Population genomic scan of endogenous retrovirus insertions revealed cryptic drivers of selective sweep in wild house mouse genomes

Tomoki Yano[1], Toyoyuki Takada[2], Kazumichi Fujiwara[3], Dai Watabe[4], Shutaro Hirose[5], Hiroshi Masuya[2], Toshinori Endo[1, 4, 5], Naoki Osada[1, 4, 5*]

1. Graduate School of Information Science and Technology, Hokkaido University
2. Integrated Bioresource Information Division, RIKEN, BioResource Research Center
3. Mouse Genomics Resource Laboratory, National Institute of Genetics
4. Faculty of Information Science and Technology, Hokkaido University
5. School of Engineering, Hokkaido University

Naoki Osada: nosada@ist.hokudai.ac.jp

Abstract

Detecting selective sweeps in population genomic scans is often complicated by the presence of "missing drivers of hitchhiking", lacking known functional genes or mutations in the target regions. Non-reference genomic insertions, such as endogenous retroviruses (ERVs), are common in mammalian genomes and are potential drivers of genetic hitchhiking. In this study, we performed a population genomic scan of non-reference ERV insertions in wild house mice (*Mus musculus*). Using a newly developed bioinformatics pipeline, we identified and characterized >100,000 non-reference ERV insertions from short-read sequencing data across 163 wild house mouse genomes. Our analysis revealed substantial variation in ERV insertion patterns among wild house mice.

Focusing on the *Fv4*, a known ERV-derived defense gene against murine leukemia viruses (MLVs) and mainly restricted in the range of *M. m. castaneus*, we detected evidence of adaptive introgression from *M. m. castaneus* to *M. m. musculus*, particularly in a part of northern East Asia. Selective sweep analysis in Korean *M. m. musculus*, where *Fv4* is prevalent despite a predominantly *musculus*-like genetic background, indicated that the spread of *Fv4* has been driven by recent positive selection presumably owing to the protection against MLVs. Additional genomic scans identified several other ERV insertions associated with high-frequency selective

sweeps, suggesting that ERVs serve as cryptic drivers of adaptive evolution.

Our findings highlight the role of ERV as genetic elements subject to selection, serving as potential contributors to host-pathogen interactions. This study provides novel insights into the evolutionary dynamics of ERV insertions and their impact on genome evolution.

Introduction

Identifying the molecular mechanisms underlying adaptive evolution is a fundamental challenge in evolutionary biology. One widely used approach in the population genomics era is to detect signatures of selective sweeps, characterized by a reduction in allelic and haplotypic diversity surrounding a target adaptive mutation (Smith and Haigh 1974; Sabeti, et al. 2002; Nielsen, et al. 2005), using genome-wide single nucleotide polymorphism (SNP) data sampled from populations. Such studies have identified numerous candidate loci undergoing recent adaptive evolution, including the lactase gene (*LCT*) in humans (Bersaglieri, et al. 2004) and domestication-related genes in rice (Huang, et al. 2012). These findings offer profound perspective on how organisms have adapted to environments through adaptive evolution. Increasing evidence indicates that immune-related loci are among the most significant targets of adaptive evolution, driven by the ongoing host-pathogen arms race (Sabeti, et al. 2002; Barreiro and Quintana-Murci 2010; Fagny, et al. 2014).

One challenge in the genomic scan studies is that many candidate loci showing signatures of selective sweeps often do not contain apparent functional genes or mutations. The pattern may complicate the interpretation of population genomic scans for selective sweeps, given the relatively high false positive rate associated with the approach (Mallick, et al. 2009). In addition,

regions under selective sweeps are inherently large, frequently on the order of tens to hundreds of kilobases. Although unannotated functional elements within the sweep region may contribute to adaptive evolution, genomic elements absent from the reference genome, such as transposable element insertions/deletions, could also be responsible for the selective sweep (Weedall, et al. 2020). A recent study by Rogers et al. (2023) found a relative deficiency of transposable element insertions in regions under selective sweep in *Megalonaias nervosa* (Rogers, et al. 2023). However, different classes of transposable elements may drive adaptive evolution in various organismal systems.

In this study, we focus on non-reference polymorphic insertions of endogenous retroviruses (ERVs) in the genomes of wild house mice (*Mus musculus*). ERVs are transposable elements derived from RNA viruses that encode reverse transcriptase and can become endogenized in host genomes when integrated into germline DNA. ERVs are classified into three major classes (Class I to III) (Stoye and Coffin 1987; Stocking and Kozak 2008) and collectively constitute approximately 10% of the house mouse genome (Mouse Genome Sequencing, et al. 2002). In laboratory mice, a wide variation in ERV insertion patterns has been documented, with some ERV and ERV-like elements significantly influencing the level of protection against retroviral infections (Ikeda, et al. 1981; Best, et al. 1996; Dodding Mark, et al. 2005; Wu, et al. 2005). Although ERVs

can influence organismal phenotypes through various mechanisms, including gene silencing, insertional disruption, and nearby gene activation via long-terminal repeats (LTRs), previous studies in wild house mice have suggested that the distribution of these restriction factors may reflect adaptation to pathogen infections through various mechanisms (Boso, et al. 2021; Fujiwara, et al. 2024). The advantageous effect of ERV insertions for the resistance of infection has been also debated in humans, but there has been no direct evidence. To address these questions, a comprehensive survey of ERV insertions using population genomic data is essential.

House mice are genetically classified into three major subspecies: *M. m. castaneus*, *M. m. domesticus*, and *M. m. musculus* (Boursot, et al. 1996; Fujiwara, et al. 2022), and the classical laboratory strains of house mice predominantly possess a *domesticus*-like genetic background (Keane, et al. 2011). Since the reference genome of house mice is derived from the laboratory strain C57BL/6J (Mouse Genome Sequencing, et al. 2002), it lacks a substantial portion of the ERV insertions present in wild populations. Given the extensive genetic diversity in wild house mice and many presently active ERV families, analyzing the non-reference ERV insertions in their genomes offers valuable insights into the population genetic dynamics of ERV variation. In particular, house mice are highly susceptible to various active Murine Leukemia Viruses (MLVs), which belong to class I retroviruses, including Friend and Moloney leukemia viruses. We

anticipate that the interplay between active retroviruses and ERVs may leave signatures of selective sweeps in wild house mouse genomes.

