DNA transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans

Shengjun Tan\textsuperscript{1,2,14}, Huijing Ma\textsuperscript{1,2,14}, Jinbo Wang\textsuperscript{1,2,14}, Man Wang\textsuperscript{4,14}, Mengxia Wang\textsuperscript{1,2,3}, Haodong Yin\textsuperscript{1,3}, Yaqiong Zhang\textsuperscript{1,2}, Xinying Zhang\textsuperscript{1,3}, Jieyu Shen\textsuperscript{1,2,3}, Danyang Wang\textsuperscript{3,5}, Graham L. Banes\textsuperscript{6,7}, Zhihua Zhang\textsuperscript{3,5}, Jianmin Wu\textsuperscript{4}, Xun Huang\textsuperscript{3,8}, Hua Chen\textsuperscript{3,9,10}, Siqin Ge\textsuperscript{1,3}, Chun-Long Chen\textsuperscript{11,12} & Yong E. Zhang\textsuperscript{1,2,3,10,13} \textsuperscript{✉}

Despite long being considered as “junk”, transposable elements (TEs) are now accepted as catalysts of evolution. One example is Mutator-like elements (MULEs, one type of terminal inverted repeat DNA TEs, or TIR TEs) capturing sequences as Pack-MULEs in plants. However, their origination mechanism remains perplexing, and whether TIR TEs mediate duplication in animals is almost unexplored. Here we identify 370 Pack-TIRs in 100 animal reference genomes and one Pack-TIR (Ssk-FB4) family in fly populations. We find that single-copy Pack-TIRs are mostly generated via transposition-independent gap filling, and multicopy Pack-TIRs are likely generated by transposition after replication fork switching. We show that a proportion of Pack-TIRs are transcribed and often form chimeras with hosts. We also find that Ssk-FB4s represent a young protein family, as supported by proteomics and signatures of positive selection. Thus, TIR TEs catalyze new gene structures and new genes in animals via both transposition-independent and -dependent mechanisms.

\footnote{1 Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. 2 State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. 3 University of Chinese Academy of Sciences, Beijing, China. 4 Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Center for Cancer Bioinformatics, Peking University Cancer Hospital & Institute, Beijing, China. 5 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, and China National Center for Bioinformation, Chinese Academy of Sciences, Beijing, China. 6 Wisconsin National Primate Research Center, University of Wisconsin–Madison, Madison, WI, USA. 7 CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China. 8 State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. 9 CAS Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, and China National Center for Bioinformation, Chinese Academy of Sciences, Beijing, China. 10 CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. 11 Curie Institute, PSL Research University, CNRS UMR 3244, Paris, France. 12 Sorbonne University, Paris, France. 13 Chinese Institute for Brain Research, Beijing, China. 14 These authors contributed equally: Shengjun Tan, Huijing Ma, Jinbo Wang, Man Wang. E-mail: chunlong.chen@curie.fr; zhangyong@ioz.ac.cn}
Transposable elements (TEs), including retrotransposons and DNA transposons, occupy a significant portion of eukaryotic genomes\(^2\). Although long considered "junk DNA"\(^2\), TEs are now widely accepted as catalysts of genetic innovations by directly contributing to regulatory or coding sequences\(^3,4\) and mediating sequence changes such as duplications or deletions\(^5,6\). The mechanism responsible for the generation of duplicates affects their evolutionary trajectories\(^7,8\). Duplicates generated by TEs are more likely to evolve new structures or functions due to the formation of chimeric transcripts or changes in the regulatory context\(^9,10\). Therefore, the mechanism through which TEs mediate duplications is of broad interest.

Studies, including ours, have shown that long terminal repeats (LTRs) and non-LTR retrotransposons, such as L1 or SVA elements, mediate the retroduplication of host messenger RNAs (mRNAs) in animals and plants\(^10,12-15\). Among DNA transposons, Helitrons duplicate non-TE sequences in animals and plants\(^16,17\), whereas terminal inverted repeat TE (TIR TE)–mediated duplications have often been studied in plants but not in animals\(^18,19\). Two anecdotal studies in animals show that P elements in Drosophila capture sequences via two mechanisms upon artificial activation of the transposase. First, P elements together with flanking sequences are subject to transposition under transduction (also called the end bypass model, Supplementary Fig. 1a) and thereby accidentally use the downstream sequence as the TIR\(^20\). Second, P elements capture sequences under the gap-filling model\(^21\). In this model, double-strand breaks (DSBs) occur in two scenarios: (1) the internal sites are broken, possibly induced by secondary structures\(^22,23\) (Supplementary Fig. 1b); or (2) complete TEs are excised due to transposition\(^24\) (Supplementary Fig. 1c). During the repair, the template could switch from the sister strand to adjacent external sequences (called fillers) in 3D proximity, leading to the capture of fillers\(^21,25\). The whole process in the former scenario (Supplementary Fig. 1b) is transposition-independent.

In contrast, TIR TE-capturing sequences have been extensively studied in plants, particularly rice: one type of TIR TE called Mutator-like element (MULE) generates Pack-MULEs (duplicated internal sequences together with flanking MULEs)\(^26,28\). One-third of Pack-MULEs are multicopy due to several rounds of transposition, as indicated by distinct target site duplications (TSDs, a hallmark of transposition)\(^26\). The internal sequences are often duplicated in trans and derived from interchromosomal sequences\(^27,29\). Because the source or parental copies are not linked with MULEs or Pack-MULEs, the origination mechanism of Pack-MULEs is incompatible with the end bypass model. The gap-filling model has, therefore, been proposed\(^27\), but whether duplication is associated with transposition remains unknown\(^18\). Functionally, 40% of Pack-MULEs are transcribed\(^28\) and possibly encode small RNAs or contribute to the 5’ untranslated region (UTR) of host genes\(^29,30\). Protein-coding Pack-MULEs are rare, and only one protease duplicated by a MULE (KI-MULE) has been characterized. However, KI-MULE is likely not functional given its repressed expression and heterochromatic location\(^31\).

Here, we consider whether Pack-TIRs (non-TE sequences with flanking TIR TEs) are present in animals, and if so, how they emerge and whether they are functional. To address these questions, we focus on young Pack-TIRs, which more likely retain sequence features indicating their origination mechanism compared with older Pack-TIRs. We identify a conservative dataset that includes 370 Pack-TIRs in 100 animal reference genomes and one Pack-TIR (Ssk-FB4, the gene Ssk amplified by one TIR TE called FB4) family in D. melanogaster populations. Sequence analyses of these Pack-TIRs suggest that single-copy Pack-TIRs are mostly generated via a transposition-independent gap-filling process, whereas the birth of multicopy Pack-TIRs is compatible with a new model, which we describe as replication Fork Stalling, Template Switching and Transposition (FoSTeST). Furthermore, we find that an appreciable proportion of Pack-TIRs are transcribed and often chimeric with neighboring genes, and Ssk-FB4s may represent one of the youngest functional protein families supported by unique mass spectrometry (MS) peptides and signatures of positive selection. In summary, TIR TEs generate new gene structures and new genes in animals via both transposition-independent and -dependent mechanisms.

## Results

### The distribution, copy number, and origination timing of Pack-TIRs in reference genomes suggest a transposition-independent birth process.

