. Correspondingly, we included 2,025 unique house mouse-specific retroCNVs for further analysis (after collapsing the same retroCNV alleles detected in multiple house mouse individuals) (*SI Appendix*, Fig. S6 and Dataset S3). Note that reliable SNP calling depends also on the need for unique mapping of reads (i.e., the reduced set is directly comparable with high-quality SNP data).

Rapid Loss of retroCNVs. With SNP calling data from the same set of house mouse wild-derived individuals (*Materials and Methods*), we were able to explore the retroCNV variation at different levels, in direct comparison with the SNP variation. For both retroCNV and SNP alleles, the frequency was computed by counting individuals with positive evidence of each allele, without distinguishing the homozygous and heterozygous genotype status. If one assumes that the SNPs are mostly neutral, they can be used as expectation for the demographic drift effects in the dataset. Of the 76,882,435 house mouse-specific SNPs, 16.3% are found in all three house mouse subspecies (Fig. 3*A*), about 11% segregate in all nine populations (Fig. 3*B*), and 6.6% are found in all 96 tested house mouse individuals (Fig. 3*C*). Among the entire 2,025 different house mouse-specific retroCNV alleles with mapped insertion site (Dataset S3), only 71 (3.5%) are found in all three house mouse subspecies (Fig. 3*A*), and only about 1% segregate in all nine populations (Fig. 3*B*), while none are found in all tested house mouse individuals (Fig. 3*C*). An additional analysis using a subset of 1,551 retroCNVs (Dataset S3) showed positive evidence of retroCNV presence (i.e., detectable retroCNV allele) as well as positive evidence of retroCNV absence (i.e., reliable alignments to span the retroCNV allele). This was the case in all 96 tested house mouse individuals (*Materials and Methods*) and confirmed the same observation that retroCNVs are more skewed toward singletons than are SNPs (*SI Appendix*, Fig. S11). This suggests that retroCNVs are removed not only through drift but also through negative selection in the different lineages. This selective purging has the effect of underestimating the prevalence of retroposition rates when compared at the species or subspecies level only. In the