One of the drawbacks in conducting a genomic scan for ERV insertions in non-human organisms is the lack of appropriate data and methodologies. Although long-read sequencing technologies such as Oxford Nanopore and PacBio can unambiguously detect long polymorphic indels in complex genomic regions (Logsdon, et al. 2020; Chu, et al. 2021), the high sequencing cost remains a limitation for large-scale population studies. Sequencing more than 100 individuals with these technologies is still impractical, whereas short-read whole-genome sequencing datasets are more widely available across many organisms. Several algorithms have been developed to identify ERV insertions using short-read genome data. However, most of these tools are optimized for human datasets with limited sample sizes and tend to have a relatively high false-positive rate (Keane, et al. 2012; Gardner, et al. 2017; Chen and Li 2019; Bowles, et al. 2022; Kojima, et al. 2023). In particular, searching for a wide range of transposable elements using typical ERV detection algorithms often requires substantial computational time, making it practically challenging to apply to a large number of genomes. To address the issue, we developed ERVscanner, a new ERV detection pipeline optimized for high computational speed and a low false discovery rate by streamlining key bioinformatics processes. To evaluate the speed and

accuracy of ERVscanner, we compared its results for the wild-derived laboratory strain JF1 mouse (Koide, et al. 1998; Takada, et al. 2013) using both short- and long-read sequencing data.

In order to benchmark our detection method and scan for adaptive ERV insertions, we focused on *Fv4*, a known restriction factor for MLV but missing from the house mouse reference genome. *Fv4* encodes an envelope (env) protein derived from a class I retrovirus (Odaka, et al. 1978; Inaguma, et al. 1991) and exhibits polymorphism for presence and absence and is predominantly found in the subspecies *castaneus*. In East Asia, a broad hybrid zone formed by secondary contact between the subspecies *castaneus* and *musculus* has been observed (Fujiwara, et al. 2022). Secondary contact can serve as an efficient mechanism for introducing potentially adaptive genetic variation into different populations or species (Huerta-Sanchez, et al. 2014). Although our previous studies showed that the genomic background of Korean samples is highly *musculus*-like, many Korean individuals have been shown to harbor *Fv4* (Inaguma, et al. 1991; Boso, et al. 2021). If the widespread presence of *Fv4* in the Korean population is due to adaptive introgression following secondary contact between different subspecies in East Asia, similar to the case of the *Fv1* gene introgression in the Japanese population (Boso, et al. 2021; Fujiwara, et al. 2024), it would leave a signature of selective sweep in the Korean population. In this study, we examined whether polymorphic insertions of *Fv4* and the signature of selective sweep surrounding the locus could be

successfully detected by our newly developed pipeline. Our additional genealogical analysis indicated that *Fv4* was recently introgressed from the subspecies *castaneus* to *musculus*, likely as a result of interactions between the two subspecies. Three additional cases of putative adaptive introgression of ERV sequences were identified in the Korean population, highlighting the potential role of ERV insertions as a cryptic source of adaptive evolution.

Results

Evaluation of ERV insertion detection pipeline

A new bioinformatics pipeline, ERVscanner, was constructed to analyze the non-reference insertions of ERVs. The basic idea of ERVscanner is almost the same as the previously published programs for detecting mobile elements, but ERVscanner largely relies on the repetitive sequence annotation for the reference genome by Dfam database (Storer, et al. 2021). Assuming that nearly all non-reference ERV insertions have homologous sequences annotated as ERV regions in the reference genome, we can significantly reduce the computational search space. Additionally, ERVscanner avoids assembling insertion sequences, a process that is typically time-consuming, because such assemblies are often error-prone due to the nature of short-read sequences and high complexity of insertion sequences. Instead of inferring the exact sequence of insertions, ERVscanner estimates the class, family, and orientation of candidate insertions for each individual,

outputting the results in VCF format. We can also incorporate other seed sequences for potential insertions. In this study, we used sequences of *Fv4*, *RMCF1*, and *RMCF2*, which are known antiviral factors derived from retroviruses, as additional targets for non-reference insertions. Detailed methods are described in the Methods section.

We evaluated the computational speed, sensitivity, and accuracy of ERVscanner using high-coverage long-read and short-read sequences from the JF1 mouse strain. Assuming that long-read data accurately represents true insertion sequences, we identified 8,365 non-reference insertions longer than 200 bp with homology to ERV sequences. ERVscanner is designed to enhance accuracy by leveraging data from multiple individuals. However, for benchmarking purposes, we evaluated its performance using only a single individual of JF1 strain to ensure a fair comparison. For comparison, we selected ERVcaller, a tool noted for its high sensitivity and low false-positive rate in a benchmarking study (Bowles, et al. 2022). Both pipelines analyzed all classes of ERVs using reference-mapped short-read files. Utilizing eight cores of an Intel Xeon CPU E5-2620 v4 at 2.10 GHz, ERVscanner detected 3,202 insertions in approximately 3 hours. For ERVcaller, we estimated all non-redundant transposable elements rather than only ERV sequences and extracted inferred insertions matching ERVs to enhance accuracy. Using stringent criteria (support levels 4 and 5), ERVcaller identified 11,850 candidate ERV insertions in 82 hours

with the same computer used for ERVscanner.

Figure 1 presents a Venn diagram comparing the results of JF1 strain using three datasets. Here we focus on ERV insertions longer than 200 bp. As expected, many insertions identified by long-read sequencing are undetectable by short-read methods, highlighting the inherent limitations of short reads in detecting ERV insertions. ERVcaller identified more correct ERV insertions than ERVscanner; however, its false discovery rate (FDR) approached 60%. In contrast, ERVscanner exhibited an FDR of approximately 38%. While this rate remains high, it is notably lower than that of ERVcaller. The majority of mispredictions in ERVscanner are due to misclassification of the sequence class rather than incorrect insertion detection. When considering only the presence or absence of insertions longer than 50 bp, FDR of ERVscanner was only 5%, showing very high specificity for insertion detection (Supplementary Figure 1).

Population genomic search of ERV across wild house mouse genomes

We analyzed short-read mapping data from 163 wild house mice using ERVscanner (Supplementary Table 1), and identified an average of 7594, 4065, and 7706 non-reference ERV insertions on autosomes in the subspecies *castaneus*, *domesticus*, and *musculus*, respectively. The distribution of these insertions for each individual is presented in Figure 2. As expected, the subspecies *domesticus* harbored the fewest ERV insertions, because the reference genome is

primarily derived from the subspecies *domesticus*. The subspecies *castaneus* and *musculus* exhibited distinct patterns of ERV insertions. The subspecies *castaneus* had significantly more Class III ERV non-reference insertions than *musculus* ($P = 4.9 \times 10^{-10}$, Mann–Whitney U test), while possessing fewer Class I ERV non-reference insertions ($P = 1.4 \times 10^{-6}$; Mann–Whitney U test).