To produce an overview of Pack-TIRs harbored by the reference genomes, we analyzed their distribution across species and TE superfamilies. We found that 370 Pack-TIRs were scattered across 55 species with a median number of 4 (Fig. 1a and Supplementary Data 1). The number of Pack-TIRs was correlated with that of consensus TIR TEs in each animal, with approximately two Pack-TIRs per 10,000 TIR TEs (Fig. 1b, \(R^2 = 0.60\)). Consistent with the literature\(^34\), zebrafish, western clawed frogs, and American alligators were found to be the top three species, with each including more than 120,000 consensus TIR TEs (Supplementary Data 2). These species accordingly encode relatively more Pack-TIRs (8, 25, and 38, respectively). In contrast, probably due to the low content of TIR TEs in birds and insects (median number of 268 and 36, respectively), no Pack-TIR was identified within these species. We found that the distribution of Pack-TIRs across TIR TE superfamilies was analogous to the number of Pack-TIRs predicted based on consensus TIR TEs in each superfamily (Fig. 1c, \(R^2 = 0.94\)). Most (323 or 87.2%) Pack-TIRs were associated with KI-MULEs (8, 25, and 38, respectively), no Pack-TIR was identified within these species. We found that the distribution of Pack-TIRs across TIR TE superfamilies was analogous to the number of Pack-TIRs predicted based on consensus TIR TEs in each superfamily (Fig. 1c, \(R^2 = 0.94\)). Most (323 or 87.2%) Pack-TIRs were associated with the top two most common superfamilies, hAT (208) and TcMariner (115) (Supplementary Data 1), which jointly contributed 81.8% of the TIR TE content in the 100 species (Supplementary Data 2). Although MULEs are active in rice, they are rare (median number of 0) in animals, and no Pack-MULE has been identified in reference genomes.

Because different TE families could exhibit different extents of transposition activity, the linear relationship between the number of Pack-TIRs and consensus TIR TEs appears to suggest a transposition-independent duplication process. More direct evidence came from the inactivity of many consensus TEs\(^35,36\). In actuality, the consensus sequences of TEs tend to be short, with a median length of only 374 base pairs (bp). Moreover, individual TE copies are different from the consensus by a median divergence value of 18.3% (Supplementary Data 1). These two patterns suggest that TEs likely represent degenerate ancient relics. We thus hypothesized that the majority of Pack-TIRs...
Pack-MULEs observed in rice (36%, 481/1337) markedly lower (Fisher's exact test, $P < 2.2 \times 10^{-16}$).

In summary, the linear relationships among the numbers of Pack-TIRs and consensus TEs, the prevalence of single-copy Pack-TIRs, and the origination timing of Pack-TIRs in primates jointly suggest that these were mainly generated via a transposition-independent mechanism.

The location and sequence features of single-copy Pack-TIRs indicate a gap-filling process. We further examined the location and sequence features of Pack-TIRs and found three patterns supporting the gap-filling model. First, the gap-filling model predicts a cis duplication bias in which TEs preferentially capture nearby sequences as fillers. Consistently, for 281 unique Pack-TIRs in the reference genomes (after controlling for redundant ones in primates and multicopy cases, Supplementary Data 1), TIR TEs capturing sequences from the same chromosome harbors 33 Pack-TIRs, including 12 cases directly called by our pipeline (Fig. 1a) and 21 cases called at orthologous loci in nonhuman primates but exhibiting slightly lower alignment quality in humans (not passing the cutoffs, Supplementary Data 3, "Methods"). For these 33 cases, the majority (22 or 66.7%) were generated when DNA TEs were inactive (Fig. 1e), which invalidates the transposition-dependent model.
(intrachromosomal duplication) were widespread in animals (59.8%, Fig. 2a). This proportion could be even higher because different scaffolds possibly belong to the same chromosome in species with only draft assemblies. The density of these intrachromosomal Pack-TIRs was negatively correlated with their distance to the parental copies (Fig. 2b), which is similar to P element-induced gap repair21. Furthermore, 160 (56.9%) Pack-TIRs were situated in the proximity of parental copies (<15 kilobases or kb, Fig. 2a, b), and the distance in the majority of these cases was less than 5 kb. Such a cis bias applies for both species currently with active TEs, such as frogs (60.9%), and species currently with active TEs, such as rabbits (67.6%) Pack-TIRs (Fig. 2c). Such a pattern applies for both intrachromosomal Pack-TIRs and single-copy or multicopy Pack-TIRs (Fig. 2d). We took the breakpoints of 33 human Pack-TIRs as examples to examine whether the microhomology length could be explained by the random association of TEs and fillers (“Methods”). Most (84.4%) did not adhere to the chance effect (P < 0.05, Supplementary Fig. 3), which supports the significance of microhomology during template switching.

Third, the gap-filling process induced by the fragile site (Supplementary Fig. 1b) might indicate a longer waiting time between TE insertion and non-TE sequence capture compared with that obtained with the process induced by active transposon excision (Supplementary Fig. 1c). Thus, TIR TEs without duplications should be present at the orthologous locus of outgroup species if duplications are specific to the focal species or lineages. We tested this hypothesis in primates due to their densely sampled phylogeny (Fig. 1e). The majority (91 out of 97, or 94%) fit the expected scenario where orthologous TIR TEs without duplications existed in the outgroup species (Supplementary Data 3). Taking Pack-Tigger1 (a TcMariner element) in gibbons as an example, all four outgroup primates encode orthologous Tigger1 (Fig. 2e). The analysis of five other Pack-TIRs revealed that no orthologous TEs or TSDs (excision relics) could be detected in related primates (Supplementary Fig. 4a–e). Because these five cases were from sparsely sampled lineages (e.g., tarsier, Fig. 1f), the absence of homologous TEs could be
explained by the lack of closely related outgroups. The last Pack-TIR was shared by primates (Supplementary Fig. 4f), which suggested its ancient origin. Thus, we could not accurately infer its origination process.

Finally, we analyzed whether DSBs induced by breaks within TEs (Supplementary Fig. 1b) or by TE excisions via transposition (Supplementary Fig. 1c) initiated the gap-filling process. We expected that the abortive gap repair in the latter scenario would generate larger deletions with breakpoints biased to the terminals of TEs compared with those obtained in the former scenario. We found evidence compatible with the former scenario: the distances between two breakpoints only account for a median of 4.6% TE sizes (Fig. 2f), and the breakpoints were moderately but significantly skewed toward the internal region of TEs compared with the random uniform distribution (Fig. 2g, "Methods"). The skew is likely due to the enrichment of internal fragile sites.

Thus, the observations of cis duplication bias, prevalence of microhomology at breakpoints, post-transposition duplication, and biased distribution of breakpoints toward internal rather than terminal TEs support the notion that a transposition-independent gap-filling process underlies most single-copy Pack-TIRs encoded by animal reference genomes.

Recurrent transposition generates multicopy Pack-TIRs. We extended the comparative methodology to four multicopy Pack-TIRs in frogs and alligators (Fig. 2a and Supplementary Fig. 2). Contrary to most single-copy Pack-TIRs in primates, we could not find orthologous TIR TEs for all four Pack-TIRs. Instead, the syntenic regions do not encode Pack-TIRs together with nearby flanking regions. Taking Pack-hAT-N8_XT (Fig. 1d) as an example, all three derived Pack-TIRs or hAT-N8_XT without duplication were absent in the orthologous loci of the outgroup, the African clawed frogs (Supplementary Fig. 5), which diverged from the western clawed frogs 57 Mya. Thus, either the multicopy cases do not fit the gap-filling model or the outgroup species is too divergent to provide sequence information indicating the origination process of these Pack-TIRs.

We thus analyzed the multicopy polymorphic Ssk-FB4s using the D. melanogaster reference genome as the outgroup. Similar to four multicopy Pack-TIRs encoded by frog and alligator reference genomes (Fig. 2a and Supplementary Fig. 2), Ssk-FB4s show features of recurrent transposition and long distances between parent and Pack-TIRs. Specifically, the protein-coding gene Ssk (Snakeskin, essential for intestinal barrier function) encoded in chromosome 3L (coordinate, 20.2 Mb) was almost completely (5’ upstream together with the genic region except ~200-bp 3’ UTR) captured by FB4, which was annotated as a TcMariner element in Repbase (Fig. 3a). In DGRP lines, we found three Ssk-FB4 loci, including one on chromosome X (coordinate, 2.7 Mb) and two on chromosome 3R (coordinates, 14.3 Mb and 17.7 Mb, respectively) (Supplementary Data 5). The locus at chr3R: 17.7 Mb was further tandemly duplicated. Because DGRP represents a northern American population, we additionally surveyed the global diversity lines (GDLs) of D. melanogaster and confirmed that these populations also harbor only these four copies (Supplementary Data 6). All three loci of Ssk-FB4 share the same chimeric structure but distinct TSDs (TACATATAG at chr3R: 14.3 Mb, AAAATTTAAC at chr3R: 17.7 Mb, and CATGTTAGCG at chrX: 2.7 Mb), which suggests recurrent transpositions. Consistently, FB4 is known as a nonautonomous element transposed by an unknown transposase. Notably, since FB4 has long TSDs (9 to 10-bp) and TIRs (~700 bp), it more likely belongs to MULE rather than TcMariner superfamily, which generally have short TSDs (2-bp TA) and TIRs (<100 bp). Thus, Ssk-FB4s represent Pack-MULEs in Drosophila.