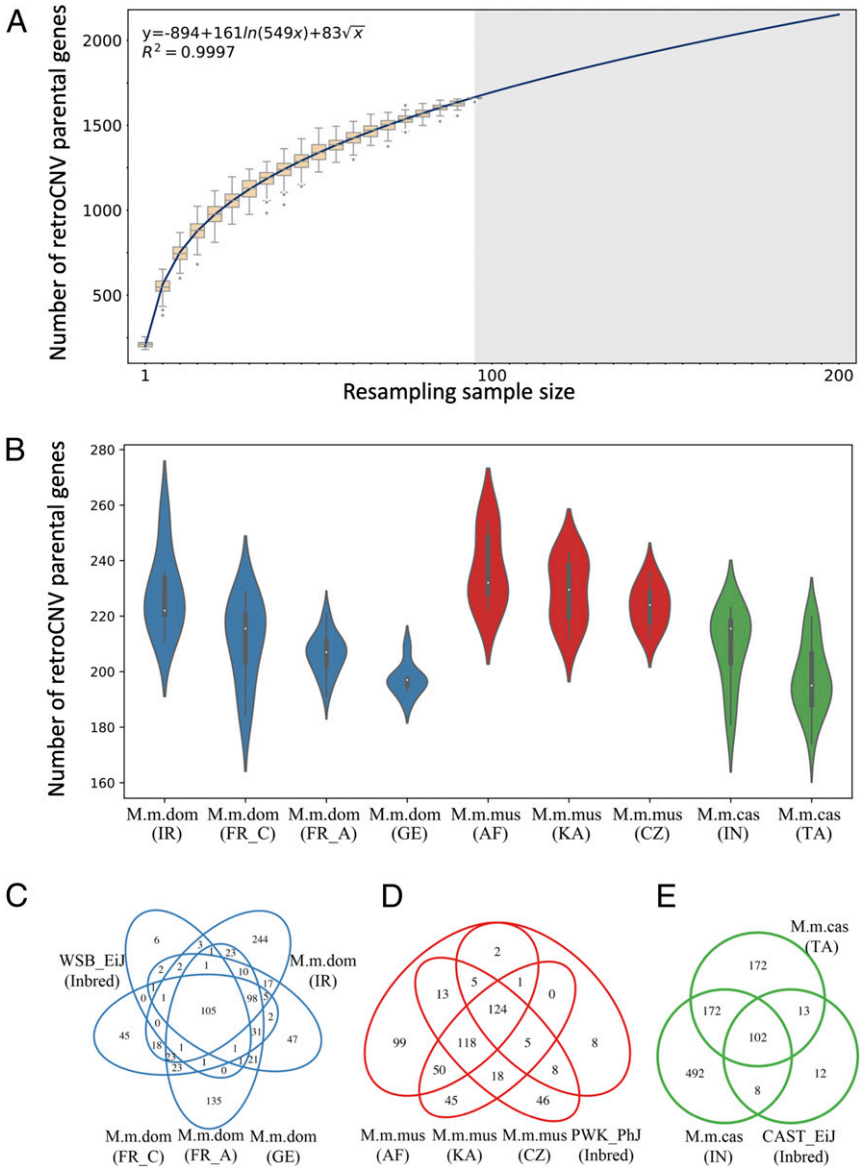

Fig. 2. Distribution of the number of detected retroCNV parental genes across house mouse populations. Only *M. musculus*-specific retroCNV parental genes are included in this analysis. (*A*) Number of detected retroCNV parental genes with increasing random resampling sample sizes. The resampling subsample sizes were selected from 1 to 95, with step size of 5. Data points represent the average number of detected retroCNV parental genes of 100 replicates for each subsample; whiskers represent the standard variance of the mean deviation. The gray area shows the prediction after doubling the number of current sampling of house mouse individuals. (*B*) Distribution of the number of detected retroCNV parental genes within each house mouse natural population (*SI Appendix*, Fig. S10 shows a corresponding depiction of retroCNVs). (*C*–*E*) Depiction of the overlap of detected retroCNV parental 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*), and CAST_EiJ (*M. m. castaneus*). (Fig. 1*A* shows a geographic representation.) AF, Afghanistan; CZ, Czech Republic; IN, India; KA, Kazakhstan; TA, Taiwan.

following, we provide an estimate for the most recent population splits in our dataset.

The Western European *M. m. domesticus* populations are derived from Iranian populations and invaded Western Europe about 3,000 y ago, where they quickly radiated. The split from the Iranian population would have occurred no more than 10,000 y ago (30, 31). This provides a time line to estimate retroCNV emergence rates by comparing the population and lineage-specific retroCNVs, under the assumption that they represent mostly new retroposition events in their lineage. For this, we used the Massif Central of France (FR_C), Germany (GE), and Iran (IR) populations 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 (Dataset S3). Assuming these populations split soon after their arrival, this would suggest on the order of 200 new retroCNV events in 10,000 y. In the IR population, we found 284 private retroCNVs (Dataset S3) (i.e., assuming a separation of 10,000 y, this would be of the same order).

A systematic comparison between primate species had suggested an birth rate of 21 to 160 retrocopies per million years (16), while our data suggest an about two orders of magnitude higher primary rate, due to looking at a recent split, as well as population samples rather than single individuals. Indeed, when we increase the population sample, we find even more population-specific retroCNVs, as is evident in the comparison between FR_C (n = 8) with Auvergne–Rhône–Alpes/France (FR_A; n = 20), where we found 60 vs. 136 population-specific retroCNVs (Dataset S3).