We tested whether global population structures could be reconstructed using the pattern of non-reference ERV insertions. Principal Component Analysis (PCA) was performed using ERV insertions as markers, and the results reflected the subspecies population structures of wild house mice (Figure 3), consistent with previous findings using SNP markers. Further clustering with PCA-UMAP demonstrated that most samples from the same countries clustered tightly (Supplementary Figure 2), indicating that the features of ERV insertions are region-specific.

Functional effect of ERV insertions

In total, we identified 175,957 insertion loci that were not present in the reference genome. As previously observed, there was a strand bias of ERV insertions in intronic regions. We observed 61,689 insertions overlapped with the transcribed regions, and among them, 40,485 ERVs were inserted in a reverse direction to the transcription. We further investigated the relationship between integration sites and local recombination rates. We performed randomization test for each

chromosome, but except for a few chromosomes showing very weak positive correlation, we did not observe robust signal of correlations between them (Supplementary Table 2). In contrast, we found a statistically significant negative correlation between insertion allele frequency and recombination rates, suggesting that common insertions preferentially reside in low-recombination regions (Spearman correlation coefficient: −0.0238; *P* value: 1.90×10^{-23}; Supplementary Figure 3).

We identified 299 genes with ERV insertions disrupting protein-coding regions in *M. musculus* samples (Supplementary Table 3). Functional enrichment analysis using Metascape software showed the most significantly enriched category was detection of chemical stimuli, including 58 olfactory receptor (OR) genes and 14 vomeronasal receptor (*Vmnr*) genes (Supplementary Figure 4). Other overrepresented categories involved immune system-related genes, such as genes related to macrophage migration, monocyte chemotaxis, and leukotriene metabolism.

Polymorphic insertion of restriction factor Fv4

To examine the causal relationship between ERV insertions and natural selection, we focused on the *Fv4* gene, a major restriction factor for MLV infection that is absent from the reference genome. Long-read sequencing data revealed that the JF1 mouse strain possesses *Fv4* at chr12:80,797,701 in the autosomal regions of GRCm38 reference genome. ERVscanner

successfully inferred that *Fv4* had presence/absence polymorphism in wild house mouse samples at the same position. The geographic distribution of *Fv4* (Supplementary Figure 5) was consistent with the findings of Boso et al. (2021) but offers higher geographic resolution. In general, East Asian house mice exhibit a *musculus*-like genetic background in northern China and a *castaneus*-like background in southern China, with the Yangtze River serving as an approximate boundary. *Fv4* is primarily restricted to the subspecies *castaneus*, but some East Asian *musculus*-like individuals along the eastern coastal regions, including Northern China (Shandong and Hebei Provinces), Korea, Japan, and Russian Primorye, also harbor *Fv4*. Among these, Japanese and Russian samples show mixed subspecies origins. In contrast, seven individuals from northeastern China, which have a strongly *musculus*-like genetic background, do not harbor *Fv4*.

The distribution pattern of *Fv4* suggests that this gene may have introgressed from the subspecies *castaneus* to *musculus* during secondary contact in East Asia. To investigate whether the spread of *Fv4* has been driven by adaptive evolution, we conducted a genomic scan for positive selection in Korean samples. These samples predominantly exhibit a *musculus*-like genetic background, yet nearly all individuals possess *Fv4*. We plotted nucleotide diversity and Tajima's *D* statistics surrounding the *Fv4* insertion sites (Figure 4) using SNP data. Additionally, using northeastern Chinese samples, which do not possess *Fv4*, as a reference, we calculated and

plotted F_{ST} and cross-population normalized selection scores (XP-nSL) (Figure 4). The results indicate a reduction in polymorphisms, an excess of rare alleles, high differentiation between populations, and homogeneous haplotype structures, all signatures consistent with a selective sweep. These findings support the hypothesis that *Fv4* has undergone adaptive introgression from the subspecies *castaneus* to the subspecies *musculus* in East Asia, likely facilitated by secondary contact between these subspecies. The presence of *Fv4* in Korean *musculus*-like populations, despite their genetic background, suggests positive selection has favored the retention of this gene, potentially due to its role in conferring resistance to MLVs.

We also inferred the timing and strength of selection using genealogy-based methods implemented in the RELATE and CLUES software (Speidel, et al. 2019; Stern, et al. 2019). Combining all individuals of the subspecies *musculus*, we estimated that the increase of *Fv4* insertion frequency occurred after approximately 1000 generations ago, with a selection coefficient of 0.47%. Log-likelihood ratio against the neutral model was 8.50. The trajectory of *Fv4* insertions over time in the subspecies *musculus* is shown in Supplementary Figure 6.

Potential candidates of drivers for adaptive ERV insertions

We conducted a genomic scan for selective sweeps in the Korean population using nSL statistics. From this analysis, we selected the top 0.1% of SNPs with the highest nSL scores. Recognizing

that SNPs with elevated nSL scores often cluster together, we merged SNPs within 10 kbp genomic regions, resulting in 607 high-nSL SNPs. Among these SNPs, 65 SNPs lacked adjacent annotated exons within 10 kbp but were flanked by ERV insertions within 1 kbp. Among the candidate insertions, three insertions present at high frequencies in the Korean population but low in the subspecies *castaneus*. In all three cases, the JF1 mouse strain possesses the corresponding insertions, and we inferred the sequence contents using available long-read sequencing data.

Table 1 presents a list of insertions that have potentially reached high frequency in the Korean population by adaptive evolution. Among these, one locus (chr1:3723618) is a solo LTR derived from an IAP element; another (chr7:42156902) is a nearly full-length RLTR10 insertion lacking fully encoded proteins; and a third (chr18:21435569) encodes a 1,153-amino-acid protein in the longest open reading frame, corresponding to ERV gag and pol proteins.

Discussions

Evaluation of methods

In this study, we conducted a comprehensive search for non-reference ERV insertions in wild house mouse genomes and identified 175,957 potential insertions on autosomes across individuals. To facilitate this, we developed a scalable pipeline designed to handle large datasets efficiently.

Our pipeline is optimized for population genomic data, leveraging shared insertions among individuals to enhance accuracy. Some insertions might result from somatic integration of exogenous retroviruses such as MLVs. However, 60,619 loci were shared among multiple individuals, suggesting they originated in germline cell lineages.

Genomic pattern of ERV insertions

Multiple genomic factors may contribute to the distribution patterns of ERVs across genomes. ERVs and other transposable elements tend to integrate into regions of open chromatin, while recombination may facilitate their removal (Campos-Sánchez, et al. 2016) (Kent, et al. 2017). Additionally, some ERVs contain sites that promote recombination (Edelmann, et al. 1989), whereas ERV integration into functional genomic regions can negatively impact fitness and may be selectively removed from the genome (Campos-Sánchez, et al. 2016). These opposing forces make it difficult to predict whether ERVs preferentially integrate into regions with high or low recombination rates. Our results, which focus on polymorphic ERVs, did not show a robust correlation between ERV insertions and local recombination rates. However, we observed a significant negative correlation between allele frequency of insertions and local recombination rate, indicating that younger insertions tend to occur in regions of low recombination. This observation can be explained by either negative natural selection against insertions or the high

removal activity of insertions through recombination.

