



Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in *Drosophila* and humans

Dan Zhang^{1,2,9}, Liang Leng^{3,9}, Chunyan Chen^{1,2,8}, Jiawei Huang^{1,2}, Yaqiong Zhang¹, Hao Yuan^{1,2}, Chenyu Ma^{1,2}, Hua Chen^{2,4,5,6} and Yong E. Zhang^{1,2,5,7}✉

Despite polymorphic duplicate genes' importance for the early stages of duplicate gene evolution, they are less studied than old gene duplicates. Two essential questions thus remain poorly addressed: how does dosage sensitivity, imposed by stoichiometry in protein complexes or by X chromosome dosage compensation, affect the emergence of complete duplicate genes? Do introns facilitate intergenic and intragenic chimaerism as predicted by the theory of exon shuffling? Here, we analysed new data for *Drosophila* and public data for humans, to characterize polymorphic duplicate genes with respect to dosage, exon-intron structures and allele frequencies. We found that complete duplicate genes are under dosage constraint induced by protein stoichiometry but potentially tolerated by X chromosome dosage compensation. We also found that in the intron-rich human genome, gene fusions and intragenic duplications extensively use intronic breakpoints generating in-frame proteins, in accordance with the theory of exon shuffling. Finally, we found that only a small proportion of complete or partial duplicates are at high frequencies, indicating the deleterious nature of dosage or gene structural changes. Altogether, we demonstrate how mechanistic factors including dosage sensitivity and exon-intron structure shape the short-term functional consequences of gene duplication.

Gene duplication has been studied for decades^{1–5} and numerous models have been formulated to explain the evolution of duplicate genes^{2,6,7}. Some models such as the subfunctionalization or neofunctionalization models generally assume the initial redundancy of the polymorphic copies^{1,7–10}, while other models assume that the copies are not redundant because of dosage effects or incomplete duplications^{7,11}.

Selection for higher dosage occurs when increased abundance is favoured^{18,12}. However, expression increases can disrupt dosage-sensitive genes: (1) the stoichiometric balance of members in protein complexes (assemblies of interacting proteins¹³) could be lost and rebalancing via downregulation could be required, as previously described^{14–17}; and (2) in the X chromosome, dosage compensation equalizes the gene products between males and females^{18,19} and new X-linked duplicates may not be as well compensated²⁰.

Non-redundancy can also be a consequence of incomplete duplications: 5' or 3' terminal duplicates can form chimaeras with flanking genes^{21–23}, whereas internal duplicates may harbour new exons, introns or alternative isoforms^{24–26}. The recombination of pre-existing genic fragments is often named exon shuffling. Both exon shuffling and alternative splicing were originally proposed by Gilbert on the basis of gene structure ('gene in pieces') and the possibility that exons act as functional units (domains)^{27,28}. Decades of research identified pervasive alternative splicing^{29,30} and provided evidence for exon shuffling, including the excess of symmetrical intron phase (maintaining the coding frame) and the correspondence between borders of exons and domains^{31–34}. The significance

of introns for exon shuffling is further supported by genomes with more introns (for example, in humans) often harbouring more intragenic duplications^{25,26} or showing stronger exon–domain correspondences^{31,33}.

Most previous work has focused on recently fixed or older duplicate genes. The polymorphic stage is rarely studied due to the difficulty in analysing similar polymorphic duplicates⁸. With improving sequencing techniques, polymorphic duplicates are extensively studied as copy number variants (CNVs)^{20,35–40}. However, although tandem duplicates constitute most polymorphic duplicates^{20,41}, they were explicitly studied in only a few genome-wide studies (for example, refs. 42–44). These efforts found evidence for positive selection targeting high-frequency complete duplicates⁴² and chimaeric structures generated by incomplete duplications⁴³ in *Drosophila*, and for fitness costs associated with dosage increase in *Caenorhabditis elegans*⁴⁴.

Because of the focus on old duplicate genes, two questions remain open: (1) how does dosage sensitivity affect the occurrence of duplicates? (2) how do different genomic features (intron-rich versus intron-poor) shape the landscape of intergenic and intragenic exon shuffling? The second question especially benefits from being studied in the context of polymorphic duplicates because of the rapid evolution of introns⁴⁵ in old duplicates and thus the difficulty in evaluating the role of introns in mediating the initial shuffling. To answer both questions, we identified polymorphic duplicates and analysed transcriptome data on the basis of new and public data in two species with different genomic architectures: humans (intron-rich and gene-sparse) and flies (intron-poor and

¹Key Laboratory of Zoological Systematics and Evolution & State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. ²University of Chinese Academy of Sciences, Beijing, China. ³Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China. ⁴CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. ⁵CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. ⁶China National Center for Bioinformation, Beijing, China. ⁷Chinese Institute for Brain Research, Beijing, China. ⁸Present address: School of Ecology and Environment, Northwestern Polytechnical University, Xi'an, China. ⁹These authors contributed equally: Dan Zhang, Liang Leng. ✉e-mail: zhangyong@ioz.ac.cn

gene-dense)^{46–48}. We analysed complete, 5', 3' and internal partial duplications with respect to dosage, gene structures and allele frequencies. We found that complete duplications are a minority of gene duplications and the only category to show strong expression increases and be subject to dosage constraints induced by stoichiometric balance while being potentially tolerable by X dosage compensation. We found that partial duplications, especially 5' duplications (possibly with promoters duplicated), are frequently transcribed as chimaeric genes. In the case of gene fusions and intragenic duplications, breakpoints are often exonic in flies, while almost entirely intronic in humans. However, for both species, the preduplication transcripts tend to remain as the major alternative isoforms maintaining the ancestral functionality. Finally, most duplications show low allele frequencies possibly because of their deleterious effects (for example, dosage increase). Taken together, our work identifies the short-term landscape of duplications as being shaped by dosage constraints and exon–intron architectures.

Results

High-quality datasets of polymorphic gene duplications in *Drosophila* and humans. For flies and humans, we compiled datasets of polymorphic tandem duplicate genes (Fig. 1a; Methods). We mainly relied on short-read data due to their accuracy and low cost^{49,50}.

In flies, we performed resequencing to increase the accuracy of the detection of polymorphic duplications and the determination of their breakpoints. Specifically, we sequenced genomes from one population, that is, *Drosophila melanogaster* Genetic Reference Panel (DGRP)⁵¹, for which CNVs have been identified^{52,53} and for which essential parameters (for example, heterozygosity) have been estimated⁵³. Among the 40 core lines, we selected six relatively unrelated and homozygous lines (Extended Data Fig. 1). We generated high-quality resequencing data (Supplementary Table 1): the depth reaches 160–220× to increase the accuracy of the read-depth-based methods and the length of paired-end reads is 250 base pairs (bp) with insert sizes >560 bp to boost the resolution of the split-read or discordant read-pair methods for inferring the breakpoints.

By extending previous work^{42,52,53} to balance specificity and sensitivity, we performed an integrated pipeline to detect duplications (Methods). Specifically, we implemented five methods developed by us and others^{54–58} that represent various strategies (for example, read depth). We combined these five calls with two published datasets^{52,53} to obtain seven datasets (Fig. 1a). Finally, we extracted 179 duplication blocks involving 270 genes (one duplication possibly involving more than one gene; Supplementary Table 2) with the following criteria: (1) duplications overlapped exons or introns of protein-coding genes; (2) the duplications were supported by at least three datasets; (3) the duplications were inferred to be derived on the basis of outgroups and thus not deletions in the reference genome (Supplementary Table 3); and (4) the duplications ranged between 50 bp and 25 kilobases (kb) because larger duplications could represent false positives⁴².

To evaluate changes in gene dosage or gene structures, we generated 84 RNA-seq datasets. Considering transcriptome diversity, we targeted five adult tissues and male and female larval fat bodies (Fig. 1a; Methods). We generated two biological replicates for each tissue across all lines. This dataset is of high quality, as revealed by the mapping rate (median percentage, 99.3%) and by the correlation between replicates (median Spearman ρ , 95.5%; Supplementary Table 4). Notably, 108 out of 270 genes show tissue-specific expression (Extended Data Fig. 2; Methods), supporting the importance of profiling diverse tissues.

Two validations show the high quality of our fly duplicate gene dataset. First, PCR validations indicate that the false positive rate (FPR) is 1.6% and the false negative rate 0.5% (Supplementary Table 5;

Methods). Second, de novo assembled transcripts around incomplete duplications show that the breakpoints of the duplications and those of the transcripts only differ by a median of 2 bp (Fig. 1b and Supplementary Table 6).

For humans, we took advantage of extensively validated (FPR < 10%) duplication calls and transcriptome data generated by the Genotype-Tissue Expression⁵⁹ (GTEx) project^{36,59}. GTEx covers 145 individuals and a median of nine tissues or cell lines (Fig. 1a). GTEx calls consist of one dataset (964 duplicate genes) based on the read-depth and split-read strategies and one dataset (188 duplications) based on read depth. We focused on the former dataset due to its larger sample size and used the latter when necessary (Methods). Similarly to flies, we found that most GTEx calls represented derived duplications (76.9–100%; Supplementary Table 3) and that most assembled transcripts were compatible with the breakpoints of the duplications (96.2%; Fig. 1b and Supplementary Table 6).

Collectively, we generated high-quality datasets of polymorphic duplicate genes in flies and humans in terms of FPR and breakpoint resolution. To make these datasets accessible, we uploaded them to the UCSC genome browser (see Data and code availability for this article).

Most duplications are partial in flies and intronic in humans.