To examine the origination process of Ssk-FB4s, we determined which might be the first Ssk-FB4 to initially emerge and which FB4 element might be involved. By split-read-based genotyping ("Methods"), we found that the X-linked copy was present in relatively more lines (24.2% vs. 6.8–16.5%, Fig. 3b), which suggests its earlier origin. Consistently, by reconstructing the phylogenetic tree of the four copies with Ssk as the outgroup, we found that the X-linked copy was the most similar to the parent, as shown by the shortest branch length (Fig. 3c). Thus, we inferred that X-linked Ssk-FB4 is the founding member of the whole multicopy family. If the aforementioned transposition-independent gap-filling model works, an FB4 element would be present around chrX: 2.7 Mb in fly populations. However, we did not find FB4 insertion or TSD (excision relic) in the fly reference genome or populations including DGRP and GDL, which suggested that FB4 is not situated in this locus. We, therefore, inferred that Ssk-FB4 did not fit the gap-filling model. Among the six full-length FB4s present in reference genome, we deduced that the one at chr3L: 20.8 Mb likely mediated the emergence of Ssk-FB4 given its highest sequence similarity relative to Ssk-FB4 (Supplementary Fig. 6) and the high population frequency (present in five out of six sampled lines, Supplementary Data 7).

Motivated by the concurrence of replication and transposition and a replicative mechanism shaping structural variations, i.e., replication Fork Stalling and Template Switching (FoSTeS), we hypothesized that a new model, named replication Fork Stalling, Template Switching and Transposition (FoSTeST), drove the recombination of Ssk at chr3L: 20.2 Mb and FB4 at chr3L: 20.8 Mb and the subsequent shuffling to chrX: 2.7 Mb. We found two lines of evidence supporting this model. Specifically, the TIR of FB4 is repetitive and associated with frequent rearrangements. Consistently, we found that all six FB4s harbor structural variations (e.g., tandem duplications, Supplementary Data 7). Thus, replication fork stalling and DSBs occur in these repetitive regions, which leads to frequent template switching. Furthermore, similar to the cis duplication bias in the gap-filling model (Fig. 2a, b), Ssk and FB4 (chr3L: 20.8 Mb) are spatially close due to high-order DNA folding, which is supported by DNA interaction data (Fig. 3d and Supplementary Table 1, "Methods"). A 2-bp microhomology (AA) at one breakpoint further facilitates the switching, although a 45-bp de novo insertion was situated at the other breakpoint (Supplementary Fig. 7). A transposon captured this newborn Ssk-FB4 and moved it to chrX, whereas the 3L-linked FB4 was repaired because Ssk-FB4 could not be detected in this locus within fly populations. Subsequently, the X-linked Ssk-FB4 was subjected to further transpositions.

Taken together, the absences of orthologous TEs without duplication across multicopy Pack-TIRs in frogs, alligators, and flies suggest that multicopy Pack-TIRs were likely generated through the FoSTeST process.

An appreciable proportion of Pack-TIRs are transcribed and the majority of these are expressed as chimeric RNAs with flanking sequences. We then inferred whether Pack-TIRs are potentially functional by analyzing their sequence and expression features. First, the size of the duplicated regions was small, with a median size of 224 bp (Fig. 4a), which was similar to that of Pack-MULEs (305 bp). The size of multicopy Pack-TIRs appeared to be larger, although this difference was not statistically significant (Fig. 4a). Despite their small size, 13.4% of Pack-TIRs captured exonic sequences, and this value was 3.2-fold higher than the...
background proportion (FET P < 0.001, Supplementary Fig. 8). This excess could be underestimated because most animal genomes are likely underannotated. Consistently, we found a stronger pattern (21.2% or 4.5-fold, Fig. 4b) in the human genome, which was better annotated. Disproportionate exonic duplications suggest that they are more likely expressed and maintained by natural selection. By analyzing a transcriptome dataset covering four tissues, we found a consistent pattern: five out of eight (62.5%, Fig. 4b) exonic duplications were transcribed, as supported by at least five unique mapping reads (“Methods”), which was significantly higher (FET P = 0.036) than intronic and intergenic duplications (5/18 and 0/7, respectively).

We reconstructed the gene structure of 10 transcribed Pack-TIRs in humans via targeted de novo transcriptome assembly of strand-specific reads (“Methods”) and found that the captured sequences were always cotranscribed with at least one flanking TE and that the majority (8) were further fused with adjacent genes. Specifically, two Pack-TIRs with exonic sources and four with intronic sources serve as part of retained introns (Fig. 4c and Supplementary Fig. 9a). This pattern is consistent with the prevalence (75% of genes) of intron retention in mammals, which regulates transcription and splicing. A second scenario is 3’ UTR elongation: a coding exon together with part of neighboring introns of ERPP4 was coopted as part of the 3’ UTR of general transcription factor IIB subunit 5 (GTF2H5, Fig. 4d), and one intronic region of H2B clustered histone 15 (H2BC15) was duplicated and coopted as part of its own 3’ UTR. Considering the essentiality of GTF2H5 and H2BC15, modification of their 3’ UTRs might have a functional consequence. Note: the transcription of these eight cases is likely driven by promoters of host genes because their transcriptional orientation is the same as that of the host genes in seven cases and their expression breadth is identical to that of the host genes in six cases (e.g., GANQ-MER5A, Fig. 4c). The last two cases are relatively long: a retrogene ARFG was almost completely captured into the MER20 transposon, but this duplication could also reflect a
Fig. 4 Sequence and transcriptional features of Pack-TIRs. a Length distribution of the internal sequences (n = 281 samples). The distribution is shown as a violin plot in Fig. 2f. The four red dots represent the multicity Pack-TIRs. b Distribution of parental sequence types in humans. "Observed" shows the actual numbers of Pack-TIRs divided by the functional types of their parental sequences; "Expected" shows the numbers calculated based on the proportion of all annotated exons and introns; and "Transcribed" shows the numbers of transcribed Pack-TIRs in each category. c Transcribed human Pack-TIRs. Host genes indicate the genes into which Pack-TIRs were inserted. “B”, “H”, “L”, and “T” indicate the brain, heart, liver, and testis, respectively. d One human example showing 3' UTR chimerism. The figure convention follows that of Fig. 1d with the exception that the host gene is marked in orange and the transcription direction is shown by thick arrows. e One worm example showing 5' UTR chimerism. The parental gene is close to the host gene at a 4-kb distance. The boxes connected by dashed lines indicate the spliced RNA-sequencing (RNA-seq) reads. Blank arrows indicate paired-end Illumina short reads. The bottom long read was generated on the Nanopore platform (Methods). f Gene structure of Ssk-FB4 in D. melanogaster at chrX: 2.7 Mb. Ssk-FB4 is highly transcribed in the midgut, and only a few representative spliced reads spanning breakpoints between the internal sequence and TE and reads supporting continuous transcription toward 165 bp (marked by a red dashed line) downstream regions are shown. Note: the reads containing poly(A) tails have non-A residues, consistent with a previous report46. g Heatmap showing the expression profile of Ssk, Ssk-FB4s, and FB4s in five tissues in D. melanogaster. The expression levels were quantified as log$_2$ transformed TPM (transcripts per million mapped reads, n = 2 biologically independent replicates).