Genes affected by ERV insersions

We also found that polymorphic ERV insertions disrupting many classes of genes. While identifying a general trend in the list is challenging, one intriguing gene family overrepresented in the list is the OR genes. OR genes are among the most frequently observed in mammalian genomes and include many pseudogenes. Because ERV insertions in OR genes are not confined to specific integration sites, their enrichment is more likely due to weak purifying selection against gene disruption than to preferential targeting of these loci.

Beyond OR genes, we also observed enrichment of genes in the pentose and glucuronate interconversion pathway (Akr1b10, Dhdh, and Ugt1a5), suggesting that ERV insertions can disrupt these loci and thereby contribute to genetic diversity in response to environmental variation in wild house mice.

In addition to direct disruption, ERV insertions may influence the expression of adjacent genes, for example by providing regulatory elements, further contributing to phenotypic diversity in these populations.

ERV insertions and adaptation

While ERV insertions can influence organismal phenotypes in various ways, one intriguing hypothesis is that ERVs may act as restriction factors against similar exogenous retroviruses (Goff 2004). However, molecular evidence for this mechanism remains scarce. In this study, we searched for retroviral-derived restriction factors not presented in the reference genome, *Fv4*, *RMCF1*, and *RMCF2*, but only *Fv4* was detected by our pipeline. *Fv4*, a restriction factor that enhances resistance to MLVs but is absent from the mouse reference genome, serves as an ideal model for studying the evolutionary dynamics of ERV insertions in mammalian immunity. Derived from MLVs, *Fv4* retains only the *env* gene in complete form, which encodes an envelope protein with several substitutions and blocks MLV receptors, thereby inhibiting infection. The gene likely originated in the common ancestor of the subspecies *castaneus*, but previous studies have identified *Fv4* in some individuals of the subspecies *musculus*.

Our global population-wide survey revealed that individuals from Northern China, Korea, Japan, and Russian Primorye carry *Fv4*. Notably, the Northern Chinese and Korean samples predominantly exhibit *M. m. musculus* genetic backgrounds, suggesting that *Fv4* spread in these populations despite restricted gene flow, likely due to its strong selective advantage. Population genetic analyses, using various summary statistics, detected strong selective sweeps around *Fv4*

insertion sites, supporting a scenario of adaptive introgression from the subspecies *castaneus* to *musculus* in this region. Genealogical analysis with RELATE software estimates the spread occurred after approximately 1000 generations ago. Notably, the signature of selection was detected using only SNP data. If we had not investigated the polymorphic insertions of *Fv4*, the driving force of strong selective sweep would have been invisible to us.

Previous study has also revealed the adaptive introgression of another well-studied restriction factor, *Fv1* in the Japanese population. *Fv1*, encoding gag protein derived from MuERV-L, is present in the reference genome, but has two distinct alleles with different protein length, B and N, showing restriction to different types of MLVs. The distribution of N and B alleles is characteristic among subspecies, and Fujiwara et al. (2024) showed that the *castaneous*-type B allele has been adaptively introgressed into the Japanese population. The examples of *Fv1* and *Fv4* indicate the adaptive introgression of ERV-derived restriction factors play a key role in the evolution of genomes under secondary contact of genetically differentiated populations. Although we lack direct experimental evidence for the functional and selective effects of the ERV insertions shown in Table 1, these candidates could serve as valuable resources for studying the relationship between ERV integrations and adaptive evolution.

Conclusions

In this study, we focused on the patterns of selective sweeps in a genome-wide scan of SNPs. However, many other studies have explored genotype-phenotype relationships using SNP data obtained from genotyping arrays and short-read sequencing. In many cases, candidate loci span a certain length, making it difficult to pinpoint exact causal mutations due to linkage disequilibrium. While some causal variants may reside in non-coding transcripts or regions with epigenetic regulatory functions, insertions of non-reference sequences are often overlooked, particularly in genotyping array-based methods. Although our approach is not without limitations, we believe it offers an efficient way to find causal variants hidden in reference genomes. Until long-read sequencing becomes more cost-effective and widely adopted in laboratories, our method provides a valuable alternative for identifying these elusive genetic variants.

Materials and Methods

Algorithm of ERVscanner

ERVscanner is a pipeline designed to detect non-reference insertions of ERV sequences using short-read mapping files (BAM or CRAM) as input. All codes are distributed at Github (https://github.com/Qtom-99/ERVscanner). Genotypes of samples are output as a single VCF file.

It consists of four main processes: read filtering, identification of read clusters around insertions, filtering based on insertion content, and genotyping. A graphical representation is provided in Supplementary Figure 7. Although ERVscanner is specifically designed for ERV sequences, it can be applied to other classes of mobile elements. The pipeline utilizes pre-defined ERV regions in the reference genome, as annotated by the Dfam database (Storer, et al. 2021), which identifies these regions using a hidden Markov model. We refer to these annotated regions as masked regions (MRs). For further processing, ERVscanner requires BED-format files containing MRs and FASTA-format sequences of MRs. The MR sequences are reformatted to the sense (+) strand representation of ERVs, and consensus sequences of ERVs are included in the FASTA file. Additionally, any missing ERV sequences that are not represented in the reference genome can be manually added to the FASTA file.

Using the BED file, ERVscanner first extracts discordantly mapped read pairs, including those mapped on different scaffolds/chromosomes, separated by >10 kbp, or mapped with incongruent orientations, after removing PCR duplicates. The discordant pairs are then classified into two categories, and reads matching these conditions are extracted. (1) One read is mapped within a MR, while the other is mapped outside of an MR. In this case, only the read mapped outside of the MR is retained if it has a high mapping quality score (MQPQ ≥30). (2) Both reads are mapped

within MRs. In this case, any reads with high mapping quality scores (MQPQ ≥30) are retained. Note that detecting insertions within MRs is often difficult when using short-read data, and the false discovery rate is expected to be relatively high. Therefore, ERVscanner independently estimates insertions located inside and outside of MRs.

If the mapping was performed using BWA-MEM or other “ALT-aware” mapping tools, and not on telomere-to-telomere reference genomes, reads mapped to both primary and alternative/unplaced assembly components may have high mapping quality scores on primary chromosomal sequences. This could increase the false discovery rate of pipeline. To mitigate this, cluster reads with high mapping quality but having secondary alignments on alternative/unplaced assembly components were removed from the clusters.