We classified 270 duplications in *Drosophila* and 964 duplications in humans into complete duplications, partial duplications overlapping exonic regions (that is, partial duplications) and intronic duplications, and examined their sizes and abundance (Fig. 2a; Methods). As expected, complete duplications tend to be larger followed by partial and intronic duplications (Extended Data Fig. 3a). In both species, complete duplications accounted for a small fraction of all duplications (15.7–17.8%, left in Fig. 2b). *Drosophila* harbours a higher proportion of partial duplications and a lower proportion of intronic duplications than humans (58.9% versus 25.2%; 23.3% versus 59.1%). The difference is probably due to two factors: (1) introns and exons contribute, respectively, to 17% and 20% of the fly genome but to 40% and 1% of the human genome^{46,47}; (2) duplication sizes could be smaller than gene sizes, especially for humans (Extended Data Fig. 3b). To test whether these two factors explain the between-species differences, we generated random control datasets by shuffling coordinates of polymorphic gene duplications (keeping the size unchanged; Methods). The observed proportions fit the distribution of random datasets where partial duplications are predominant in flies and intronic duplications in humans (Extended Data Fig. 4a). Thus, the differences in genomic architecture and distribution of duplication sizes explain the higher proportion of partial duplications in flies and of intronic duplications in humans.

The distribution of the three types of duplications is further reproduced at the level of individual lines or individual humans (right panel of Fig. 2b) and is also recapitulated in the smaller human GTEx dataset, even though it contains relatively more complete duplications (Supplementary Table 7). With the two GTEx datasets merged, complete duplications remained less abundant than partial duplications (266 versus 295). Notably, individual fly line harbours similar numbers of duplications relative to individual human (median 58 versus 78). This could be explained by multiple factors such as different gene numbers or different polymorphism levels (Supplementary Table 8) caused by distinct effective population sizes^{60,61}.

Complete duplications are under dosage constraint induced by stoichiometric balances. Next, we examined whether duplicated genes show dosage increases by calculating the median fold-change (between duplication-present and duplication-absent lines) across tissues and lines (Fig. 3a). We found that complete duplications tend to show dosage increase (83% increase in flies and 52% in humans, one-sided Wilcoxon signed rank test $P < 1 \times 10^{-7}$). The median extent (83%) of upregulation in flies was significantly higher than

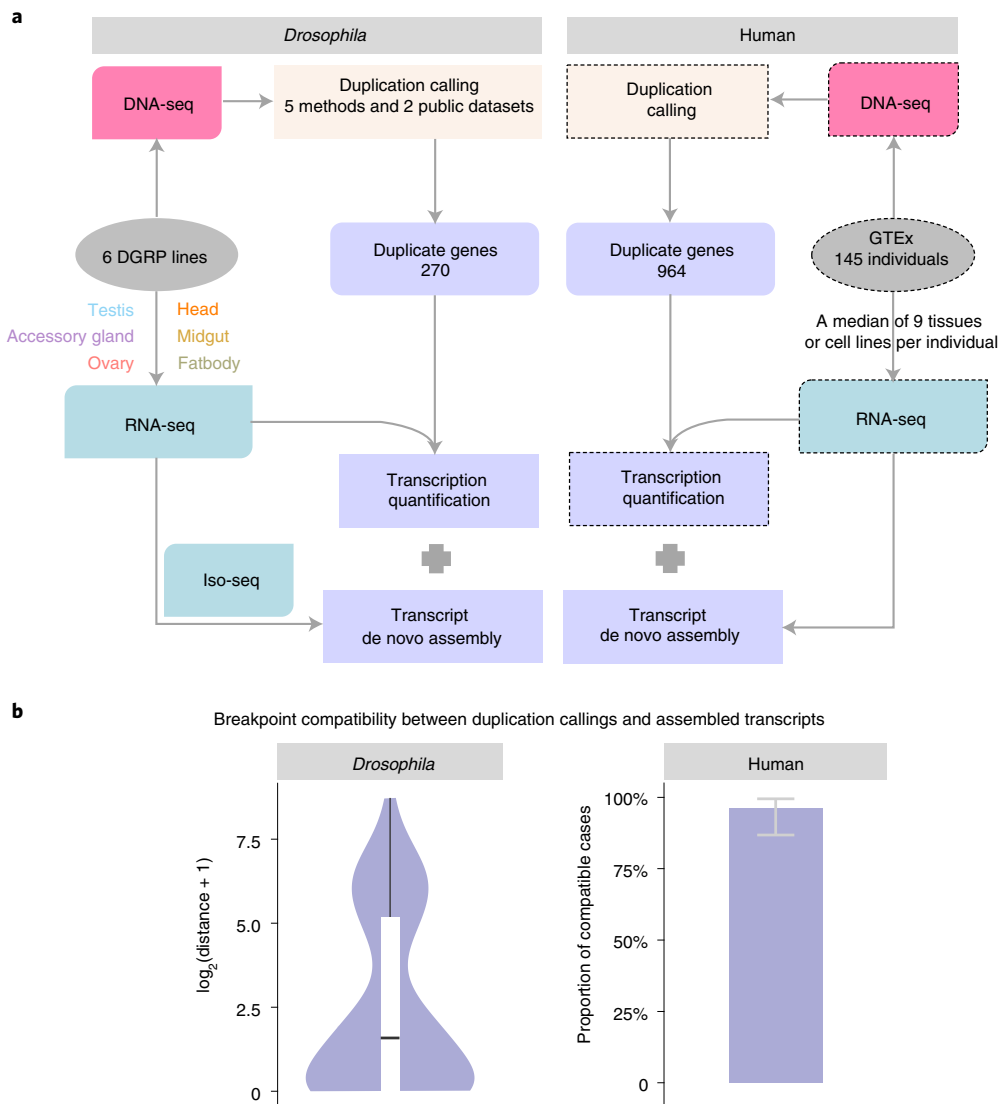


Fig. 1 | Project design and evaluation of breakpoint inference. a, Project design in flies and humans. In *Drosophila*, six DGRP lines were sequenced and seven sets of calls were merged as the final dataset (270 duplicate genes). The expression was quantified by Illumina short-read RNA-seq followed by whole-genome de novo transcript assembly. To test the quality of the assemblies, the testes of one line (RAL-379) were subjected to PacBio long-read isoform sequencing (Iso-seq). In humans, we used a dataset of 964 duplicate genes and RNA-seq data from the GTEx⁵⁹ project, which profiled 145 individuals. The expression levels were downloaded from GTEx and targeted local assemblies around duplication loci were performed subsequently. Previously published data or analyses associated with GTEx are framed by dashed lines. **b**, Breakpoint inference evaluation. Because many breakpoints in *Drosophila* reside in exons and the assembled transcripts (contigs) generally follow the pre-existing exon-intron structure, the contigs usually span these positions. So we recorded the distance between breakpoints of the duplicated loci and those harboured by the contigs. The distribution is shown as a violin plot where the curves indicate the probability density, the bar the interquartile range and the line the median. In contrast, human breakpoints tend to be intronic so the contigs do not harbour these breakpoints. We therefore examined whether the structures of contigs are compatible with the breakpoints of duplication loci and showed the proportion of compatible cases as a bar plot (right). The 95% CI, calculated via binomial distribution, is shown as the error bar.

50% ($P=0.003$; Fig. 3a) and not significantly different from 100%. In contrast, the upregulation in humans was similar to 50% and significantly lower than 100% ($P=2 \times 10^{-7}$). This difference is explained by the homozygosity of the fly lines (Extended Data Fig. 1) and the heterozygosity of humans (Supplementary Table 2). Note that for both species, some complete duplications were associated with no expression changes, downregulation or more than 100% upregulation (Fig. 3a), which could be due to postduplication feedback^{42,44,62}. We also examined each tissue individually and reproduced the overall pattern (Extended Data Fig. 5): (1) duplicate genes are generally upregulated in all tissues; and (2) the dosage increase in most tissues is not significantly different from 100% in flies and

50% in humans, although the extent varies possibly because of tissue-specific feedback.

We expected complete copies to be expressed similarly to parental copies because they have similar regulatory controls. To test this, for each gene, we calculated the expression correlation between duplication-present and duplication-absent lines, since it is difficult to directly quantify the expression of individual copies because of their sequence similarity. Consistent with our expectation, the median Spearman correlation is 0.94 for flies and 0.81 for humans (Fig. 3b). The lower correlation coefficient in humans could be due to the heterogeneity of the samples (for example, variation in sex or age).

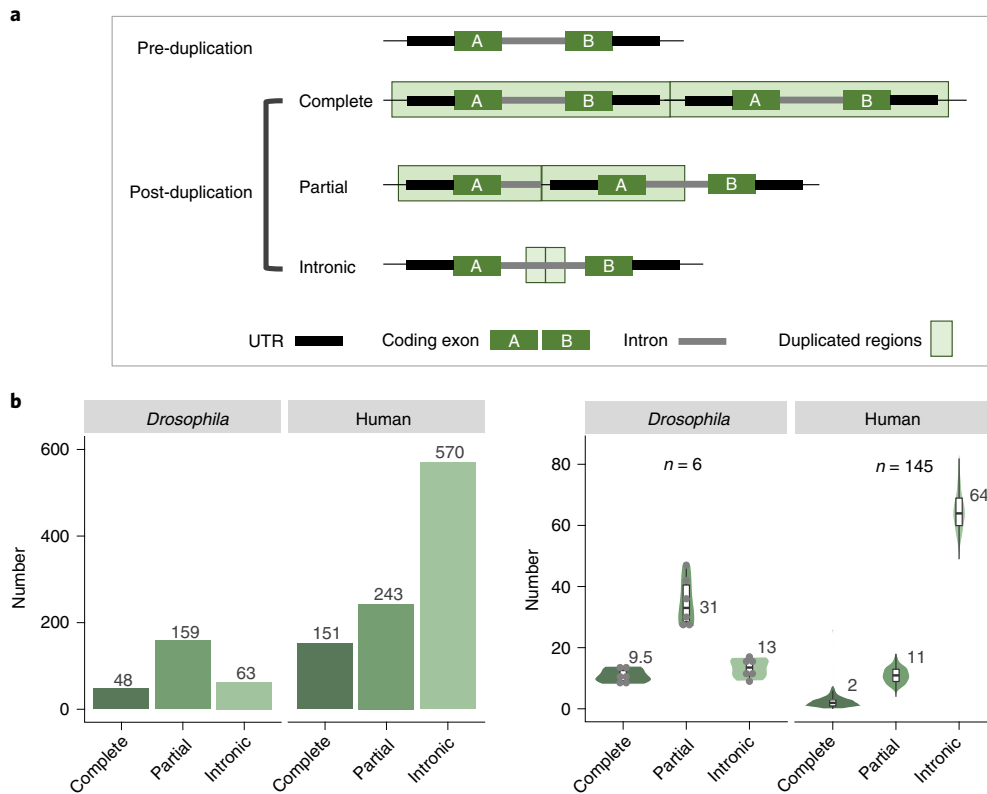


Fig. 2 | Type and distribution of duplicate genes. a, Schematic classification of duplicate genes. **b**, Abundance of complete, partial and intronic duplicates in the whole dataset (left) and in each line or individual (right). For *Drosophila*'s violin plots, all six data points are shown as grey dots. The actual numbers are labelled in the left graph while the median numbers are labelled in the right graph.