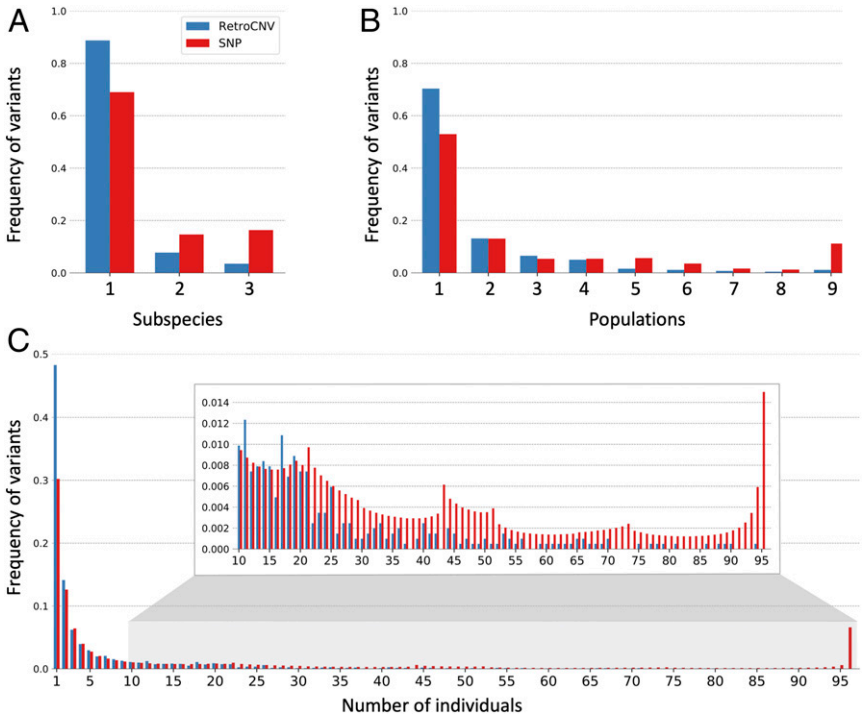

Fig. 3. Distribution of the frequency of detected retroCNVs with mapped insertion sites and SNPs (*A*) across different house mouse subspecies, (*B*) across populations, and (*C*) across individuals. *C*, *Inset* represents an enlargement with focus on the frequencies of retroCNVs/SNPs present in larger numbers of individuals.

Hence, the number of primary retroposition events could be even higher, which also explains why we do not reach saturation of retroCNV parental genes, even in our full sample set (Fig. 2*A*).

Negative selection effects can also be detected in the site frequency spectra analysis of the retroCNVs (Fig. 4) in comparison with the corresponding frequency spectra of SNP allele categories for the same population samples. Based on the functions of these SNPs, we categorized them into four distinct groups (32): 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, and 4) modifier-effect SNPs that are located in noncoding 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 $= 1.7 \times 10^{-18}$; retroCNV vs. moderate-effect SNPs: P value $= 2.6 \times 10^{-18}$; retroCNV vs. low-effect SNPs: P value $= 3.5 \times 10^{-67}$; retroCNV vs. modifier-effect SNPs: P value $= 1.3 \times 10^{-64}$]).

To test for similarity of the distributions, we used 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: $D = 0.21$). From these data, we conclude that most new retroCNVs are under negative selection (i.e., they would not only be lost by drift but also by selective purging in natural populations).

retroCNV Expression. In a previous study on the evolutionary origin of promoters of retrocopies (33), it was found that most retrocopies show low-level transcription, whereby about only 3% of them inherited the promoter from the parental gene, while the remainder recruited it from a gene in the vicinity of their insertion site (11%) or it evolved de novo from a cryptic intergenic promoter (86%). 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 GE, FR_C, and IR for which the genome sequences that we used for the retroCNV detection were also obtained (Dataset S1*B*). To combine this information, we focused on the recently originated retrocopies present in the mm10 reference genome, as annotated in RetrogeneDB version 2 (27), since full-length information for the inserted fragment is available for them. As newly originated retrocopies are usually highly similar to their parental genes (25), we implemented an effective length-based approach to calculate their specific expression (a proxy to the divergence to the parental gene), by applying a high-mismatch penalty strategy to distinguish the reads that could be perfectly and uniquely mapped to the new retrocopies (*SI Appendix*, *Materials and Methods*).