We extended analogous analyses to model invertebrates, including worms and flies, and found similar chimerism between Pack-TIRs and neighboring sequences. For four Pack-TIRs encoded by the worm reference genome, we analyzed a whole-body RNA-seq dataset and found that two were expressed as chimeric 5' UTR and ncRNA, respectively. The former case is worth noting: part of the 3' UTR and the terminal coding exon of $F43E2.6$ were copied to the upstream of the nearby essential gene, i.e., the RNA polymerase II subunit (rpb-4). Short- and long-read RNA-seq data revealed two isoforms, and Pack-TIRs provided alternative 5' UTRs (Fig. 4e). Thus, similar to GTF2H5 and $H2BC1S$, the gene structure of rpb-4 was changed, which could have a functional consequence. Note: because this Pack-TIR was inserted into the upstream region of rpb-4 with the intervening gene $mtch-1$ on the antisense strand, its transcription appears driven by the 5' flanking transposon.

The polymorphic Ssk-FB4s in flies are expected to be expressed under the joint regulatory context of Ssk and FB4 (Fig. 3a). To test this hypothesis, we analyzed RNA-seq data (“Methods”) by taking advantage of the unique nucleotides of Ssk and individual Ssk-FB4. We first reconstructed the gene structure of the copy with the highest frequency, i.e., X-linked Ssk-FB4 (Fig. 3b). We found numerous reads compatible with the original gene structure of Ssk and poly(A) tail-containing reads, which indicated a new transcription termination site 165 bp downstream of Ssk-FB4 (Fig. 4f). Thus, due to loss of the 3' terminus, X-linked Ssk-FB4 was fused with the adjacent sequence. Analogously, with the exception of Ssk-FB4 at chr3R: 14.3 Mb, for which we lacked the corresponding DGRP lines, we identified reads supporting the recurrent retrotransposition event (Supplementary Fig. 9b, “Methods”); the 3' UTR of SP1 was duplicated with MER20 as a noncoding RNA (ncRNA, Supplementary Fig. 9c). In both cases, their expression appears shaped by flanking host regions and/or MER20 transposons with known promoter activity.31

| Parental copy | Host gene | Pack-TIR | Strandness |
|---------------|-----------|----------|------------|
| Gene name     | Region    | Expression | Region     | Expression |
|               |           |           | Strandness |
| GNA2          | Exon      | Intron    | B, H, L, T | Same       |
| RAB6A         | Intron    | Intron    | B, H, L, T | Same       |
| MYO1B         | Intron    | Intron    | H, L       | Same       |
| ARFGEF1       | Intron    | Intron    | B, H, L, T | Same       |
| RAR8          | Intron    | Intron    | Intron T   | Same       |
| PTPRE         | Intron    | Intron    | L, T       | Same       |
| ERPF44        | Exon+Intron| T      | Exon T     | Same       |
| H2BC15        | Intron    | Intron    | B, H, L, T | Same       |
| ARF6S         | Exon      | Exon+Intron| T         | Both strands |
| SP1           | Exon      | Exon      | Intergenic | T           |

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-24585-9 | www.nature.com/naturecommunications
Ssk-like gene structure for Ssk-FB4 at chr3R: 17.7 Mb and a tandemly duplicated locus (Supplementary Fig. 10). Based on Ssk-derived regions, we quantified the expression of Ssk and three Ssk-FB4s across five representative tissues and identified four patterns (Fig. 4g): (1) Ssk-FB4s are similarly expressed across lines and across copies (Pearson $r^2 = 0.93$, $P = 0.023$), suggesting analogous transcriptional control; (2) similar to Ssk, Ssk-FB4s are predominantly expressed in the midgut with a median expression reaching 120, which is higher than 95.0% of genes (Supplementary Fig. 14a). The lower divergence data of Ssk-FB4s and Ssk. The upper polymorphism data are based on the codon-level alignment of Ssk-FB4s (Supplementary Fig. 14a). The lower divergence data of Ssk are based on the whole Drosophila genus (Supplementary Fig. 14e). Each cell is shown as the “number of changed sites (observed/expected proportion of changed sites)”. $P$-values were calculated for the regions of interest with the one-sided proportion test. (d) Allele frequency spectrum of FB4s and Ssk-FB4s in 83 DGRP lines. (e) Transcription of Ssk, mesh, and Tsp2A in the gut across lines with (n = 15 samples) or without Ssk-FB4s (n = 23 samples). The distribution is shown as violin plots in Fig. 2f. Because the expression profiles of Ssk and Ssk-FB4s are largely similar, we simply merged the expression of Ssk and Ssk-FB4s. The one-sided Wilcoxon rank-sum test was used to calculate the $P$-value. (f) Relative abundance of Ssk and Ssk-FB4 in RAL-399 cells before and after co-immunoprecipitation with an antibody against Mesh. Dots indicate two biologically independent replicates.

Fig. 5 Ssk-FB4s represent a young rapidly evolving protein family. a Unique peptides of Ssk and Ssk-FB4s. Three Ssk-FB4s are color-coded as in Fig. 3c. The top row shows the sequence of Ssk, whereas the distinct residues between Ssk and individual Ssk-FB4s are marked in gray. b Abundance (intensity-based absolute quantification, iBAQ) of Ssk and X-linked Ssk-FB4 in the RAL-399 line. Dots indicate 2 technically independent replicates. c Numbers of nonsynonymous and synonymous changes across three functional regions in Ssk-FB4s and Ssk. The upper polymorphism data are based on the codon-level alignment of Ssk-FB4s (Supplementary Fig. 14a). The lower divergence data of Ssk are based on the whole Drosophila genus (Supplementary Fig. 14e). Each cell is shown as the “number of changed sites (observed/expected proportion of changed sites)”. $P$-values were calculated for the regions of interest with the one-sided proportion test. d Allele frequency spectrum of FB4s and Ssk-FB4s in 83 DGRP lines. e Transcription of Ssk, mesh, and Tsp2A in the gut across lines with (n = 15 samples) or without Ssk-FB4s (n = 23 samples). The distribution is shown as violin plots in Fig. 2f. Because the expression profiles of Ssk and Ssk-FB4s are largely similar, we simply merged the expression of Ssk and Ssk-FB4s. The one-sided Wilcoxon rank-sum test was used to calculate the $P$-value. f Relative abundance of Ssk and Ssk-FB4 in RAL-399 cells before and after co-immunoprecipitation with an antibody against Mesh. Dots indicate two biologically independent replicates.

Ssk-like protein family evolving under positive selection. We finally examined whether Ssk-FB4s can encode proteins. We collected protein bands with molecular weights in the range of 15–20 kDa (the mass of Ssk or Ssk-FB4 is ~17 kDa) and used a mass spectrometer to search for peptides ("Methods"). We identified high-quality unique peptides encoded by three Ssk-FB4 copies in available fly lines (Fig. 5a and Supplementary Fig. 12, and Supplementary Table 2). We quantified the expression53 of X-linked Ssk-FB4, which exhibited the highest frequency, and found that its protein level largely mirrors its transcriptional pattern (Fig. 4g and 5b): (1) Ssk-FB4 is upregulated in the midgut relative to its expression in the head; (2) the expression intensity of Ssk-FB4 is 5.5-fold higher than that of Ssk in the RAL-399 line midgut; and (3) Ssk-FB4 exhibited higher intensity than Ssk in the head of both the RAL-399 and RAL-427 lines.

Because translation does not necessarily mean function affecting fitness34, we analyzed whether Ssk-FB4s were subject to positive selection by examining the substitution patterns and frequency distribution. First, because Ssk encodes a membrane protein38, substitutions at approximately neutral sites, i.e., synonymous sites, would even accumulate55,56, in intracellular, transmembrane, and extracellular regions (Supplementary Fig. 13a). Consistently, there is no enrichment of synonymous substitutions in three regions of Ssk or Ssk-FB4s (Fig. 5c). In contrast, substitutions at functional sites, i.e., nonsynonymous sites, were overrepresented in the transmembrane domains of Ssk and the extracellular regions of Ssk-FB4s, respectively (Fig. 5c and Supplementary Fig. 14b, c). Given the even distribution of synonymous substitutions, these overrepresentations suggest weaker purifying selection or stronger positive selection in the corresponding functional regions. Second, polymorphic FB4s show low allele frequencies, with 65.8% as singletons (Fig. 5d, "Methods"). This phenomenon suggests that the spread of FB4s is repressed by negative selection especially considering the generally deleterious mutagenic nature of TEs57 (Supplementary Fig. 13b). In comparison, the four Ssk-FB4s have a higher frequency than FB4s (Fig. 5d). This contrast was confirmed by independent frequency data of FB4s (Supplementary Fig. 15a). The increase in the Ssk-FB4 frequency is either because these are less deleterious than FB4s or because they are subject to positive selection. We performed a test by searching for the signal of selective sweep58, where haplotypes harboring focal mutations would have no time to accumulate many mutations if positive
selection drives the rapid increase of these haplotypes (Supplementary Fig. 13c). For two high-frequency copies applicable for this analysis, we indeed found the depletion of linked single nucleotide polymorphisms (SNPs, Supplementary Fig. 15b, c) that could not be explained by chance (P < 0.05, Supplementary Fig. 15d, e).