Compared with complete duplicates, partial duplicates only showed minor dosage increases (4–5%; Fig. 3a), while intronic duplications did not lead to notable expression changes in humans or flies (Fig. 3a). Given that complete duplicates show strong dosage increases, we predicted that members of protein complexes would only be depleted among complete duplicates due to the disruption of stoichiometric balances^{15–17}. Consistently, we found that members of protein complexes are depleted among complete duplicates when compared to the genomic background: 9.8% versus 18.3% in flies (one-sided proportion test $P=0.114$) and 8.9% versus 16.9% in humans ($P=0.017$; Fig. 3c). In contrast, the proportions of protein-complex members among partial and intronic duplicates were close to the expectation in both species.

We further examined a second type of dosage-sensitive genes, X-linked genes^{18–20}, and asked whether they are less likely to be completely duplicated compared to autosomal genes. We found that the number of complete duplicates on the X chromosome is proportional to the total number of X-linked genes in flies but potentially smaller in humans (Extended Data Fig. 6a; 16.7% versus 15.8% in flies; 2.0% versus 4.2% in humans, one-sided proportion test $P=0.125$). In flies, we took advantage of expression data for both males and females (for the fat body) and asked if X-linked complete duplicates were associated with dosage increases and if the expression levels were similar between the sexes. We found that complete duplicates led to dosage gains of X-linked genes just like other genes (Fig. 3a) and that X-linked genes generally show similar expression levels between males and females (Fig. 3d), with a between-sex expression ratio close to 1 (median ratio: 0.92 in duplication-absent lines, 1.09 in duplication-present lines). Since dosage compensation in *Drosophila* relies on the protein complex binding on high affinity sites (HASs)⁶³ and genes closer to HASs are better

compensated^{19,64,65}, we expected that genes with similar postduplication expression across sexes would be closer to HASs. Although there are only eight X-linked complete duplicates, random sampling detected a moderate signal ($P=0.126$; Extended Data Fig. 6b; Methods).

For humans, we only identified three X-linked complete duplicate genes: *PUDP*, *STS* and *VCX*. All three genes are situated in a young evolutionary strata⁶⁶. Young strata are enriched with genes escaping from X inactivation⁶⁷ and, consistently, *PUDP* and *STS* are known escapees^{68,69} with female-biased expression profiles. Considering that only 15% of X-linked genes are escapees⁶⁷, the fact that two out of three duplicated genes are escapees is significant (one-sided proportion test $P=0.045$). Moreover, *VCX* is a gene specifically expressed in testis⁷⁰ and thus all three duplicated genes do not need dosage compensation.

In summary, complete duplicates show dosage increases and are subject to dosage constraints imposed by stoichiometry in protein complexes while potentially being tolerable by the X dosage compensation system.

Incomplete duplications lead to chimaeric genes and exon shuffling. We then focused on partial duplications, which can induce exon shuffling. Because the postduplication changes in gene structure depend on the boundaries of duplications^{21,25,26}, we classified duplication loci involving partial duplications into six possible scenarios (Fig. 4a) and examined their distributions. We found that duplications potentially leading to gene fusions (5'–3', 5'–5' and 3'–3') are more frequent in flies than in humans (9.9–17.1% versus 6.1–10.7%; upper part of Fig. 4b), which is consistent with the higher gene density in flies^{46,47}. We note that only the proportion of the 3'–3' arrangement is statistically significant (one-sided Fisher's exact test, FET, $P=0.002$), which could be caused by the higher density of convergent gene pairs

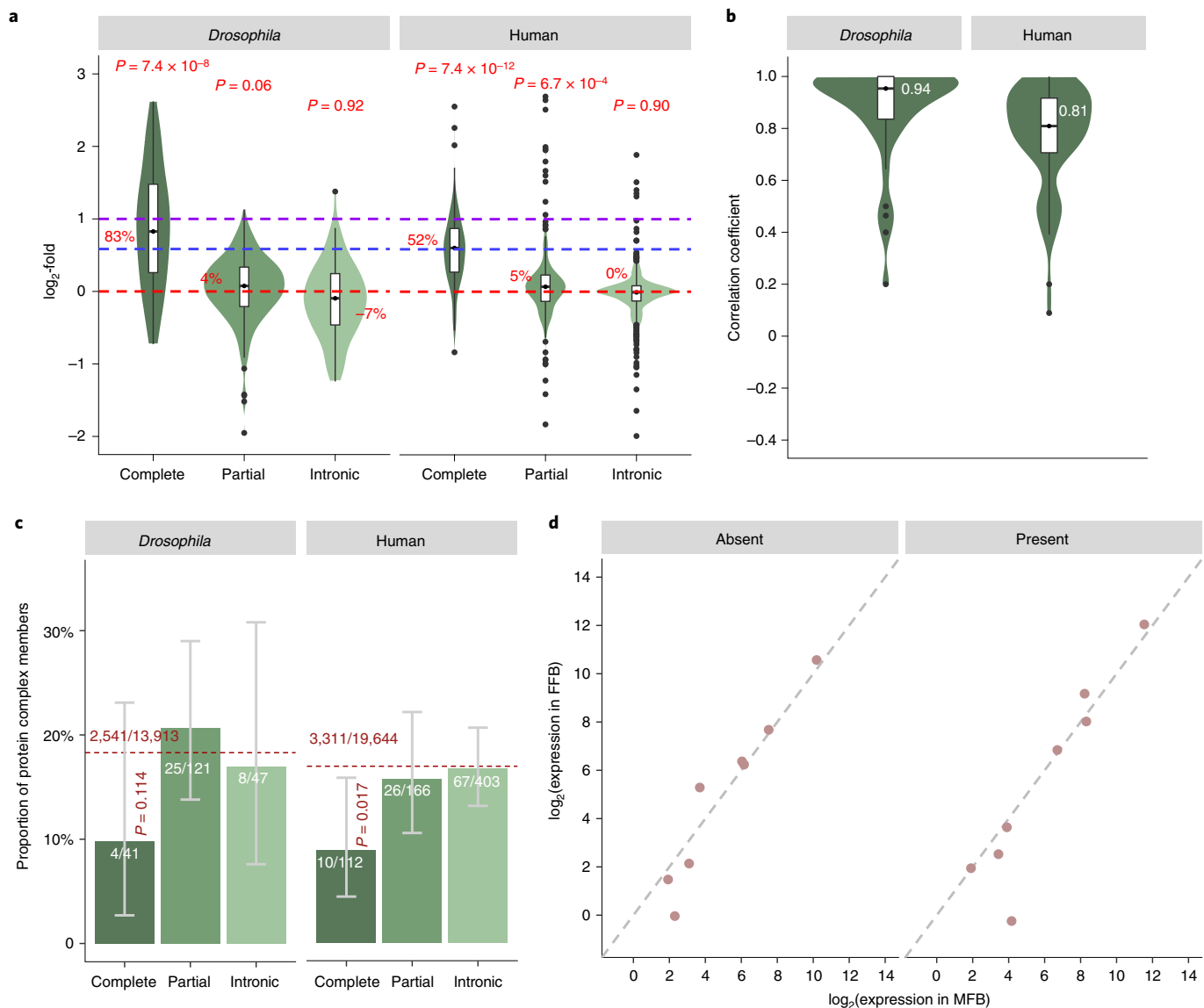


Fig. 3 | Expression changes and dosage constraints of duplicate genes. a, Transcriptional fold-change distribution for complete, partial and intronic duplicates. For each gene, we calculated the median fold-change across all tissues per line/individual and then recorded the median of the median ratio across lines/individuals as the summary statistics. The median raw percentage increases are labelled. The absence of a change and changes with 1.5- and 2-fold upregulation are marked by red, blue and purple lines, respectively. A one-sided Wilcoxon signed rank test was used to estimate the significance relative to the absence of a change in expression. To zoom onto the range with the high data density, outlier values beyond the range of the y axis (-2, 3) are omitted, which includes eight data points in flies (-5.4, 5.5) and five data points in humans (-2.5, 3.5). **b**, Distribution of expression correlation of complete duplicates. For each gene, we calculated the Spearman correlation between expression level of duplication-present and duplication-absent lines. The medians are labelled. **c**, Proportion of genes annotated as members of protein complexes among complete, partial and intronic duplications. The numbers are shown in each bar. The dashed line marks the genome-wide proportion of protein-complex members with the numbers shown above the line. A one-sided proportion test was used to estimate the significance relative to the genome-wide proportion. **d**, A scatter plot showing expression of eight X-linked complete duplicates in duplication-present and duplication-absent lines. The dashed diagonal line marks the expected identical expression between female fat bodies (FFB) and male fat bodies (MFB).

in flies⁷¹. In contrast, humans showed a marginal enrichment of internal duplications compared with flies (24.5% versus 16.2%, $P=0.059$), probably due to the greater gene size relative to duplication size found in humans (Extended Data Fig. 3b). Both the under-representation of gene fusions and the over-representation of internal duplications in humans were confirmed by the read-depth-based GTEx calls (Supplementary Table 7).