Fifty-nine retrocopies with nonzero effective lengths across the three *M. m. domesticus* populations were included for this analysis. It should be noted that these retrocopies with nonzero effective lengths will be more diverged from the parental copy than those with zero effective length, but the expression levels of the latter ones cannot be quantified since it is not possible to distinguish the reads that reliably map to the retroCNVs and those to the parental copy. We found that most of them (55 of 59) are expressed in at least one tissue or at least one population (Dataset S4; summarized in *SI Appendix*, Table S2). Most are expressed in multiple tissues, whereby the expression levels usually differ between the populations. This confirms the notion that the majority of retroCNV copies become transcribed after their insertion, 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

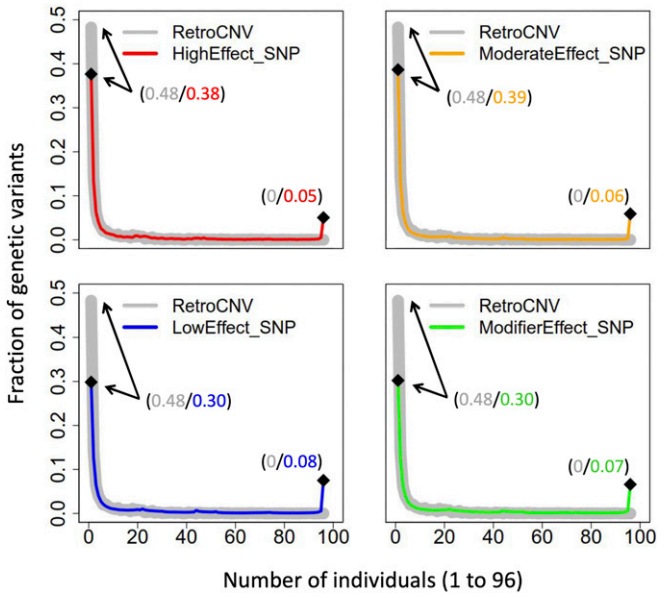

Fig. 4. Comparison of the frequency spectrum of retroCNVs with the site frequency spectra of SNPs. High-effect SNPs: the ones causing the gain/loss of start/stop codons or change of the splicing acceptor/donor sites; moderate-effect SNPs: the ones resulting in a different amino acid sequence; low-effect 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 gray) and SNPs (colors corresponding to SNP categories) that are individual private or reach fixation in all 96 tested house mouse individuals, respectively. UTR: untranslated region.

of the parental gene in the same individual. To avoid any potential bias from population structure, we only performed this line of analysis for individuals from the same populations (FR_C and GE populations separately). As these wild mice individuals were collected via a carefully designed sampling procedure, any possible effect from the genetic relatedness (or population substructure) among individuals should also be minimized (23, 24). We also restricted this analysis to the animals with singleton retroCNVs in each population (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, whereby all combinations of test vs. reference individuals can occur. We used a Wilcoxon rank sum test to ask whether the presence of a retroCNV led to a significant expression change in the respective individual.

To begin, we focused this analysis on the loss of expression, for which most likely antisense transcription of retroCNVs would silence the parental gene's expression level (13). We found that 22% (GE) and 31% (FR_C) of the singleton retroCNVs have in at least one tissue a significant negative effect (False discovery rate, i.e., FDR $\leq$ 0.05) on the expression of their parental gene (Table 1 and Dataset S5 *A* and *B*).

Around three-quarters of these retroCNVs (GE: 55/74; FR_C: 57/71) show truncated exons compared with their parental gene (Dataset S5 *C* and *D*), and this allowed us to explore up-regulation effects on parental gene expression since the expression level can be explicitly quantified based on the read fragments mapped to the exons that are unique to the parental gene. We found that about 7% of the singleton retroCNVs in both GE and FR_C populations have, in at least one tissue, a significant up-regulation effect (FDR $\leq$ 0.05) on the expression of their parental gene (Table 1 and Dataset S5 *C* and *D*). This hints that retroCNVs could also functionally interfere with their parental gene expression through sponging regulatory microRNAs (15).

Strand-Specific Expression of retroCNVs. To further assess whether the deleterious effects of 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 (50 retroCNVs occurred in the 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 toward sense transcripts (Table 2 and Dataset S6). This applies not only to the number of transcribed retroCNVs per tissue but also to the level of transcription (Dataset S6). Since only a low fraction (~3%) of retrocopies in mammals is expected to have inherited the promoter from the parental gene (33), it is unlikely that the direction of integration into the chromosomes could be biased to this extent. Hence, we interpret this finding as a strong selection against retroCNV copies that showed antisense transcription, implying that they are affecting their parental genes via dsRNA (double-stranded RNA) silencing (13).