After filtering the reads, we identified clusters indicative of non-reference ERV insertions. We expected reads to cluster with concordant orientations (forward-reverse pairs) flanking the insertion sites. For each forward and reverse read cluster, we merged regions covered by reads if they were separated by 200 bp or less. If the number of reads within a cluster exceeded the threshold, we defined the locus as a forward or reverse cluster. If forward and reverse clusters were detected within 200 bp of each other with the correct orientation, we classified them as candidate locus of non-reference insertions.

For each locus, we estimated the insertion content by re-mapping the reads that were paired with the cluster reads using BWA, against 361 non-redundant transposable element sequences in genus *Mus* in Dfam (Supplementary Table 4). We further re-assigned Long_Terminal_Repeat family to ERV1 and MaLR family to ERV3, according to the previous studies (Stocking and Kozak 2008; Kawase and Ichiyanagi 2023). Hereafter we refer to these sequences as the Dfam NRPH (non-redundant profile HMM) consensus. The Dfam NRPH consensus represents the consensus sequences of transposable elements without overlapped annotations. To improve accuracy, we gathered short-read sequences from multiple individuals, assuming that polymorphic insertions at the same loci should share identical insertion contents. When using the Dfam database, users can choose to classify insertions at the class level (e.g., ERV1–3) or at the family level (e.g., RLTR4_Mm). We recommend classification at the class level, as Dfam entries are often highly similar and difficult to distinguish. Clusters were considered valid if at least 70% of the mapped reads matched the same ERV class with the correct orientation.

For each valid cluster, we genotyped the insertions to determine whether insertions were heterozygous or homozygous. Within each valid cluster, clipped reads were extracted using CIGAR information to identify precise breakpoints. If clipped reads from forward and reverse clusters overlapped around a breakpoint, we assigned the minimum position as the breakpoint to

maintain consistency with the VCF format generated by general genotyping software. Heterozygous insertions were identified when at least three breakpoint-crossing reads covering a 30-bp window centered on the breakpoint. Valid clusters lacking a sufficient number of clipped reads were genotyped as '1/.' in the output VCF file, indicating the presence of an insertion but uncertainty regarding zygosity.

Evaluation of insertion calling using long-read sequences

Variant data based on long-read sequencing of the JF1 strain were downloaded from a publicly available website (https://molossinus.brc.riken.jp/pub/MoGplus30data/). The short-read data (DRA006630) was randomly down sampled to 30x coverage. Since the variant data was based on GRCm39, the coordinates of variants were converted to GRCm38 using the liftOver tool. To infer insertion contents from long-read sequencers, we extracted insertion sequences longer than 200 bp and performed a homology search using the BLASTN program (Altschul, et al. 1990) against Dfam NRPH consensus sequences. For each insertion, we identified the longest hit with ≥70% query coverage to classify the insertions by class and family. Insertions assigned to the ERV classes were selected for further analysis.

ERVcaller 1.4 was run using Dfam NRPH consensus sequences as seed sequences for insertions (Chen and Li 2019). Because we were unable to complete the analysis of the JF1 strain using the

default ERVcaller parameters, we reduced the internal parameter specified by -c in BWA-MEM (Li and Durbin 2009) for read extraction to 1/10 of the original value (from 100,000 to 10,000). This modification significantly reduced the running time but may have affected accuracy to some extent. After running ERVcaller, we extracted insertions classified as the ERV classes. The output VCF file is accessible at Zenodo (DOI: 10.5281/zenodo.17128997).

ERVscanner was used to estimate ERV insertions as described above. When the estimated breakpoints were within 40 bp, we considered them to be the same insertion.

Analysis of 163 house mouse short-read data

Short-read alignment files of 163 wild mouse samples determined by Fujiwara et al. (2024) were used for the analysis. Briefly, the file was constructed high-coverage whole-genome short-read sequences of 163 *M. musculus* by mapping to the GRCm38 reference genome.

We downloaded the annotation of ERV sequences for GRCm38 from Dfam database. Segments classified as ERV1 and ERV2-group were considered as candidate ERV sequences. The ERV2-group class includes ERV2, 3, 4, and Lentivirus families. In addition, RLTR4_MM-int family, which is not classified as the ERV class but classified as the Long_Terminal_Repeat_Element class, was included in the analysis because previous study classified RLTR4_MM-int as MLV-related sequences (Kawase and Ichiyanagi 2023). Using the

information, we constructed BED and FASTA files to represent the MRs. To further characterize known ERV-derived restriction factors, we added the sequences of *Fv4*, *RMCF1*, and *RMCF2* in the public database (DDBJ/EMBL/Genbank accession numbers: AF490353, AF490352, AY999005, AH001845, M33884, and AH001894)

We inferred insertions sites using ERVscanner, explained above, using five reads for the thresholds for each forward- and reverse-cluster, and three reads for the thresholds of breakpoint-crossing reads to judge the zygosity. PCA and PCA-UMAP were performed using scikit-learn library of Python. For the PCA-UMAP analysis, the first 50 principal components were used with minimum distance of 0.1 and number of neighbors of 15 for UMAP.

Detection of selective sweep

Population genetics summary statistics were calculated to detect the signature of selective sweep surrounding the non-reference ERV insertions. Nucleotide diversity (π), Tajima's *D, and* F_{ST} between Korean (11 individuals from Korean peninsula) and Northeastern Chinese (five individuals from Liaoning, Jiling, and Heilongjiang Provinces) samples were calculated using VCFtools (Danecek, et al. 2011), for 5-kbp-lengh windows with step size of 1000 bp.

For the calculation of nSL and XP-nSL statistics, we assigned ancestral alleles using the genotypes of seven *M. spretus*, and modified the VCF file using the information. Only the

monomorphic sites in *M. spretus* were used for the analysis. Both statistics were calculated using selscan 2.0 software with default parameters (Szpiech 2021). For XP-nSL, we used Korean samples for the test population and Northeastern Chinese samples for the reference population.

Genealogies surrounding the insertion sites of *Fv4* were inferred using RELATE software, using the same process described in Fujiwara et al. (2024), but in this study, we re-inferred population size change using only the chromosome 12 sequences of subspecies *musculus* samples with five rounds of iterations. In total 100 branch lengths were resampled using RELATE for CLUES.

Candidate of ERV insertions under adaptive evolution

We selected SNPs from Korean populations located within regions under selective sweep, identified using nSL statistics. SNPs in the top 0.1% of nSL scores were then pruned using PLINK with the --indep-pairwise 20 5 0.8 option. After pruning, we selected SNPs that (1) lacked exons within 10 kb and (2) were located within 1 kb of ERV insertions. Loci were retained if the allele frequency of the ERV insertion was equal to or greater than 0.5 in the Korean population and higher than that in the subspecies *castaneus.*

Acknowledgments

This work was supported by MEXT KAKENHI (grant numbers 22H02707 and 23H04846 to N. O.)