To test whether the genomic architecture and distribution of duplication sizes contributed to the between-species difference, we used the aforementioned controls and found that gene fusions are

more abundant in flies and internal duplications are more abundant in humans (Extended Data Fig. 4b). Although the difference in the observed data is recapitulated by the control data, the distribution of the six types of partial duplicates, especially that of internal duplications, differs between observation and control (16.2–24.5% versus 45.6–61.1%). This is probably a consequence of internal duplications often disrupting exon–intron structures and thus being removed by selection.

To examine whether incomplete duplications caused the structural changes predicted by the gene arrangements, we performed

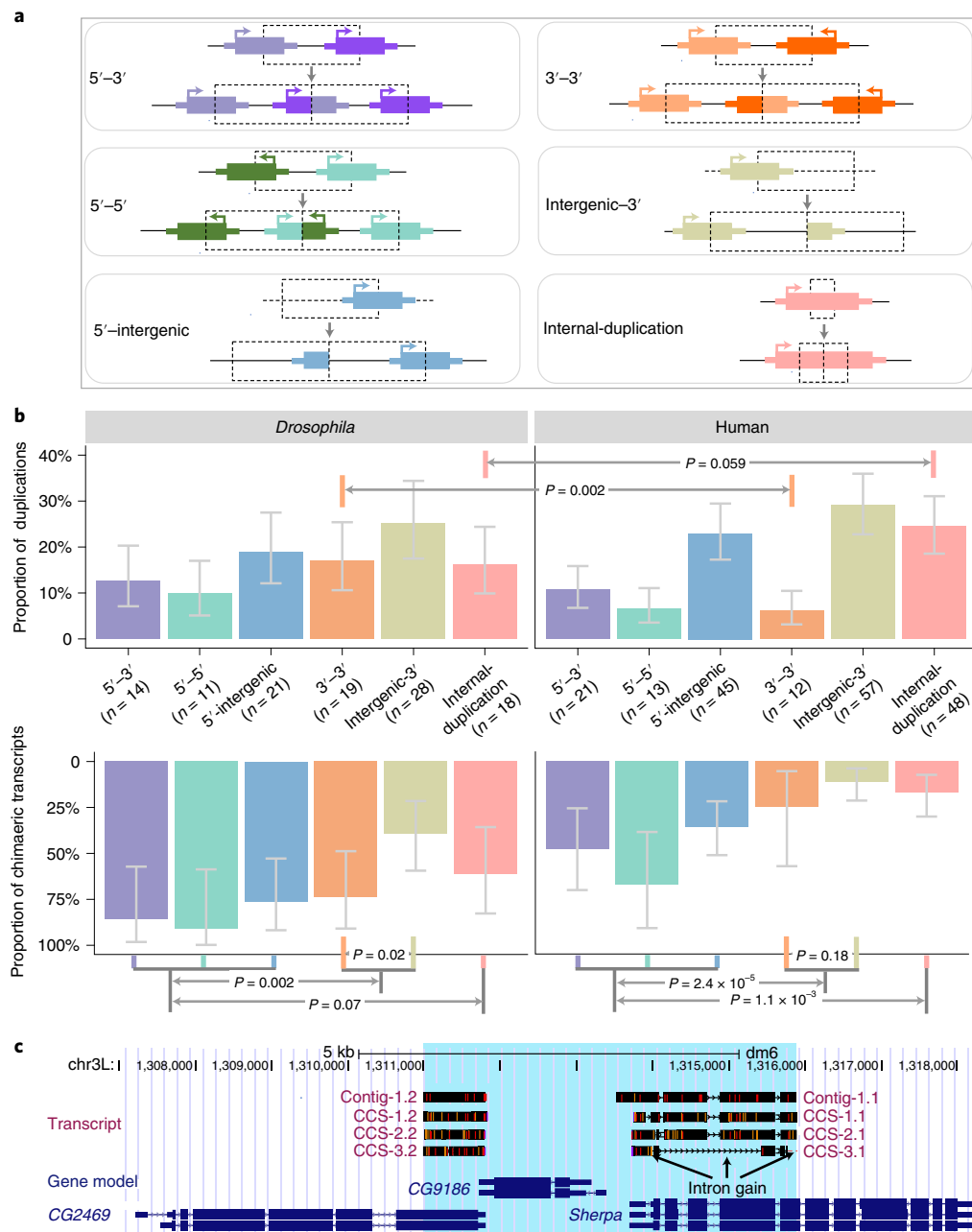


Fig. 4 | Pervasive chimaerism due to incomplete duplications. **a**, Schematic overview of six different chimaerism. For each type, the pre- and postduplication status is shown at the top and bottom, respectively. The gene models are drawn in a similar manner as in Fig. 2a except that arrows indicate the transcriptional orientations. **b**, Proportion of chimaerism predicted by breakpoint locations (top) and supported by de novo assemblies (bottom). Flies and humans show different biases with respect to 3'-3' and internal duplication arrangements (top). In the lower part, we compared the proportions of 5' duplicates relative to 3' and internal duplicates and the proportions of 3'-3' arrangements relative to intergenic-3' arrangements. One-sided Fisher's exact tests were performed. The error bar represents the 95% CI calculated via binomial distribution. **c**, Snapshot of the CG2469, CG9186 and *Sherpa* locus at the UCSC genome browser. The duplication block is highlighted in blue. One assembled contig and three PacBio CCS reads are presented as the transcript track, whereas FlyBase-annotated genes are presented as the gene model track. In the transcript track, mismatches are marked as red lines; CCS reads harbour more mismatches due to sequencing errors. The contig and three CCS reads were split into two parts according to the breakpoint and separately mapped with the 5' and 3' segment labelled as '1' and '2', respectively. CCS-3 harbours three new introns that are absent in the preduplication gene models. In the gene model track, the thinner and thicker boxes represent UTRs and coding exons, respectively, and the connecting lines represent introns with zigzags indicating transcriptional orientations. The transcript track is similarly shown, except that UTRs and coding exons are not differentiated.

de novo assembly and evaluated the quality of the assembled contigs. For 111 loci harbouring partial duplicates in flies, we assembled chimaeric transcripts spanning the breakpoints of 74 (66.7%) loci (lower part of Fig. 4b and Supplementary Table 6). In addition to

the consistency between the breakpoints of the chimaeras and the breakpoints of the duplication blocks (Fig. 1b), the detection of chimaeras in multiple tissues (Supplementary Table 9; Methods), validation in PCR with reverse transcription (RT-PCR) experiments

(Supplementary Table 10) and their presence in PacBio long-read transcriptome data (Extended Data Fig. 7a–e) all support that chimaeras show bona fide expression rather than leakage^{72,73}. Among 196 duplication loci involving partial duplicates in humans, we assembled chimaeras for 52 (26.5%) loci (lower part of Fig. 4b and Supplementary Table 6). This proportion is lower than that of flies, possibly due to the high gene density and thus high density of transcriptional activators in the latter species^{46,47}. Due to a lack of samples, we could not perform experimental evaluations in humans. However, the bona fide status of the chimaeras was still supported by the compatibility between the chimaeras' breakpoints and the boundaries of the duplication blocks (Fig. 1b and Supplementary Table 6) and the presence of chimaeras across multiple tissues (Supplementary Table 9).

The assembled transcripts or long reads show that chimaeras generally follow the expected exon–intron structure. Taking *Sherpa*–CG2469 in flies as an example (Fig. 4c), the 5' region of *Sherpa* was fused with the 3' region of CG2469, with the splicing structure of *Sherpa*–CG2469 being largely inherited from the ancestral sequence. Note that one PacBio circular consensus sequencing read (CCS-3) showed three potential intron gains, as supported by raw subreads (Extended Data Fig. 7e), which could be facilitated by cryptic regulatory sites. Changes in gene structure have also been observed for old duplicates^{26,74}.

We examined how 126 (74 in flies and 52 in humans) assembled loci were distributed across the six categories (lower panel of Fig. 4b). Similar patterns were detected across the species. Specifically, duplicates involving 5' genes (potentially including promoters) are the most likely to be transcribed as new chimaeras ($P \leq 0.07$), followed by 3' and internal duplications. Compared with 3'–intergenic duplicates, 3'–3' duplicates are more likely to be expressed. This result is probably due to the higher gene density of 3'–3' duplicates, for which promoter exaptation is more likely to occur²¹. Internal chimaeras are seldom detected because either nonsense-mediated messenger RNA decay leads to the degradation of the transcripts⁷⁵ or alternative splicing maintains the ancestral single-copy gene structure²⁵.

Therefore, flies and humans show different biases in chimaeric arrangements and distinct arrangements have distinct propensities to undergo exon shuffling.

Shuffling often uses intronic breakpoints and generates in-frame proteins in humans. We expected more exon shuffling with intronic breakpoints in humans due to the intron-rich architecture^{46–48}. We tested this with 5'–3' and intragenic duplicates (Fig. 4a; Methods) since only these two arrangements could shuffle protein-coding exons with introns as spacers.