Ssk mainly exerts its function by forming a protein complex with Mesh and Tsp2A at smooth septate junctions (sSJs)\(^ \text{52,59} \). Possibly due to the stoichiometric balance between protein complex members\(^ \text{60} \), these sSj proteins appear to be subject to coregulation and thus maintain a dosage balance. For example, in response to environmental changes, Ssk and mesh are consistently up- or downregulated\(^ \text{61} \). Therefore, we analyzed whether Ssk-FB4s are coregulated with mesh or Tsp2A. We first found that the expression of mesh or Tsp2A was not upregulated despite the extra dosage of Ssk-FB4s (Fig. 5e and Supplementary Fig. 16). Moreover, after co-immunoprecipitation with an antibody against Mesh, we quantified the protein intensity of Ssk-FB4 and Ssk. Despite the 500% abundance of X-linked Ssk-FB4 relative to that of Ssk in the fly midgut, the former abundance dropped to 70% of the latter in the immunoprecipitates (Fig. 5f). Thus, Ssk-FB4s appear not to be involved in the protein complex of Ssk/Mesh/Tsp2A or only weakly interact with this complex. Together with the contrastive substitution pattern (Fig. 5c), Ssk-FB4s and Ssk very likely have different functions.

Therefore, Ssk-FB4s might have rapidly acquired a novel functional role by modifying extracellular regions under the action of positive selection.

Discussion

Our study identified hundreds of Pack-TIRs in animals and provided insights into the mechanism underlying their emergence and how they drove functional evolution.

Specifically, our results, particularly the findings that the emergence of Pack-TIRs when DNA TEs were inactive in pri-mates (Fig. 1e and Supplementary Data 3) and that the distribution of breakpoints was biased to internal regions (Fig. 2q), indicate that the transposition-independent gap-filling model (Supplementary Fig. 1b) underlies the formation of single-copy Pack-TIRs. Certainly, the transposition-dependent gap-filling model (Supplementary Fig. 1c) might also generate a small proportion of single-copy Pack-TIRs. Their copy numbers did not increase, possibly due to low activity of the corresponding TEs or negative selection\(^ \text{62} \). Our analyses also hint toward a new FoSTeST mechanism generating multicopy Pack-TIRs. The working model of FoSTeST is as follows (Fig. 6): (1) replication fork stalling; (2) template switching; (3) transposition; and (4) potential amplification via a new round of transposition.

Whether the gap-filling or FoSTeST model also applies to the formation of Pack-MULEs in plants is worth exploring, particularly considering that the comparative framework used in this study is becoming applicable to rice given the recent accumulation of rice genomic data\(^ \text{65,66} \). Compared with animals, the FoSTeST model is likely more important in plants where TIR TEs are more active (e.g., MULEs in rice) and more inter-chromosomal and multicopy Pack-MULEs exist\(^ \text{26,67} \). Moreover, similar to Pack-MULEs\(^ \text{27} \), multicopy Pack-TIRs also tend to capture high GC-content regions, although single-copy Pack-TIRs do not show this pattern (Supplementary Fig. 17).

With respect to the evolution of new gene structures and new genes in animals, TIR TEs are underappreciated. It could be argued that the rate of TIR TE-mediated duplication was low (Fig. 1a). However, because we used multiple criteria to identify young cases associated with parental sources and flanking TIR TEs, our list of Pack-TIRs is highly conservative (“Methods”). Specifically, our survey identified only 12 Pack-TIRs in humans (Fig. 1a). However, manual curation showed that we missed 21 cases identified in other primates (Fig. 1e), mainly because orthologous cases in humans are slightly below the cutoffs (e.g., identity cutoff with parental copies). Moreover, when we adopted relaxed criteria by removing the requirement of the presence of a parental copy, we identified 199 (including 12 in the original scan) Pack-TIRs in humans. Third, we required the presence of TEs on both sides, whereas they rapidly degenerated via subsequent mutations. Taking one multicopy Pack-TIR in frogs as an example, three out of five copies already lost or gained a TE on one side despite their young ages, as suggested by the high identity detected (Supplementary Fig. 18).

Similar to other duplication mechanisms\(^ \text{9} \), TIR TE-mediated mechanisms, including both the gap-filling and FoSTeST models, affect the functional trajectories of duplicates. Such a predisposing effect could be demonstrated at both the mRNA and protein levels. Specifically, due to the fragmented nature of Pack-TIRs, they mainly fine-tune the structure of host genes at the insertion site. Pack-TIRs expand the 5′ modification exerted by Pack-MULEs\(^ \text{29} \) by adding intron retention and 3′ UTR elongation. Certainly, Pack-TIRs could be transcribed as antisense transcripts for the parental genes, which occur in two bidirectional cases derived from PTPRE and ARF6 in humans (Fig. 4c). Thus, regulation as small noncoding RNAs due to sense/antisense pairing, as observed with Pack-MULEs\(^ \text{30} \), could also occur for Pack-TIRs. Because KI-MULE appears to be unexpressed\(^ \text{31} \), Ssk-FB4s are the first identified potentially functional coding Pack-TIRs (Fig. 5), and their evolution appears facilitated by the FoSTeST model. With increases in the copy number, they likely evolve under the adaptive radiation model or innovation, amplification, and divergence (IAD) model\(^ \text{7-68} \). In other words, Ssk exhibits trace-level beneficial side activity (innovation) but cannot be optimized due to the antagonistic constraint of its main function. With
amplification to increase the mutational targets, positive selection specifically targets the side function of the derived copies and leads to divergence. Consistently, substitutions mainly occur in transmembrane (improving the main function) and extracellular (improving the side function) regions of Ssk and Ssk-FB4s, respectively; and Ssk-FB4s weakly interact with Msh or Tsp2A. Furthermore, Ssk does not accumulate any nonsynonymous SNPs or coding indels but three synonymous SNPs in the DGRP population. In contrast, Ssk-FB4s harbor an excess of amino acid changes (10 nonsynonymous SNPs and two in-frame coding indels but three synonymous SNPs, FET = 0.044; Supplementary Fig. 14a and 14d).

The side function of Ssk or the major function of Ssk-FB4s is intriguing. Although Ssk is known to be involved in the intestinal epithelial barrier, gut homeostasis and immune response, no studies have pinpointed the key residues underlying these processes; Therefore, we could not infer the function of Ssk-FB4s based on amino acid differences relative to Ssk. Nonetheless, Ssk-FB4s are likely involved in similar processes as Ssk given their similar expression pattern. To infer the functionality of Ssk-FB4s, we analyzed public genome-wide association study (GWAS) datasets (“Methods”) from flies subjected to microbial infection and their gut homeostasis or immune response was challenged. We found that the presence of Ssk-FB4s was always associated with stronger resistance against viral, bacterial and fungal infection in all datasets; the comparison reaches marginal significance in one set with female flies subjected to bacterial infection (P = 0.06, Supplementary Fig. 19), and significance in another two sets with flies subjected to fungal infection (P = 0.002 and P = 0.02, respectively). Whether and how they shape these phenotypes warrant further studies.