Tables

Table 1

Position	Class	Name	Length	Freq. Korea	Freq. *castaneus*
chr1:3723618	ERV2	IAPLTR1a_Mm	764	0.89	0
chr7:42156902	ERV2	RLTR10	2685	0.94	0.08
chr18:21435569	ERV1	RLTR6B_Mm	3387	0.76	0.17

Figure Legends

Figure 1

Venn diagram of detected non-reference ERV insertions in the JF1 strain genome. Long-read represents high-coverage PacBio data, while ERVcaller and ERVscanner utilized 30x short-read data.

Figure 2

Histogram of loci with non-reference ERV insertions among all wild house mouse individuals. The labels of subspecies were determined by the majority of genomic components inferred by ADMIXTURE software (Fujiwara et al. 2024).

Figure 3

PCA plot for 163 wild house mice, using non-reference ERV insertions as genetic markers. PC1 and PC2 were plotted.

Figure 4

Genomic signature of selective sweep around the insertion of *Fv4* in the Korean population. The vertical dashed line in the center of plot represents the estimated insertion site of *Fv4*. Dashed line in red and purple show the 0.01% thresholds for the top and bottom distributions, respectively.

For the XP-nSL and F_{ST} calculation, northeastern Chinese individuals, which also possess strong *M. m. musculus* genetic background, were used as the control population.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nature Reviews Genetics 11:17-30.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics 74:1111-1120.

Best S, Tissier PL, Towers G, Stoye JP. 1996. Positional cloning of the mouse retrovirus restriction gene Fvl. Nature 382:826-829.

Boso G, Lam O, Bamunusinghe D, Oler AJ, Wollenberg K, Liu Q, Shaffer E, Kozak CA. 2021. Patterns of Coevolutionary Adaptations across Time and Space in Mouse Gammaretroviruses and Three Restrictive Host Factors. Viruses 13:1864.

Boursot P, Din W, Anand R, Darviche D, Dod B, Von Deimling F, Talwar GP, Bonhomme F. 1996. Origin and radiation of the house mouse: mitochondrial DNA phylogeny. Journal of Evolutionary Biology 9:391-415.

Bowles H, Kabiljo R, Jones A, Al Khleifat A, Quinn JP, Dobson RJ, Swanson CM, Al-Chalabi A, Iacoangeli A. 2022. An assessment of bioinformatics tools for the detection of human endogenous retroviral insertions in short-read genome sequencing data. bioRxiv:2022.2002.2018.481042.

Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. 2016. Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. PLoS Computational Biology 12:e1004956.

Chen X, Li D. 2019. ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. Bioinformatics 35:3913-3922.

Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, Park PJ. 2021. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nature Communications 12:3836.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker

RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Dodding Mark P, Bock M, Yap Melvyn W, Stoye Jonathan P. 2005. Capsid Processing Requirements for Abrogation of Fv1 and Ref1 Restriction. Journal of Virology 79:10571-10577.

Edelmann W, Kröger B, Goller M, Horak I. 1989. A recombination hotspot in the LTR of a mouse retrotransposon identified in an in vitro system. Cell 57:937-946.

Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets. Molecular Biology and Evolution 31:1850-1868.

Fujiwara K, Kawai Y, Takada T, Shiroishi T, Saitou N, Suzuki H, Osada N. 2022. Insights into Mus musculus Population Structure across Eurasia Revealed by Whole-Genome Analysis. Genome Biology and Evolution 14.

Fujiwara K, Kubo S, Endo T, Takada T, Shiroishi T, Suzuki H, Osada N. 2024. Inference of selective forces on house mouse genomes during secondary contact in East Asia. Genome Research 34:366-375.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Consortium TGP, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Research 27:1916-1929.

Goff SP. 2004. Retrovirus Restriction Factors. Molecular Cell 16:849-859.

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497-501.

Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194-197.

Ikeda H, Sato H, Odaka T. 1981. Mapping of the Fv-4 mouse gene controlling resistance to murine leukemia viruses. International Journal of Cancer 28:237-240.

Inaguma Y, Miyashita N, Moriwaki K, Huai WC, Jin ML, He XQ, Ikeda H. 1991. Acquisition of two endogenous ecotropic murine leukemia viruses in distinct Asian wild mouse populations. Journal of Virology 65:1796-1802.

Kawase M, Ichiyanagi K. 2023. Mouse retrotransposons: sequence structure, evolutionary age, genomic distribution and function. Genes and Genetic Systems

98:337-351.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289-294.

Keane TM, Wong K, Adams DJ. 2012. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29:389-390.

Kent TV, Uzunović J, Wright SI. 2017. Coevolution between transposable elements and recombination. Philosophical Transactions of the Royal Society B: Biological Sciences 372:20160458.

Koide T, Moriwaki K, Uchida K, Mita A, Sagai T, Yonekawa H, Katoh H, Miyashita N, Tsuchiya K, Nielsen TJ, et al. 1998. A new inbred strain JF1 established from Japanese fancy mouse carrying the classic piebald allele. Mammalian Genome 9:15-19.

Kojima S, Koyama S, Ka M, Saito Y, Parrish EH, Endo M, Takata S, Mizukoshi M, Hikino K, Takeda A, et al. 2023. Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. Nature Genetics 55:939-951.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754-1760.

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. Nature Reviews Genetics 21:597-614.

Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. Genome Research 19:922-933.

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

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Research 15:1566-1575.

Odaka T, Ikeda H, Moriwaki K, Matsuzawa A, Mizuno M, Kondo K. 1978. Genetic Resistance in Japanese Wild Mice (Mus musculus molossinus) to an NB-Tropic Friend Murine Leukemia Virus2. JNCI: Journal of the National Cancer Institute 61:1301-1306.

Rogers RL, Grizzard SL, Garner JT. 2023. Strong, Recent Selective Sweeps Reshape Genetic Diversity in Freshwater Bivalve Megalonaias nervosa. Molecular

Biology and Evolution 40.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetical Research 23:23-35.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics 51:1321-1329.

Stern AJ, Wilton PR, Nielsen R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genetics 15:e1008384.

Stocking C, Kozak CA. 2008. Endogenous retroviruses. Cellular and Molecular Life Sciences 65:3383-3398.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12:2.