Consistently, for 12 5'–3' duplicates in flies, 20 exonic breakpoints were used (Extended Data Fig. 8), while in humans, all nine 5'–3' duplicates used intronic breakpoints (FET $P = 2.1 \times 10^{-8}$). Across the two species, for 16 of the 21 duplicates, there are fusions of coding sequences (CDS). Except for eight cases disrupted by retained introns or codon phase incompatibility (for example, Extended Data Fig. 7f), the others encode in-frame chimaeras, with a higher proportion in humans (though not statistically significant, 62.5% versus 37.5%; Extended Data Fig. 8). Domain shuffling occurred in five of the eight in-frame duplicates (62.5%) and mid-domain breaks were found for four cases. This phenomenon has been reported for functional fusion genes^{22,76}, suggesting that new functions could emerge via partial domain combinations.

Nineteen internal exonic duplications with assembled chimaeras could be divided into four scenarios (Supplementary Table 6): (1) for duplicates with at least one breakpoint in 5' untranslated regions (UTRs), a longer inclusive transcript with a premature termination codon (PTC) or an alternatively spliced exclusive transcript is expressed (Fig. 5a); (2) for coding exon duplications with two

intronic breakpoints, both inclusive transcripts with two copies of exon(s) and two mutually exclusive transcripts could be detected (Fig. 5b); (3) for coding exon duplications with at least one exonic breakpoint, transcripts with a chimaeric exon emerge and the exclusive transcript is also present (Fig. 5c); and (4) for duplications with at least one breakpoint in the 3' UTR, 3' UTR elongation occurs (Fig. 5d). The distribution of the four types differs between flies and humans, with humans harbouring more duplications with intronic breakpoints (FET $P = 0.008$; Fig. 5e and Extended Data Fig. 4c).

We predicted that some cases could encode chimaeric proteins facilitated by the intronic breakpoints. For ten out of 19 internal duplications involving the UTRs (for example, Fig. 5d and Extended Data Fig. 9a), the open reading frames (ORFs) were not affected. The remaining nine cases could be divided into three scenarios: (1) intron retention or exon elongation occurs in all fly duplicates (Fig. 5a,c and Supplementary Table 11) and in one human duplicate (Extended Data Fig. 9b) where the chimaeric intron could not be recognized by the splicing machinery, possibly due to a misinteraction of splicing signals³⁰; (2) two human duplicates show phase incompatibility (Extended Data Fig. 9c); and (3) two human duplicates encode an in-frame chimaeric transcript, among which C6 was the only case with domain shuffling and a partial domain was duplicated. Thus, the proportion of in-frame shuffling is 40% in humans (two out of five) and 0% in flies (zero out of four).

Since internal duplications are selected against (Extended Data Fig. 4b) and the corresponding chimaeric transcripts are relatively less frequently detected (Fig. 4b), we predicted that: (1) alternative splicing might occur to rescue the preduplication ORFs²⁵; and (2) the short exclusive transcript (the equivalent to the ancestral transcript) represents the major isoform due to their better-optimized splicing signals and/or due to the decay of frame-disrupted inclusive transcripts. Consistently, for eight of nine cases (Fig. 5f), including all seven cases with ORF disruption, the exclusive transcript was assembled. Moreover, read-depth analyses revealed that inclusive transcripts always show lower expression than the corresponding exclusive transcripts (median ratio = 0.2; Fig. 5f and Extended Data Fig. 9d). For three out of four cases in flies, we confirmed this pattern (median ratio = 0.06; Extended Data Fig. 9e) by quantitative PCR (qPCR; Methods). The failed case is *Myo81F*, which may not be quantifiable by qPCR due to its low expression. The final case with only an inclusive in-frame transcript assembled involves human *SPG11*. Even for this case, read-depth analyses suggest that the exclusive transcript should be present and that its expression level is higher than that of the inclusive transcript (Fig. 5f and Extended Data Fig. 9d).

We additionally analysed the previously generated control data to test whether the genomic architecture and duplication sizes explained the between-species differential use of intronic breakpoints. Our observed data fit the control data, with random partial duplications more likely to harbour intronic breakpoints in humans than in flies (Extended Data Fig. 4c). Thus, the disproportionate use of intronic breakpoints in humans is mainly a consequence of the intron-rich genomic background and small duplication size.

Altogether, 5'–3' gene fusions and internal duplications more often use intronic breakpoints in humans than in flies and for both arrangements more in-frame shuffling has occurred in humans.

Complete and partial duplicates rarely show high frequencies.

Finally, we examined the allele frequency distributions of polymorphic duplicates in flies and humans. We expected duplications, especially large ones, to show low frequencies because CNVs are generally subject to purifying selection^{20,52,77}. In agreement, we found that the allele frequency spectra of complete and partial duplicates in flies and humans were left-skewed compared with those of nearly neutral synonymous mutations and similar to those of PTCs (Fig. 6a). Within each type of duplicate, those larger have lower frequencies

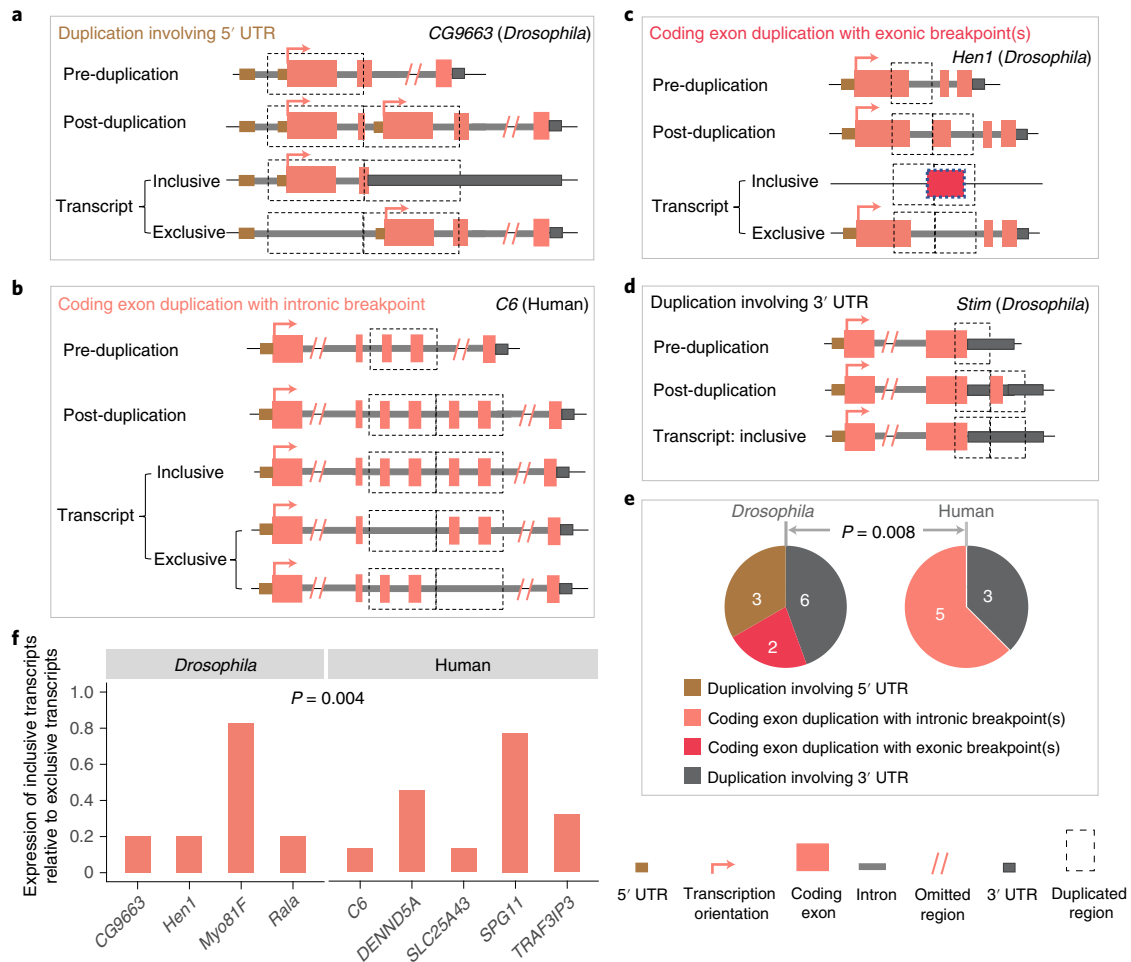


Fig. 5 | Internal duplications often lead to alternative splicing. **a–d**, Representative examples involving 5' UTR (**a**), intact coding exon (**b**), partial coding exon (**c**) and 3' UTR (**d**). In **a** and **b**, we show both an inclusive longer transcript spanning the tandemly duplicated regions and an exclusive transcript spanning only one copy. In **a** and **c**, the coding exon is elongated by fusion with neighbouring introns as indicated by a long transcript (**a**) or a partially assembled transcript (**c**, highlighted by the dashed frame). **e**, Distribution of four arrangements. P was calculated with Fisher's exact test. **f**, Relative expression level of isoforms. For nine internal duplications with the frame potentially disrupted, we calculated the expression ratio between inclusive and exclusive isoforms. P was calculated with a Wilcoxon signed rank test against the expected median of 1 across nine cases.

(Extended Data Fig. 10a). Thus, our data support that the functional consequences (for example, dosage increase) of duplications are generally deleterious.

The most notable between-species difference was found in intronic duplications: in flies, these duplications are selected against, like complete or partial duplications; but in humans, they appear to be neutral, like synonymous mutations (skewness 0.18 versus 0.14; Fig. 6a). Consistently, intronic CNVs are subject to purifying selection in flies but are tolerable in humans^{20,78}. Moreover, possibly due to the compact gene structure of flies, intronic duplications can cause minor downregulation in flies but lead to no expression changes in humans (Fig. 3a). Since one duplication block can involve both complete and partial duplicate(s) (for example, Fig. 4c), we repeated these analyses with duplications spanning a single gene and obtained similar patterns (Extended Data Fig. 10b).