We noted that studies of new genes mainly focused on fixed genes. The youngest or most polymorphic new genes have rarely been studied, with the exception of a few tandemly duplicated or retroduplicated cases, such as amylase, Sdic or retro-FGFR77–79. In this respect, the Ssk-FB4 family represents the only reported case mediated by DNA transposons. Despite their young age, their possible function suggested by proteomic, evolutionary, and association analyses adds to the expanding picture in which new genes could rapidly gain functionality80,81.

Finally, our results together with those obtained in prior studies16,12,14,15 show that both DNA transposons and retrotransposons catalyze duplications in animals. Despite their different transposition mechanisms and distribution across species, DNA transposons and retrotransposons share four features. First, the end bypass model (Supplementary Fig. 1a) can apply to the 3′ transduction mediated by TIR TE20, Helitron82, L1s15,83, and SVAs14. Second, template switching during the gap-filling (Supplementary Fig. 1b, c) or FoSTeST model (Fig. 6) is also commonly observed in the duplication process mediated by Helitron84, L1s85 and LTR retrotransposons10. Third, duplicates flanked by TEs could act as pseudo TEs and amplify via further transpositions, such as the Ssk-FB4 family described here, the AMAC family mediated by SVA14,86, and the CG17604_r family mediated by LTR retrotransposon10. Fourth, chimerism between duplications, flanking TEs or insertion sites is widespread not only for Pack-TIRs but also for Helitrons16, LTRs10, L1s85,87, and SVAs14. All of these features confer TEs with a strong capability of shuffling genetic materials and endorse TEs as a vibrant force in shaping the evolution of new genes and new gene structures in animals.

Methods
Identification and analyses of Pack-TIRs in 100 animal reference genomes.

We identified young Pack-TIRs (Supplementary Fig. 20a) with unambiguous parental copies by modifying a previous method87. Specifically, the UCSC Genome Browser database hosts 106 animal species (https://genome.ucsc.edu, December 2016), in which TEs have been annotated in 100 genomes. After downloading the sequence alignments and TE annotation files for these species, we searched for these sequences flanked by two TIR TEs belonging to the same element where the internal sequences were longer than 100 bp and shorter than 5,000 bp by following routine practices87,88. Because Pack-MULEs are short (~300 bp), 5,000 bp is sufficient to contain bona fide Pack-TIRs. Taking humans as an example, the largest case is ARF3 (3,585 bp). We also tried an alternative cutoff of 10,000 bp in humans and could not find additional Pack-TIRs.

To ensure that one TIR TE element recently captured a sequence with an unambiguous source, we implemented the following procedures using custom scripts written on Perl vs 25.2.6 (MetaMeta Perl Language): Pack-TIRs were filtered to cover at least 10 bp of the terminals of the consensus TEs; (2) we searched the internal sequences against the genomes by running BLAT v3589 with the parameter “minScore = 100, minIdentity = 90.” (3) to ensure that the capture event was recent and did not include additional deletions or insertions, we required the ratio of the length between internal sequences and parental copies to be between 80 and 120%; (4) we required that the top BLAT hit was unique such that the possible second hit had a lower score; and (5) we manually checked the candidates on the UCSC Genome Browser to exclude cases containing sequencing gaps in the parental loci. To ensure bona fide DNA TE-mediated duplication, we then excluded cases with parental copies harboring TEs in the flanking 100-bp regions to control for segmental duplications. We also discarded cases with parental copies encoding multiexonic genes and captured sequences harboring intronless copies to control for retroduplications80.

Herein, we followed the routine practice for Pack-MULEs in plants60,67 and searched Pack-TIRs with parental copies not associated with TIR TEs to control for small-scale duplications. Using our pipeline, we may identify a directly generated by the end bypass model for which TEs are linked with parental copies.

Our initial search found the multicopy family in Fig. 1d. We further searched all 370 candidate Pack-TIRs by lowering the cutoffs, e.g., incompleteness of TIR repeated. We then found another three multicopy cases whose members were closely similar to the one initially detected with our pipeline (see one example in Supplementary Fig. 18).

To date Pack-TIRs along the primate phylogenetic tree, we followed our previous practice31,32 and deduced the orthology of Pack-TIRs based on the UCSC whole-genome syntenic alignments across primates. We inferred their evolutionary ages according to maximum parsimony.

We followed a previous method10 to estimate whether microhomology could be observed by chance for 33 Pack-TIRs in humans. At the 66 breakpoints of these Pack-TIRs, 32 left- or right-side breakpoints harbored microhomology, whereas the others were blunt or linked with short de novo insertions and were therefore excluded from our simulation. For 32 cases, we took the parental copy together with its 5′ and 3′ flanking regions (half length of the parental copy) as the template (Supplementary Fig. 3a). We randomly selected 100 fragments from the parental locus with the same length as the captured sequence of the corresponding Pack-TIR and randomly selected the substitution points in the consensus TIR. We recorded the size of the microhomology at the breakpoint of interest and counted the proportion of replicates with equal or longer microhomology as the empirical P-value. The microhomology length distribution of 100 replicates is depicted as boxplots in Supplementary Fig. 3b. Taking case #2 as an example, only six replicates showed equal or longer microhomology than the length of observed microhomology, and we thus labeled this example as “16.0”.

Regarding the positions of breakpoints, because they can be closer to either the 5′ or 3′ end, we normalized them to the 5′ side. For example, if the relative position (proportion) of the breakpoints are 0.8–0.9, they would be normalized as 1.0–1.0. To test against the null (excision) model via transposition, we conservatively assumed the breakpoint closer to the terminal as the DSB site and thus recorded a value of 0.1. We generated 281 (the actual total number of Pack-TIRs) random samples of breakpoints based on uniform distributions.

Identification and analyses of polymorphic Pack-TIRs in flies.

We searched for polymorphic Pack-TIRs (Supplementary Fig. 20b) based on our PacBio data across six DGRP lines (RAL-208, -379, -399, -427, -517, and -799). First, we mapped the long subreads to the reference genome (UCSC dm6) via BLASR v1.3.14224493. We detected structural variations via SMRT-SV v1.94 and then locally assembled them as contigs via Canu v1.895. By searching against TEs annotated in Repbase96, we then found another three multicopy cases whose members were closely similar to the one initially detected with our pipeline (see one example in Supplementary Fig. 18).

To date Pack-TIRs along the primate phylogenetic tree, we followed our previous practice31,32 and deduced the orthology of Pack-TIRs based on the UCSC whole-genome syntenic alignments across primates. We inferred their evolutionary ages according to maximum parsimony.

We followed a previous method10 to estimate whether microhomology could be observed by chance for 33 Pack-TIRs in humans. At the 66 breakpoints of these Pack-TIRs, 32 left- or right-side breakpoints harbored microhomology, whereas the others were blunt or linked with short de novo insertions and were therefore excluded from our simulation. For 32 cases, we took the parental copy together with its 5′ and 3′ flanking regions (half length of the parental copy) as the template (Supplementary Fig. 3a). We randomly selected 100 fragments from the parental locus with the same length as the captured sequence of the corresponding Pack-TIR and randomly selected the substitution points in the consensus TIR. We recorded the size of the microhomology at the breakpoint of interest and counted the proportion of replicates with equal or longer microhomology as the empirical P-value. The microhomology length distribution of 100 replicates is depicted as boxplots in Supplementary Fig. 3b. Taking case #2 as an example, only six replicates showed equal or longer microhomology than the length of observed microhomology, and we thus labeled this example as “16.0”.

Regarding the positions of breakpoints, because they can be closer to either the 5′ or 3′ end, we normalized them to the 5′ side. For example, if the relative position (proportion) of the breakpoints are 0.8–0.9, they would be normalized as 1.0–1.0. To test against the null (excision) model via transposition, we conservatively assumed the breakpoint closer to the terminal as the DSB site and thus recorded a value of 0.1. We generated 281 (the actual total number of Pack-TIRs) random samples of breakpoints based on uniform distributions.
supporting the presence of Ssk-FB4 in a line of interest. Through this process, we found a third insertion site at chr3R:14,340,248. We also noticed that Ssk-FB4 was heterozygous at this site, where we found a chimeric structure of the insertion site as well as discordant or split-mapped reads (Supplementary Data 5). Consistently, heterozygous sites are always situated in known inversions of DGRP lines\(^{48}\). Analogously, we performed genotyping in 90 GDL lines\(^{29}\) and did not find any new insertion sites (Supplementary Data 5).