Discussion

Our population-based retroCNV analysis allowed a much deeper insight into the retrogene formation dynamics than what has previously been possible. Most importantly, we found that the primary origination rate of retroCNVs must be orders of magnitudes higher than the one that is derived from between-species comparisons. 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., they might be considered "disease" alleles). Furthermore, 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 functional 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 (34). We have previously shown that this effect can also be traced in mitochondrial mutation patterns of mice after island colonization (35), and we observed it here for the comparisons of gene retroposition events between the most recently diverged populations.

Our rate estimates assume a more or less constant retroposition activity, rather than episodes of retropositions. As the main source of reverse transcriptase for retroposition, LINE-1 (long interspersed nuclear element-1) elements are generally continuously active in mammals (2). In line with this, we observed similar numbers of retroCNV parental genes across all nine tested house mouse populations (Fig. 2*B*) (median values range from 200 to 230), indicating a more or less constant rate of retroposition turnover activity. An episodic retroposition activity has been proposed to explain the relatively high birth rate of

Table 1. Singleton retroCNVs with significant effects on their parental genes' expression in their population

Regulation pattern and population	Total no. of singleton retroCNVs	No. of singleton retroCNVs with significant (FDR ≤ 0.05) effect on parental gene expression	Average no. of tissues affected per singleton retroCNV
Down			
GE	74	16	1.81 ± 0.98 SD
FR_C	71	22	2.14 ± 0.99 SD
Up			
GE	55	4	2
FR_C	57	4	2 ± 0.82 SD

retrocopies in ancestral primates (36, 37), but in the light of our results, an alternative interpretation would be an enhanced retention rate of these retrocopies, possibly because some of them may have become involved in primate-specific adaptations.

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 nonrecessive effects. Most intriguingly, our data directly show the impact of new retroCNVs on the transcription of their parental genes; 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 and 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 the 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 samples since they would be most quickly purged.

But 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 microRNAs (15). In our analysis of retroCNV effects in GE and FR_C populations (Table 1), we also found around 7% having a possibly negative effect caused by the up-regulation of the expression of their parental genes. This is further supported by the observation that retroCNVs transcribed on the sense strand and antisense strands share the similar pattern of allele frequencies (*SI Appendix*, Fig. S12). Hence, while many previous reviews on retrogenes have focused of the evolutionary potential generated by retrogenes, these apparently strongly deleterious effects have been overlooked.

Implications for Human Genetic Disease Studies. Our analysis 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 (38). We found around 200 retroposition events per mouse genome, of which a substantial fraction is likely to have direct deleterious effects. This includes most of the transcribed antisense copies, but also a fraction of the sense copies, given that we observed the strong purging of retroCNVs in comparison with SNPs. Most importantly, if the negative effects of retroCNVs are related to their transcription, only one allele would suffice to cause the effect (i.e., the negative effects are not recessive). Accordingly, the retroCNV mutational load can be expected to be at least as large as that caused by (mostly recessive) SNPs.

A comparable retroCNV study in human populations (12) revealed also a very high rate of new retroCNVs, although about three times less (1,663 retroCNV parental genes in house mouse populations vs. 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 nonrecessive effect, this means that retrocopies may be equally likely to cause a genetic disease as new SNP mutations. Genome-wide association mapping studies of complex genetic diseases often find SNP associations in intergenic regions that are interpreted as regulatory variants. It is possible that some of these SNPs are in close linkage to an undetected retroCNV exerting a transregulatory 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 does not yet include specific pipelines for retroCNV analysis (39). Although there are a few known cases where retroCNVs have caused a genetic disease through direct inactivation of genes (3, 40), a much more systematic approach to trace events caused by the transcriptional activity of retroCNVs seems warranted.

Materials and Methods

Genome Datasets. We obtained the mouse reference genome sequence (mm10/GRCm38) and gene annotation data from Ensembl version 87 (41). We also retrieved the genome assembly sequence data for two out-group sister species (SPRET_EiJ_v1: *M. spretus*; GCA_003336285.1: *M. spicilegus*) from the National Center for Biotechnology Information GenBank database (42, 43).