The similar allele frequency spectra of the three types of duplications in flies predicted that their relative abundance among different frequency ranges should be constant. In contrast, we expected intronic duplications to increase in proportion from moderate to high frequency in humans. To test this, we focused on a subset of fly duplications shared by one published dataset, which covers 38 DGRP lines and provides high-frequency resolution⁵². We

confirmed our two expectations (Fig. 6b): (1) the proportions of complete, partial and intronic duplicates are roughly constant in flies, with partial duplicates always being the major group; and (2) intronic duplications only account for 38.6% of the singleton group in humans but become the majority (>75%) with increasing frequencies. Analyses based on the frequency data from six fly lines and analyses based on duplications spanning only one gene showed similar results (Extended Data Fig. 10c).

For partial duplicates, we examined the differences in allele frequency spectra between different arrangements (Fig. 4a) or between partial duplicates with assembled transcripts and the other partial duplicates. We expected each arrangement to have the chance of spreading in the population. Consistently, despite small sample sizes, we detected non-singleton cases for most arrangements (Fig. 6c). Within each species, various arrangements showed similar proportions and internal duplications consisted of a relatively higher proportion of non-singletons. Thus, although internal duplications are generally deleterious and selected against (Extended Data Fig. 4b), these observed duplications are less deleterious because of rescue effects through alternative splicing or because some inclusive isoforms (Fig. 5) are beneficial. We also expected that partial duplicates supported by assembled chimaeras would exhibit higher frequencies

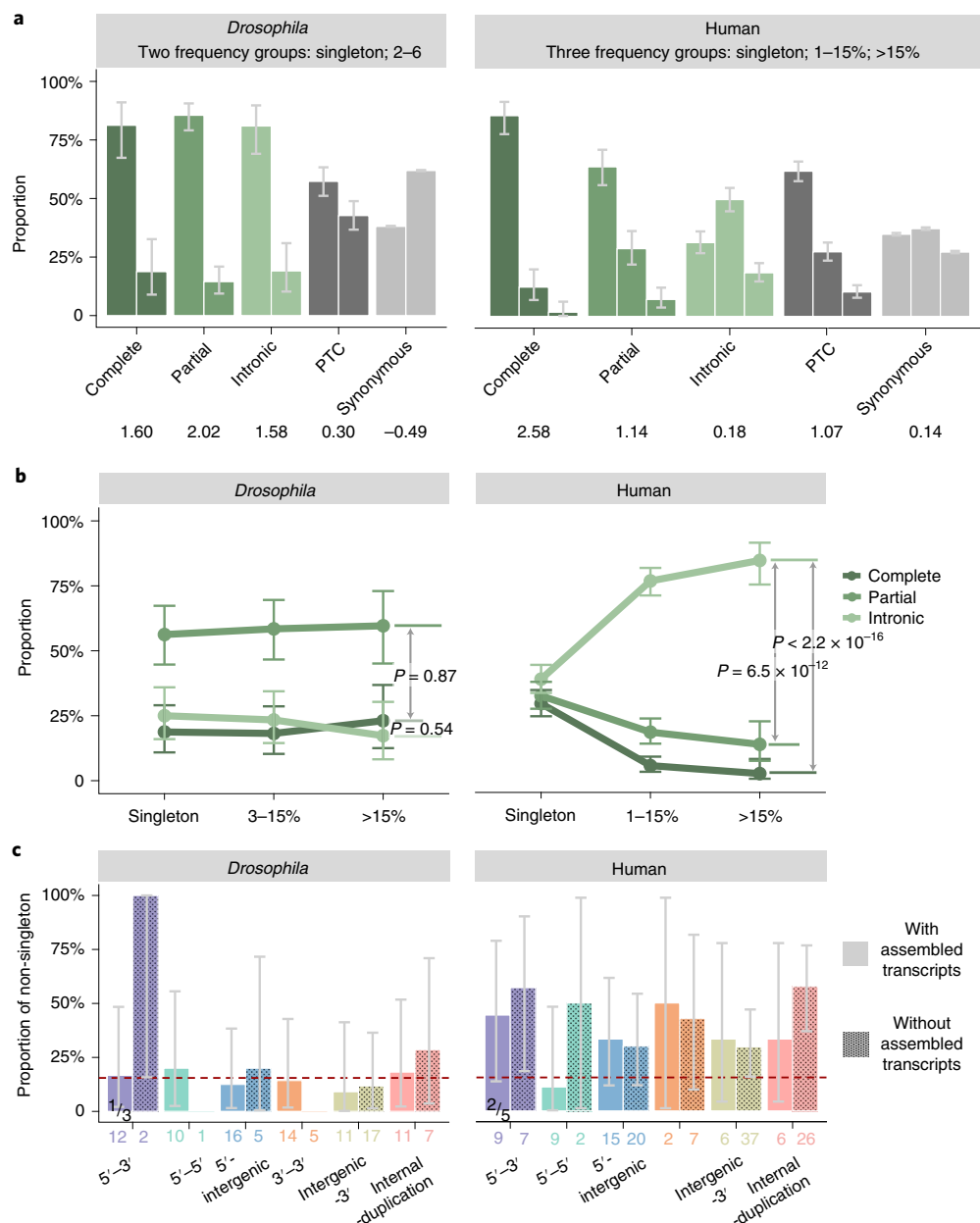


Fig. 6 | Distribution of duplicate genes with respect to allele frequencies. **a**, Allele frequency spectra of complete, partial and intronic duplicates, and of PTCs and synonymous substitutions in *Drosophila* and humans. The skewness values are presented under each group. **b**, Proportion of complete, partial and intronic duplicates across different frequency groups. The P values were calculated by one-sided Fisher's exact tests. **c**, Proportion of non-singleton partial duplicates with or without assembled chimaeric transcripts. The median proportions across all 12 bars are plotted as dashed lines and the total numbers of each scenario are shown at the bottom. For 5'-3' fusions with assembled transcripts, the number of non-singleton in-frame coding fusions and the total number of in-frame coding fusions are also shown (for example, one out of three for flies). Note that singleton corresponds to an allele frequency of 16.7% in flies and 0.5% in humans, respectively. For all panels, the 95% CI, calculated via binomial distribution, is shown as the error bar.

if they had required time to evolve the chimaeric expression. But this was not the case as each type of arrangement, including 5'-3' in-frame fusions, shows a similar proportion of non-singletons in expressed and non-expressed duplicates (Fig. 6c). Thus, chimaeric expression may emerge immediately on duplication.

In summary, except for human intronic duplications, duplications are generally selected against.

Discussion

Here, we generated a comprehensive portrait of polymorphic duplicates in flies and humans. Our results unmasked dosage sensitivity as a key constraint on the origination of complete duplicates and

identified the importance of introns for intergenic and intragenic exon shuffling.

First, we revealed that only complete duplicates show strong dosage increases. These results suggest that the previous conflict^{12,15,17,42–44} over whether or not gene duplications are associated with expression increases could be the result of the misclassification of complete and partial duplicates caused by the uncertainty of duplication boundaries for old duplicates and for polymorphic duplicates identified in low-quality resequencing data. Moreover, we found that these polymorphic complete duplicates are subject to dosage constraints caused by stoichiometry in protein complexes. Such data concur with the dosage rebalancing process observed in

old duplicates^{14–17}. As for dosage sensitivity of X-linked genes, our data indicate that X dosage compensation systems can cope with duplication via various strategies. That is, fly duplicates tend to be closer to HASs and thus well compensated, and human duplicates play sex-biased roles and therefore do not need to be compensated. However, due to the small numbers of X-linked duplicates, these patterns need future testing.

Second, we showed that most exon shuffling in humans uses intronic breakpoints, which often generates in-frame proteins. Despite the small numbers of polymorphic duplications, the significance of introns for exon shuffling is in line with previous reports that intron-rich genomes show stronger exon–domain correspondence^{31,33} and harbour more old intragenic duplications^{25,26}. Besides shuffling, introns also enable alternative splicing, as Gilbert initially predicted²⁷. Intragenic or intergenic chimaeras thus emerge by the co-option of the splicing machinery. As the most relevant category, internal duplicates often encode pre-existing transcripts as the major isoforms and maintain the ancestral function. Thus, the link between mutually exclusive alternative splicing and old tandem exon duplications^{24,25} rapidly emerges at the population level. Given the role of alternative splicing in maintaining the preduplication function, it may be a prerequisite to intragenic shuffling, which in turn enables more complex protein structures^{24,29}. Therefore, by studying polymorphic duplicates with known preduplication exon–intron structures, we provide direct evidence that exon shuffling occurs with introns enhancing the recombination between and within genes, and that alternative splicing facilitates this process. In this aspect, the two fundamental concepts of Gilbert²⁷ are intimately linked.

Since the 1930s, complete duplications are viewed as the pre-dominant force for new gene functions^{1,5}. After Gilbert's work in 1978²⁷, exon shuffling as a new force became widely appreciated^{80,81}. By studying the earliest stage of gene duplication, we showed that both forces have been constrained by mechanistic factors, that is, dosage sensitivity and exon–intron architecture. Such patterns echo the need to anchor understanding of evolutionary processes with a mechanistic perspective⁸².

Taken together, we painted a landscape of polymorphic duplicates in two key species, which are jointly shaped by dosage sensitivity and exon shuffling.