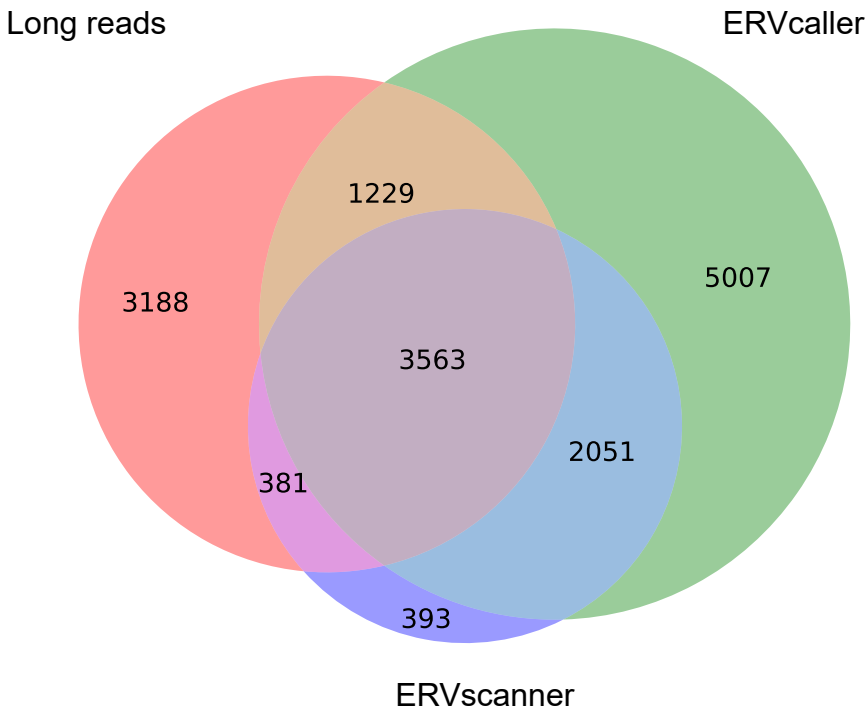
Stoye JP, Coffin JM. 1987. The four classes of endogenous murine leukemia virus: structural relationships and potential for recombination. Journal of Virology 61:2659-2669.

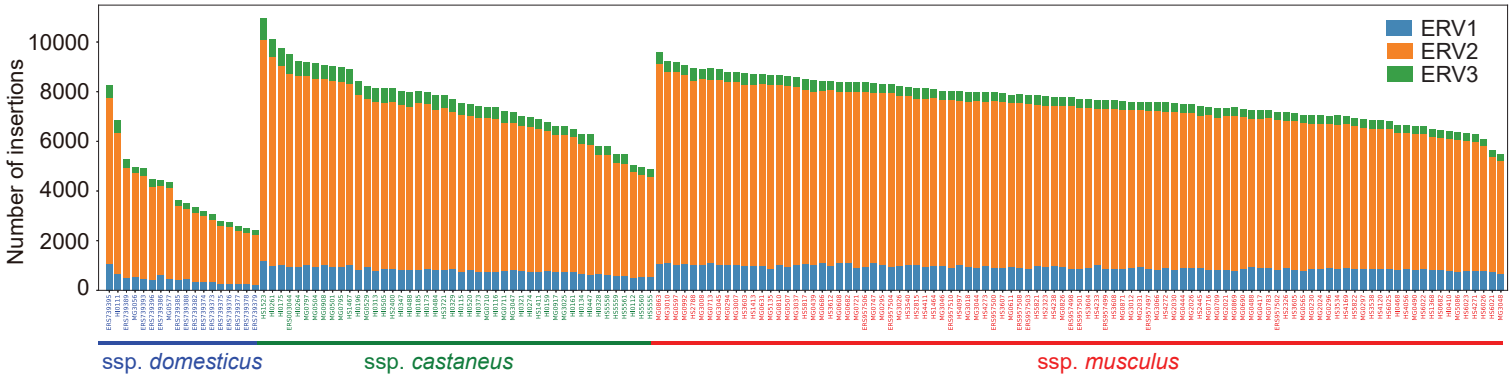
Szpiech ZA. 2021. selscan 2.0: scanning for sweeps in unphased data. bioRxiv:2021.2010.2022.465497.

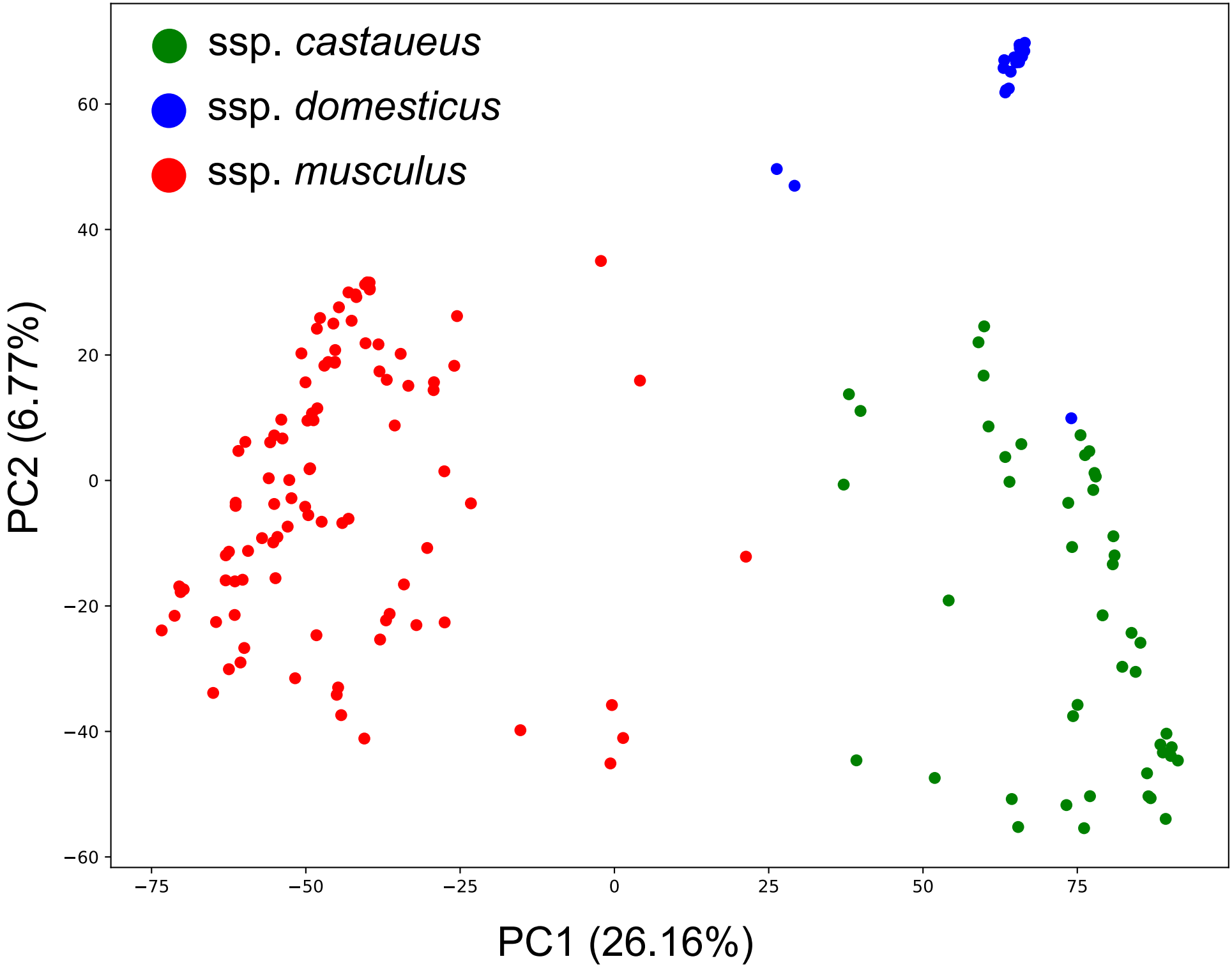
Takada T, Ebata T, Noguchi H, Keane TM, Adams DJ, Narita T, Shin-I T, Fujisawa H, Toyoda A, Abe K, et al. 2013. The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. Genome Research 23:1329-1338.