To infer the evolutionary history of Ssk-FB4s, we first curated full-length sequences of Ssk-FB4s. Specifically, our lab maintained 20 DGRP lines with nine lines containing Ssk-FB4 copies, and four of these lines were sequenced using PacBio technology. For the remaining five lines, we performed PCR for Ssk-FB4s using primers described in Supplementary Data 7 and then performed Sanger sequencing. In total, we generated 12 full-length sequences of Ssk-FB4s (GenBank accession numbers MT433937-MT433948) in these nine lines, which cover all four Ssk-FB4 copies (Fig. 3c). We then aligned the internal sequences of 12 Ssk-FB4s and two orthologous Ssk in the reference genomes of D. melanogaster and D. simulans (UCSC drosom1) via MUSCLE v3.8.31\(^{100}\) and manually polished the multiple sequence alignment using MEGA v7.2.160\(^{100}\). We used RAxML v8.2.120\(^{101}\) with the frequently used GTR+GAMMA nucleotide substitution model\(^{102}\) to construct a phylogenetic tree. To determine which of the six full-length Ssk-FB4s mediated the formation of Ssk-FB4, we performed similar phylogenetic analyses with the TIR TE phyly as described in a previous study\(^{38}\), and analyzed the enrichment of synonymous and nonsynonymous changes within each category. Second, if the spread of Ssk-FB4s was under positive selection, their frequencies were expected to be higher than that of in-gel digestion. We counted the number of polymorphic substitutions along each lineage of Ssk-FB4s (Supplementary Fig. 14d). We further divided the protein encoded by each Ssk-FB4 into three functional categories (transmembrane, intracellular and extra-cellular, Supplementary Fig. 14a) according to a previous study\(^{38}\), and analyzed the enrichment of synonymous and non-synonymous changes within each category.

Searching signal of positive selection. We deployed three strategies to detect positive selection acting on Ssk-FB4s. First, we calculated the KH ratio (the ratio between the nonsynonymous and synonymous substitution rates) for Ssk in 12 Drosophila (Supplementary Fig. 14e) using the codeml tool, which is part of the PAML v4.9h package\(^{48}\). Moreover, with Ssk as the outgroup, we used codeml to count the number of polymorphic substitutions along each lineage of Ssk-FB4s (Supplementary Fig. 14d). We further divided the protein encoded by Ssk or Ssk-FB4 into three functional categories (transmembrane, intracellular and extra-cellular, Supplementary Fig. 14a) according to a previous study\(^{38}\), and analyzed the enrichment of synonymous and non-synonymous changes within each category. Second, if the spread of Ssk-FB4s was under positive selection, their frequencies were expected to be higher than that of in-gel digestion. We counted the number of theoretically observable peptides 53. To test this hypothesis, we divided DGRP lines into one group harboring Ssk-FB4 and another group not harboring Ssk-FB4. By implementing vcftools v0.1.120\(^{100}\) with a 1,000-bp window size and a 500-bp step size, we took advantage of DGRP SNPs and calculated the nucleotide diversity (\(\pi\)) for 50-kb regions surrounding the insertion site in the two groups. We found that X-linked and 3R-linked (chr3R:17.7 Mb) Ssk-FB4s were associated with a lower \(\pi\) on the left side (10 kb) of the breakpoint and on the two sides (15 kb on one side, 30 kb in total) of the breakpoints, respectively. To see whether this trend was true among all 30,000 bp between the two sites, we divided \(\pi\) into several subsets based on the observed values. To estimate whether neutral evolution resulted in such a deviation, we used scripts (R v3.4.4) to randomly sample 1,000 synonymous sites with the same chromosome, a similar (95%–105% fold) allele frequency and a similar recombination rate (95%–105% fold) as the two Ssk-FB4 loci and examined whether a decrease in \(\pi\) across 10 or 30 kb could be observed. We count the number of replicates as empirical statistical power (more than 80 lines). These Ssk-FB4s were associated with a lower \(\pi\) than the observed values.

Transcription analyses. To determine the gene structure of Pack-TIRs in humans, we selected one public high-depth and strand-specific RNA-seq dataset\(^{106}\) and used the splice-aware mapper STAR v2.4.0k\(^{107}\). We retained only reads mapped to the human genome. We selected DEGs with a fold change of 2 between the expressed Pack-TIRs and 100-kb flanking regions. We performed de novo assembly using Trinity v2.6.5\(^{111}\) with the parameter -a conservatively required at least RF. We calculated the gene model scores for which mutations could be passed to offspring. We could only retrieve one reads to the UCSC dm6 reference genome via HiCUP\(^{119}\); removed experimental datasets were used in Fig. 5d and Supplementary Fig. 15a, respectively. Note that the splice-aware mapper STAR v2.4.0k\(^{107}\) to map reads against the human genome. We used custom scripts (Python v2.7.15) to calculate the interaction intensity.

Protein identification, quantification, and co-immunoprecipitation (Co-IP). For proteomics, each sample including 10 midguts or 10 heads was lysed with the PCTM-Magic kit (Thermo, 30 ml of lysis buffer: 1 ml lysis buffer, 0.8 ml urea and 0.1 ml dithiothreitol) supplemented with complete mini EDTA-free protease inhibitor (Roche) in a barocycler (2320EXT, Pressure BioSciences, Inc.)\(^{112}\). The lysate program consisted of 90 cycles, and each cycle comprised 25 s of high pressure at 45,000 psi and 10 s of ambient pressure at 30 °C. The protein concentration was determined using the Pierce BCA protein assay kit. After incubation with SDS-PAGE loading buffer in a boiling water bath for 10 min, the protein was separated by SDS-PAGE.

The other proteomic procedures followed routine practice. Specifically, for in-gel digestion, the gels were visualized by Coomassie blue staining. Protein bands with molecular weights in the range of 15–20 kDa (the mass of Ssk or Ssk-FB4, ~17 kDa) were excised from the gel and digested with trypsin. The digested samples were lyophilized and further suspended in MUS buffer before subsequent liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis. The samples were spiked with 15% iRT peptides (Biontin) and analyzed using a Q Exactive HF X Orbitrap mass spectrometer coupled with an EASY-nLC 1200 system. The peptides were loaded into the trap. The peptides were then washed and further suspended in MUS buffer before subsequent liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis. The samples were spiked with 15% iRT peptides (Biontin) and analyzed using a Q Exactive HF X Orbitrap mass spectrometer coupled with an EASY-nLC 1200 system. The peptides were loaded into the trap. The peptides were then washed and further suspended in MUS buffer before subsequent liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis. The samples were spiked with 15% iRT peptides (Biontin) and analyzed using a Q Exactive HF X Orbitrap mass spectrometer coupled with an EASY-nLC 1200 system. The peptides were loaded into the trap. The peptides were then washed and further suspended in MUS buffer before subsequent liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis. The samples were spiked with 15% iRT peptides (Biontin) and analyzed using a Q Exactive HF X Orbitrap mass spectrometer coupled with an EASY-nLC 1200 system. The peptides were loaded into the trap. The peptides were then washed and further suspended in MUS buffer before subsequent liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis.

PMR-MS raw files were processed in Skyline to generate an extracted ion chromatogram (XIC) and perform peak integration. The top five most intense fragment ions were used to quantify the protein peak area. The sum of all identified peptide intensities was used as the protein intensity. To perform intensity-based absolute quantification (iBAQ), protein intensities were further divided by the number of theoretically observable peptides\(^{151}\). The quantifications were performed using the Pierce ClickTag Matrix Magnetic IP/Co-ip kit. The middy were lysed in IP lysis buffer, and the debris was removed by centrifugation at 12,000 ×g and 4 °C for 20 min. The antibody against Mesh (diluted at 1:1,000) and A/G beads were added to the protein lysates. Afterward, the lysates were incubated overnight at 4 °C. The immune complex samples were washed five times with precold IP lysis buffer. All the samples were then boiled in SDS-PAGE loading buffer and analyzed by SDS-PAGE and in-gel digestion. Analogue PMR-MS runs were then performed as mentioned above.