Methods

Fly work and sequencing. We used the Illumina HiSeq 2500 platform to sequence genomes from one North American population, that is, DGRP³¹. Among a total of 205 DGRP lines, we targeted six of the 40 core lines (RAL-208, RAL-379, RAL-399, RAL-427, RAL-517 and RAL-799) for resequencing due to the following reasons (Extended Data Fig. 1): (1) core lines tend to be associated with the publicly available polymorphic duplication datasets^{52,53}; (2) lines infected with *Wolbachia* were excluded; (3) lines harbouring known inversions were excluded⁵³; (4) lines with segregating variants accounting for 2% or more sites in at least one chromosome were excluded; and (5) to control for closely related lines⁵³, representative lines were selected according to the tree that was reconstructed based on synonymous variants annotated in the DGRP Freeze 2.0 (Extended Data Fig. 1). We performed PCR and confirmed that all six lines did not show between-line contamination. That is, we amplified three highly divergent regions (Supplementary Table 12), performed Sanger sequencing and checked single-nucleotide polymorphism (SNPs) on the basis of the DGRP annotations. We also confirmed that no *Wolbachia* infection occurred in these lines using previously described primers (Supplementary Table 12)⁸³. All the lines were reared on standard yeast–glucose medium at 25 °C, with 60% humidity and under a 12-h light/12-h dark cycle.

For DNA sequencing, we used female flies because the Y chromosome was poorly assembled in the *D. melanogaster* reference genome r6.02, and so that we could have the same depth for the X chromosome as for the autosomes. We used the DNeasy Blood & Tissue Kit (QIAGEN, catalogue no. 69504) to extract DNA from 300 or more flies. RNase A (Thermo Fisher Scientific) was added for the digestion of the total RNA. Gel electrophoresis was used to check the quality of the genomic DNA and the concentration was measured using Qubit (Qubit2.0, Life Technology) before library construction and sequencing. Genome DNA samples (10 µg) without apparent degradation were used for library construction and sequencing by Illumina.

For RNA sequencing, we dissected seven fly tissues, namely, the accessory gland, female fat body, female head, female midgut, male fat body, ovary and testis. These tissues were selected on the basis of the ease of dissection and to increase the transcriptome diversity by including major somatic and germline tissues. Note that the fat bodies were dissected from L3 stage larvae and all other tissues were obtained from 2–3-day-old adults (Supplementary Table 4). The larvae were washed twice with 1× PBS before dissection to remove any residual yeast from the food. After dissection, each sample was immediately stored in 1 ml of TRIzol reagent (Life Technologies, catalogue no. 15596-018). Each sample was homogenized with a Minilys instrument at the highest speed for 1 min and then subjected to the TRIzol protocol for RNA extraction. Gel electrophoresis was used to check the RNA quality and the RNA concentration was measured by Qubit. The RNA was then used for Illumina stranded RNA-seq. In addition, RNA samples from the testes of RAL-379 2–3-day-old adults without degradation were used for PacBio Iso-seq.

Duplication calling, PCR validation, classification and population analyses in flies. We mapped the DNA-seq reads against the *D. melanogaster* genome r6.02 using NovoAlign v.3.03 (<http://www.novocraft.com/products/novoalign/>) given its high accuracy in short-read alignment⁸⁴. We specified the ‘-r Random -R 10’ option in NovoAlign to handle multiple-mapping reads. For reads with multiple equally possible alignment positions, the ‘-r Random’ option enabled the random output of a single alignment. The ‘-R’ option specified how the equally possible alignments were defined. By increasing the default value from 5 to 10, we generated a more stringent set of uniquely mapped reads.

We used four published packages to call duplications: Delly v.0.7.2 (ref. ³⁷), CNVnator v.0.3 (ref. ⁵⁶), Breakdancer v.1.2.6 (ref. ⁵⁴) and Pindel v.0.2.5.a8 (ref. ⁵⁵). Specifically, Delly was used to call duplications (-t DUP). Only ‘precise’ calls supported by both split-read and paired-end methods were retained for further analyses. The bin size in CNVnator was set to 100 bp, as previously suggested for high read-depth data⁵⁶. With Breakdancer, ITX-type calls supported by at least three split reads were retained for further analyses. For Pindel, the options were set to ‘-l false -k false -T 8 -w 1’ to save memory and runtime, as suggested by the manual. Only calls supported by at least three split reads were retained for further analyses. In addition to the four datasets called by these four programmes, we also included two public datasets covering the same six DGRP lines^{52,53}. Because these datasets were based on independent sequencing data and/or different bioinformatic pipelines, these calls were complementary to our results.

We also developed a new split-read-based pipeline that took advantage of our relatively longer read length (250 bp) because public programmes often call duplications on the basis of both split reads and discordant read pairs. Our pipeline can be described as follows: (1) we identified ‘abnormal reads’ if a large (>20 bp) soft clip was found in the read alignment; (2) we remapped the soft-clip parts against the reference genome; (3) we detected discontinuously aligned reads as those mapping to the neighbouring region with identical orientations; (4) we clustered these discontinuous reads with a confidence interval (CI) (<50 bp)-based strategy; and (5) we required at least three discontinuously aligned reads or split reads to support a duplication event and to be conservative, the innermost breakpoint was selected as the duplication boundary.

We merged the seven datasets into one non-redundant dataset. We first filtered less reliable calls by limiting the duplication range to 50 bp to 25 kb, which is similar to that used in previous studies (for example, 25 bp to 25 kb; ref. ⁴²). To balance sensitivity and specificity, we performed two rounds of CI-based merging to generate our final calls: (1) within each line, we accepted a duplication call simultaneously identified by at least three out of seven datasets if the breakpoints fell within a range of 200 bp (a reasonable cutoff given our median insert size, ~600 bp); (2) to be conservative, we specified the innermost coordinates of all supporting calls as the breakpoints of the duplication; (3) for between-line comparisons, we examined whether one call in one line was shared (defined by the CI of 200 bp) by other lines; and (4) to simplify the gene-based analysis, we manually excluded recurrent duplications, that is, duplications present in several lines but with different breakpoints across lines (for example, *Sdic*⁸⁵). After generating a merged dataset, we excluded loci with potentially unreliable read mapping where 500 bp flanking regions of two breakpoints shared similarity.

Notably, four out of 270 genes, *CG40470*, *laccase2*, *sima* and *CG45784*, harbour two distinct duplication loci. To simplify the downstream gene-centric analyses, we chose the more likely true positives: (1) we only took the high-frequency calling for *CG40470*, *laccase2* and *CG45784*; and (2) for *sima*, its two callings have the same frequency and thus we focused on the potentially more reliable shorter case.

To estimate whether our calls represent duplications in the DGRP line(s) rather than deletions in the reference line, we took advantage of the syntenic genomic region provided by the UCSC genome browser⁸⁶. That is, we performed polarization by aligning (BLAST) our duplication regions (one copy or two copies) with closely related species, including *D. simulans* and *D. sechellia*. In 166 out of 179 (92.7%; Supplementary Table 3) cases, we found that the alignment length was largely similar between one copy and two copies, which suggested a recent duplication in *D. melanogaster*. In contrast, 13 cases were unresolved due to the absence of syntenic regions in both *D. simulans* and *D. sechellia*. We concluded that the vast majority of duplications represent the derived rather than the ancestral state.

To evaluate the accuracy in calling duplications, we performed PCR experiments as in previous efforts^{20,42,52,87,88}. In these efforts, a typical design involved: (1) designing divergent primers to span the breakpoints between two tandem copies; (2) genotyping one predicted duplication-present line and duplication-absent line (sometimes the reference line) as the control; and (3) testing 20 to 70 duplication loci with validation rates per locus ranging between 50% and 96%. We improved on this previous design: (1) to perform a realistic estimation, we selected 31 duplications by requiring the median size of this test dataset to be similar to the whole dataset (2,246 bp versus 2,286 bp) since long duplications presumably have larger functional impacts and are more likely to be selected against and so these calls may represent false positives (Extended Data Fig. 10a; ref. 42); and (2) we examined whether our computational calls were consistent with PCR experiments across all six fly lines in addition to the control line (the reference line). For 31 duplications, we failed to confirm five calls (Supplementary Table 5): (1) the duplication call *Drosophila*-24 was computationally identified in RAL-427, RAL-379 and RAL-799 and we confirmed its presence in RAL-427 and RAL-799, its absence in RAL-208, RAL-399 and RAL-517, but failed to confirm its presence in RAL-379; (2) similarly, for *Drosophila*-125, we confirmed its presence in RAL-799, its absence in four lines, but failed to confirm its presence in RAL-399; (3) in a third case, we predicted a duplication across six lines and PCR experiments showed its presence in all six lines and in the reference line, thus it represents a hidden duplication that failed to be assembled in the reference genome; (4) in the fourth case, a predicted long (~22 kb) singleton in RAL-517, PCR rejected its presence in this line but confirmed its absence in the other five lines; and (5) the fifth case is similar to the third except that we predicted the absence of this call in RAL-799 and PCR detected its presence in this line. Thus, for a total of 186 (31 calls across six lines) calls, the false positive rate is 1.6% (3/186) and the false negative rate is 0.5% (1/186). If we defined the accuracy at the locus level, the aforementioned *Drosophila*-24 and *Drosophila*-125 (cases 1 and 2) would be validated as correct. Together with the fact that two calls (cases 3 and 5) represent errors in the reference genome assembly, our calling pipeline worked for 30 loci (96.8%).

Since tandem and dispersed duplications leave different sequence signatures⁸⁹, we specified the parameters of split-read- and read-pair-based calling methods to identify tandem duplications. Thus, although read-depth-based method can detect dispersed duplications, we only identified tandem duplications by extracting intersections across the methods. Consistently, the band sizes or breakpoints revealed by PCR or RT-PCR experiments demonstrated that most duplications are tandem (Supplementary Table 5 and Fig. 1b).

We extracted and classified duplications overlapping with protein-coding genes. Because a gene might encode several isoforms, we selected the longest principal transcript (the potentially functionally most important transcript) to represent the corresponding gene⁹⁰. We performed a classification (Figs. 2a and 4a) according to which part of a protein-coding gene was duplicated. Notably, the predominance of partial duplicates in this dataset was consistent with an early study, which used tiling arrays²⁰ and identified even fewer complete duplicates. This stronger depletion could be an artefact of the limits of the array, where complete duplicates were mistakenly called as partial duplicates due to random fluctuation of hybridization signals.