Table 2. retroCNV expression patterns in the strand-specific RNA-Seq dataset

	Testis	Brain	Kidney	Liver	Heart
No. of expressed retroCNVs (FPKM > 0, sense strand)	35	25	31	20	23
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)
No. of expressed retroCNVs (FPKM > 0, antisense strand)	16	10	8	6	7
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)

*Only the retroCNVs with nonzero expression were included. FPKM: Fragments per kilobase of transcript per million mapped reads.

Details for the whole-genome sequencing data from wild individuals (24) are listed in Dataset S1*A*.

Identification of House Mouse-Specific retroCNV Parental Genes. Based on previous approaches (18, 19, 25), we developed a refined computational pipeline for the discovery of retroCNV parental genes based on the short-read sequencing datasets from individual genomes (*SI Appendix, Materials and Methods*). 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. A more detailed description of the discovery of retroCNV parental genes can be found in *SI Appendix, Materials and Methods*.

Detection of retroCNV Alleles. Based on the detected house mouse-specific retroCNV parental genes, we performed detection of retroCNV alleles at individual genome level. The presence status of retrocopies that are annotated in the mm10 reference genome and the insertion sites for those retrocopies absent in the reference genome were analyzed separately (*SI Appendix, Materials and Methods*).

Comparison of the Allele Frequency Pattern between retroCNVs and SNPs. We followed the general GATK version 3 Best Practices (44) to call SNP variants (*SI Appendix, Materials and Methods*) and only kept the SNP variants with unambiguous ancestral states in out-group species. We predicted the functional effects of each SNP by using Ensembl VEP v98.2 (32), based on the gene annotation data from Ensembl version 87 (41). Further details are in *SI Appendix, Materials and Methods*.

Transcriptional Profiling of retroCNVs. We used two different sets of transcriptomic sequencing data for the transcriptional profiling of retroCNVs: 1) one nonstrand-specific RNA-Seq dataset from our previously published data (24) and 2) one strand-specific RNA-Seq dataset newly generated in the present study. The detailed description of these two datasets can be found in Dataset S1 *B* and *C*. The details on quantifying expression levels, as well as the assessment of the impact on parental gene expression from singleton retroCNVs, are provided in *SI Appendix, Materials and Methods*.

Data Availability. The raw strand-specific RNA-Seq data generated in this study are available in the European Nucleotide Archive under study accession number PRJEB36991.

ACKNOWLEDGMENTS. We appreciate Peter Keightley and Guy Reeves for reading through the manuscript and providing helpful comments. We thank Julien Dutheil for valuable suggestions on statistical analysis and 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 high-performance computing cluster of the Max Planck Institute for Evolutionary Biology. This work was supported by institutional funding through the Max Planck Society (to D.T.).

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. 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*.; 1000 Genomes Project Consortium, 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*.; Broad Institute Genome Sequencing and Analysis Program and Platform, 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. R. Nielsen, J. S. Paul, A. Albrechtsen, Y. S. Song, Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).
27. W. Rosikiewicz *et al*., RetrogeneDB-a database of plant and animal retrocopies. *Database (Oxford)* **2017**, bax038 (2017).
28. 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).
29. 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).
30. 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**, 26 (2015).
31. T. Cucchi *et al*., Tracking the Near Eastern origins and European dispersal of the western house mouse. *Sci. Rep.* **10**, 8276 (2020).
32. W. McLaren *et al*., The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
33. F. N. Carelli *et al*., The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**, 301–314 (2016).
34. S. Y. W. Ho *et al*., Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**, 3087–3101 (2011).
35. 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).
36. K. Ohshima *et al*., Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.* **4**, R74 (2003).
37. A. C. Marques, I. Dupanloup, N. Vinckenbosch, A. Reymond, H. Kaessmann, Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* **3**, e357 (2005).
38. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
39. S. S. Ho, A. E. Urban, R. E. Mills, Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
40. J. Ciomborowska, W. Rosikiewicz, D. Szklarczyk, W. Makałowski, I. Makałowska, "Orphan" retrogenes in the human genome. *Mol. Biol. Evol.* **30**, 384–396 (2013).
41. F. Cunningham *et al*., Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).
42. D. A. Benson *et al*., GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
43. J. Lilue *et al*., Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**, 1574–1583 (2018).
44. G. A. Van der Auwera, *et al*., From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).