GWAS data analyses. Since Ssk-FB4s could be involved in gut homeostasis or immune response, we performed a literature survey about public GWAS studies focusing on these traits in DGRP lines and found 13 studies (Supplementary Data 9). Considering the moderate population frequency of Ssk-FB4s, we focused on the association studies with population allele frequency (MAF) lower than 0.01. These datasets included the viability of different DGRP lines upon the infection of the West Nile virus subtype Kunjin (WNV-Kun)\(^{114}\), the bacterium Pseudomonas aeruginosa. We used multiple datasets to quantify the expression levels of Ssk, Ssk-FB4s, mesh, and Psp2A. First, for both the RNA-seq data of the guts from 38 DGRP lines and 31 D. melanogaster strains and the raw RNA-seq data from 200 Drosophila species that were processed using the same pipeline, i.e., FPKM (fragments per kilobase of exon model per million reads mapped). Because these two datasets did not differentiate Ssk and Ssk-FB4s, we interpreted the expression value of Ssk as the total expression of Ssk and Ssk-FB4s. Second, based on our own in-house RNA-seq data across five wild-type Drosophila species, we calculated the expression (TPM in each line using Kallisto v0.41.0\(^{108}\)) with the addition of sequences of Ssk-FB4s and FB4s in the transcriptome, Kallisto differentiated these paralogs via a mapping-free strategy. In parallel, we performed STAR-based alignment and loaded the mapping files into Integrative Genomics Viewer (IGV)\(^{117}\) to visualize the gene structure of Ssk-FB4s.
entomophilae” or Pseudomonas aeruginosa Pa14 and the fungus Metarhizium anisopliae Ma497. We divided the DGRP lines into two groups based on the presence or absence of Sok-FR4 and compared the viability between these two groups.

Data availability
All Sanger sequencing data generated in this study are available at the NCBI GenBank database under the accession codes MT433937-MT433961. The genome sequences and TE annotations for the 100 species are available at the UCSC Genome Browser database (https://genome.ucsc.edu).

Code availability
The code produced for this study is available at https://github.com/clamp131/pack-TIR and archived in Zenodo.

Received: 12 September 2020; Accepted: 23 June 2021; Published online: 13 July 2021

References
1. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982 (2007).
2. Makalowski, W. Not junk after all. Science 290, 1246–1247 (2003).
3. Gotea, V. & Makalowski, W. Do transposable elements really contribute to proteomes? Trends Genet. 22, 260–267 (2006).
4. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements from conflicts to benefits. Nat. Rev. Genet. 18, 71–86 (2017).
5. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41, 331–368 (2007).
6. Lisch, D. How important are transposons for plant evolution? Nat. Rev. Genet. 14, 49–61 (2013).
7. Inzé, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11, 97–108 (2010).
8. Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. Trends Genet. 20, 544–549 (2004).
9. Katju, V. In with the old, in with the new: the promiscuity of the duplication mechanism. Int. J. Evol. Biol. 2012, 341932–341932 (2012).
10. Tan, S. et al. LTR-mediated retrotransposition as a mechanism of RNA-based duplication in metazoans. Genome Res. 26, 1663–1675 (2016).
11. Zhao, D., Ferguson, A. A. & Jiang, N. What makes up plant genomes: the vanishing line between transposable elements and genes. Biochim. Biophys. Acta. 1859, 366–380 (2016).
12. Knäussmann, H., Visekrunb, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. 10, 19–31 (2009).
13. Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J. & van der Knaap, E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319, 1527–1530 (2008).
14. Xing, J. et al. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc. Natl Acad. Sci. USA 103, 17608–17613 (2006).
15. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent and archived in Zenodo.
16. Thomas, J., Phillips, C. D., Baker, R. J. & Pritham, E. J. Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. Genome Biol. Evol. 6, 2595–2610 (2014).
17. Morgante, M. et al. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. 37, 997–1002 (2005).
18. Cerbin, S. & Jiang, N. Duplication of host genes by transposable elements. Curr. Opin. Genet. Dev. 49, 63–69 (2018).
19. Catoni, M., Jonesman, T., Cerruti, E. & Paszkowski, J. Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. Nucleic Acids Res. 47, 1311–1320 (2018).
20. Tsutbota, S. I. & Huang, D. V. Capture of flanking DNA by a P element in Drosophila melanogaster: creation of a transposable element. Proc. Natl Acad. Sci. USA 88, 693–697 (1991).
21. Gloor, G., Nassif, N., Johnson-Schiltz, D., Preston, C. & Engels, W. Targeted gene replacement in Drosophila via P element-induced gap repair. Science 253, 1110–1111 (1991).
22. Fiston-Lavier, A.-S., Anxolabehere, D. & Quesneville, H. A model of segmental duplication formation in Drosophila melanogaster. Genome Res. 17, 1458–1470 (2007).
113. Roach, N. P. et al. The full-length transcriptome of *C. elegans* using direct RNA sequencing. *Genome Res.* **30**, 299–312 (2020).

114. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

115. Frochaux, M. V. et al. cis-regulatory variation modulates susceptibility to enteric infection in the *Drosophila* genetic reference panel. * Genome Biol.* **21**, 6 (2020).

116. Everett, L. J. et al. Gene expression networks in the *Drosophila* genetic reference panel. *Genome Res.* **30**, 485–496 (2020).

117. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

118. Stadler, M. R., Haines, J. E. & Eisen, M. B. Convergence of topological domain boundaries, insulators, and polytene interbands revealed by high-resolution mapping of chromatin contacts in the early *Drosophila melanogaster* embryo. *eLife* **6**, e29550 (2017).

119. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).

120. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).

121. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743–755 (2016).

122. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker’s guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).

123. Guo, T. et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.* **21**, 407–413 (2015).

124. Tan, S., Ma, H. & Zhang, Y. E. DNA transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans. *Zenodo* https://doi.org/10.5281/zenodo.4905489 (2021).

**Acknowledgements**

We thank Drs. Manyuan Long, Jinfeng Chen, Zongzhao Zhai, Qing Li, Yan Li, Tian Tang, Michael Stadler, and the Zhang laboratory members for the helpful discussions. We appreciate the Mesh antibody provided by Drs. Mikio Furuse and Yasushi Izumi. This research was supported by grants from the National Key R&D Program of China (2018YFC1406902, 2019YFA0802600), the Chinese Academy of Sciences (ZDBS-LY-SM005, No. XBZG-ZDSYS-201913), the National Natural Science Foundation of China (2018YFC1406902, 2019YFA0802600), the Chinese Academy of Sciences (ZDBS-LY-SM005, No. XBZG-ZDSYS-201913), the National Natural Science Foundation of China (2018YFC1406902, 2019YFA0802600), and the third round of public welfare development and reform pilot projects of Beijing Municipal Medical Research Institutes (Beijing Medical Research Institute, 2019-1). Additionally, the C.L.C. lab is supported by grants from the I. Curie YPI program, the ATIP-Avenir program from CNRS and Plan Cancer, the Agence Nationale pour la Recherche (ANR) and the Institut National du Cancer (INCa). G.L.B. is supported in part by the Office of the Director, National Institutes of Health, under Award Number P51OD011106. The computing was jointly supported by the HPC Platform of BIG and that of the Scientific Information Centre of IOZ.

**Author contributions**

Y.E.Z., S.T. and C.L.C. conceived and designed the study; S.T. and H.M. performed the computational analyses with help from D.W., Z.Z., and H.C.; J.W. and M.W. performed the experiments with help from M.X.W., H.Y., Y.Z., J.S., X.Z., S.G., J.M.W., and X.H.; and Y.E.Z., S.T., C.L.C., H.M., J.W., M.W., and G.L.B. wrote the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

*Supplementary information* The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-24585-9.

*Correspondence* and requests for materials should be addressed to C.-L.C. or Y.E.Z.

*Peer review information* *Nature Communications* thanks Wojciech Makalowski and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

*Reprints and permission information* is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2021