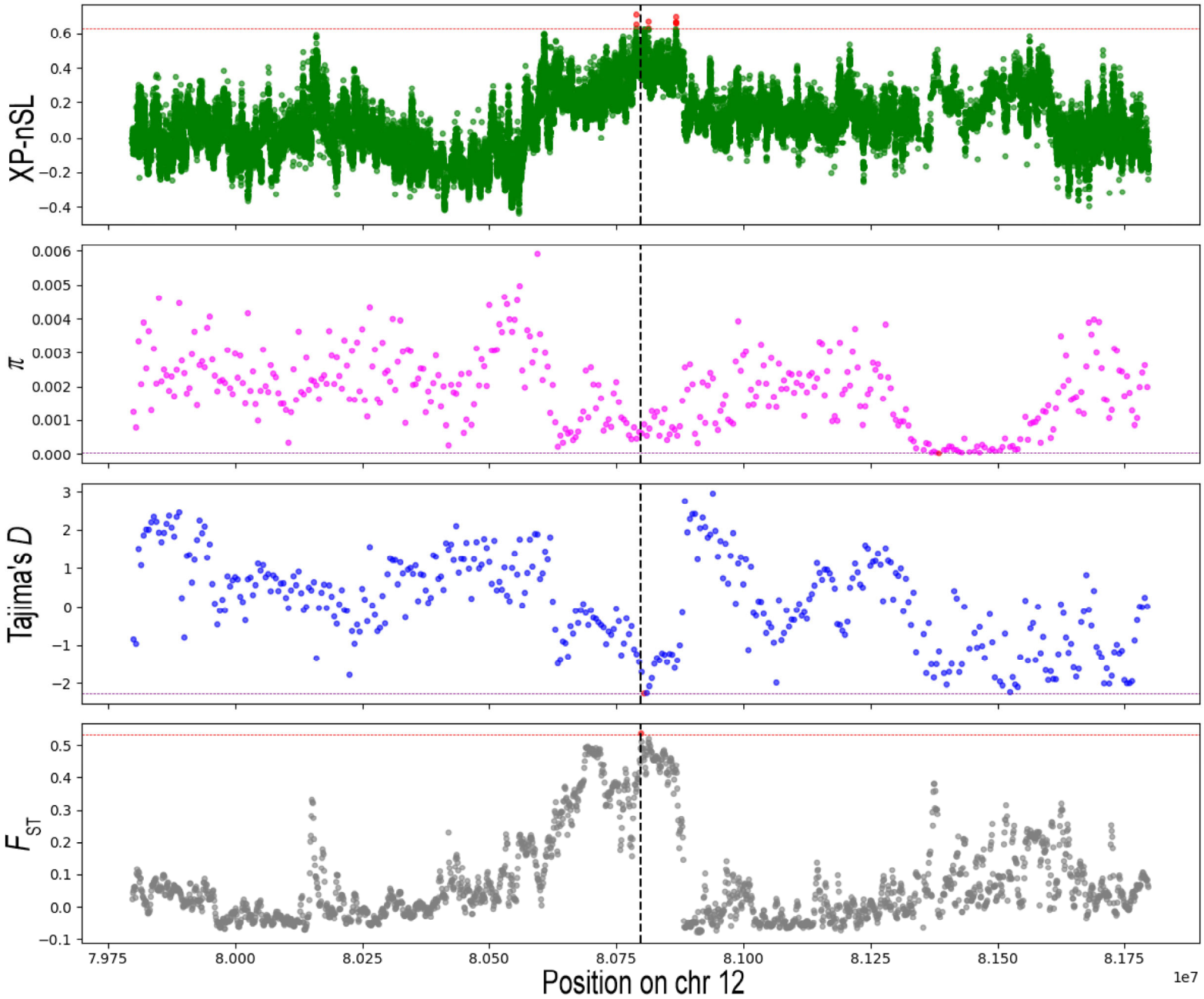
Weedall GD, Riveron JM, Hearn J, Irving H, Kamdem C, Fouet C, White BJ, Wondji CS. 2020. An Africa-wide genomic evolution of insecticide resistance in the malaria vector Anopheles funestus involves selective sweeps, copy number variations, gene conversion and transposons. PLoS Genetics 16:e1008822.

Wu T, Yan Y, Kozak Christine A. 2005. Rmcf2, a Xenotropic Provirus in the Asian Mouse Species Mus castaneus, Blocks Infection by Polytropic Mouse Gammaretroviruses. Journal of Virology 79:9677-9684.


Long reads
ERVcaller
ERVscanner
1229
3188
5007
3563
2051
381
393


Number of insertions
10000
8000
6000
4000
2000
0
ERV1
ERV2
ERV3
ssp. domesticus
ssp. castaneus
ssp. musculus


ssp. castaueus
ssp. domesticus
ssp. musculus
PC2 (6.77%)
PC1 (26.16%)
60
40
20
0
−20
−40
−60
−75
−50
−25
0
25
50
75


XP-nSL
0.6
0.4
0.2
0.0
−0.2
−0.4
π
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Tajima's D
3
2
1
0
−1
−2
F_ST
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
7.975
8.000
8.025
8.050
8.075
8.100
8.125
8.150
8.175
1e7
Position on chr 12