We examined how complete and partial duplicates were distributed with increasing allele frequency. For duplicates shared with the previous call based on 38 lines⁵², we divided the duplicates into three groups according to their frequencies: singleton, 3–15% and >15%. On the basis of these cutoffs, the numbers of complete duplicates were 15, 14 and 12. Thus, each group had a relatively large sample size. Since the inbreeding process of flies may fix slightly deleterious duplicates and remove strongly recessive deleterious duplicates, we needed to compare the allele frequency spectra of duplicates relative to those of synonymous substitutions and PTC mutations within the same dataset. We thus identified these two types of mutations in our resequencing data from the six lines. That is, we aligned reads against the reference genome using BWA mem v.0.7.10 (ref. 91) with the ‘-M’ option to prepare the input file for Picard v.1.119 (<http://broadinstitute.github.io/picard>), which was used to mark PCR duplicates with the parameter set ‘MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000’ to improve the computational performance. As described by ref. 92, we used GATK v.3.5 to recalibrate the alignment and identify high-quality SNPs. For each SNP, we inferred its functional effect with SnpEff v.4.3 (ref. 93). On the basis of the data from the six lines, we then generated allele frequency spectra across complete, partial and intronic duplicates, PTCs and synonymous substitutions.

Transcriptome quantification, de novo assembly and RT-PCR in flies. We processed the RNA-seq data using the STAR v2.5 (ref. 94) (mapping) and RSEM v.1.3 (quantification) pipelines given their high accuracy^{95–97}. For RSEM⁹⁸, we calculated the transcripts per million (TPM) values. We used the default parameters with the exception of the option ‘--forward-prob=0’, which was compatible with our strand-specific RNA-seq libraries. On the basis of transcriptome data across seven tissues, we calculated the τ index to quantify expression bias⁹⁹ and took 0.9 as a conservative cutoff¹⁰⁰ to define tissue-specific genes.

For each gene in each tissue and in each line, we calculated the average expression across two replicates as the expression level in this tissue of this line. For

duplicate genes with a median expression level in a tissue across duplication-absent lines higher than 1, we quantified the change in expression as a ratio:

$$\frac{\text{median TPM value of lines with the duplication}}{\text{median TPM value of lines without the duplication}}$$

Here, we used a cutoff of 1 to remove the lowly expressed genes because their fold changes would be less reliable. We then calculated the median ratios across tissues as a summary statistic for the gene.

To test whether incomplete duplications lead to novel gene structures or chimaeric genes, we performed whole-genome de novo transcriptome assembly and analysed the differences between the assembled transcripts and the reference genome. Specifically, we assembled the whole transcriptome using Trinity v.2.6.5 (ref. 101) by specifying strand-specific RNA-seq libraries with the parameter ‘--SS_lib_type RF’. We then used BLAT v.35 (ref. 102) with the default parameters to search the Trinity contigs against a two-copy (postduplication status) library of all loci and the genome. We only retained the contigs that uniquely spanned breakpoints. We found that breakpoints harboured by contigs and those of duplication blocks only differed by a median of 2 bp (Fig. 1b). This deviation was mainly due to the read-depth and discordant read-pair methods, which did not reach base-level resolution as do split-read-based methods¹⁰³.

We further estimated the quality of contigs using three strategies.

First, chimaeras were independently detected in a median of three tissues (Supplementary Table 9). Second, among a total of 74 chimaeric loci supported by contigs, we tested 66 (89.2%; Supplementary Table 10) cases with suitable primers and available fly lines (RAL-799 died out later). That is, we performed RT-PCR experiments using two primers flanking each breakpoint, which was followed by Sanger sequencing. Twenty 2–3-day-old adults (ten females and ten males) were homogenized and total RNA was extracted. We confirmed 52 (78.8%) cases and failed in 14 cases, possibly due to primer inefficiency because the breakpoints of these duplications were also congruent with those of the duplication blocks. Third, for 11 chimaeric genes detected in the testes in the RAL-379 line, we generated PacBio Iso-Seq data and examined the gene structures. We mapped long reads to the genome using BLAT and examined whether the circular consensus sequencing (CCS) reads spanned breakpoints of duplication blocks. Despite the known bias of PacBio toward high-abundance transcripts due to low throughput of the technology^{21,50}, we were able to confirm five (45.5%) cases.

We performed a series of curations with respect to frameshifts, domain changes and alternative splicing. We examined whether frameshifts occurred by translating the sense strand of the longest assembled contigs with NCBI ORFfinder. As for alternative splicing, we directly assembled exclusive isoforms for six out of eight cases in Fig. 5f. For the remaining *Hen1* and *Myo81F*, we failed to assemble exclusive transcripts. Thus, we performed RT-PCR using primers (Supplementary Table 12) targeting two exons outside of the duplication blocks to amplify the inclusive and exclusive transcripts simultaneously.

We estimated whether inclusive or exclusive transcripts represent major isoforms by calculating the ratio of the read depth within duplication blocks to the read depth outside of duplication blocks (Extended Data Fig. 9d). The read depth was calculated using SAMtools v.1.3 (ref. 104). To take the fluctuation in the RNA-seq read depth across different parts of genes into account, we separately calculated this ratio for duplication-absent and duplication-present lines. The ratio of duplication-absent lines served as a control; for a given gene, if the duplicated region of the duplication-present line was not transcribed as a part of inclusive transcripts, the ratios of duplication-absent and duplication-present lines should be equal. We performed these calculations only for tissues in which the focal gene is expressed (TPM > 1). To increase the statistical power, read-depth ratio in each tissue replicate of each line served as one data point. We finally used a Wilcoxon signed rank test to measure whether the ratio is different between duplication-absent and duplication-present lines.

We further investigated the expression of four cases in flies by qRT-PCR. For inclusive transcripts, we designed primers spanning the breakpoint, and for inclusive/exclusive transcripts, we designed primers targeting the constitutive exons (Supplementary Table 12). We selected *SdhA* as the internal control for the qRT-PCR analysis because this gene shows constant expression levels in a narrow ΔC_t (cycle threshold) range of 11.8 to 12.3 across the lines. Total RNA was extracted from 2–3-day-old adults using the TRIzol reagent (Life Technologies, 15596-018), reverse transcribed with SuperScript III First-Strand Synthesis System (Life Technologies, 18080051) and then quantified with PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, A25743). For each case, both the duplication-present line and duplication-absent line were quantified in parallel and three technical replicates of each sample were included.

Duplication calls and targeted de novo assemblies in humans. Although several sets of population resequencing data and transcriptome data are available for humans (for example, 1000 Genomes Project¹⁰⁵), GTEx⁵⁹ is the most suitable for our purpose given its large sample size of individuals and its high transcriptome diversity (a total of 53 tissues or cell lines with a median of nine per individual). We obtained duplication datasets from ref. 36, GTEx v.6 expression quantification data from GTEx Portal and duplication genotyping data from dbGaP (<https://www>).

ncbi.nlm.nih.gov/gap/). GTEx called duplications by both LUMPY (split-read- and read-pair-based method) and Genome STRiP (based on the read depth) with a different resolution of breakpoints (96 bp and >1 kb, respectively). Because most of our analyses depend on the resolution of breakpoints and the LUMPY dataset consisted of more duplications, we mainly used this dataset. However, since the Genome STRiP dataset tended to consist of larger duplications (median size, 4,999 bp versus 576 bp)³⁶, we repeated the analyses when necessary. Similar to flies, LUMPY was used to identify tandem duplications in GTEx data³⁶.

We performed two filters for the original GTEx dataset. First, the original LUMPY dataset involves 1,079 genes, 115 (10.7%) of which harbour more than one duplication loci. Since most of them are singletons, we could not differentiate them on the basis of frequency as we did in flies. Considering their small proportion, we simply focused on the remaining cases to ensure the downstream gene-centric analyses. Similar filter was applied for Genome STRiP dataset and 71 genes were excluded. Second, the GTEx release of duplications described by ref.³⁶ includes 147 individuals. Among these, the individual GTEx-UPIC consists of 323 duplications in the Genome STRiP dataset, whereas all the others consist of less than 100 duplications. Similarly, the individual GTEx-S7SE also consists of a disproportionately high number of duplications in the LUMPY dataset. We excluded these two individuals. After these filters, duplications in the LUMPY dataset ranged between 97 bp and 1.0 Mb (800 genic duplications or 964 duplicate genes; Supplementary Table 2), while those in the Genome STRiP dataset ranged between 2.0 kb and 4.1 Mb (90 genic duplications or 188 duplicate genes; Supplementary Table 2).

For the frequency-related analyses, because the 145 human individuals were from different populations³⁶, we only retained 120 European individuals. Moreover, for simplification, we excluded X-linked variation in these frequency-oriented analyses and used the haploid copy number to calculate the frequencies for the split-read-based dataset. Similar to flies, we defined three frequency groups: singleton (one diploid allele), 1–15% (two to 36 diploid alleles) and >15% (37 or more). Because information on homozygous or heterozygous status was not available in the Genome STRiP dataset, we did not use this dataset in all frequency-related analyses.

To reduce the computational cost, we only performed polarization analyses for high-frequency duplication calls, which were more likely to represent deletions in the reference genome. That is, for 26 duplications present in at least ten individuals in the split-read-based dataset and 34 duplications present in at least ten individuals in the read-depth-based dataset, we estimated that most of them represented bona fide duplications rather than deletions in the reference genome by performing polarizations against outgroups as we did in flies (Supplementary Table 3).

Because GTEx generated thousands of RNA-seq libraries, we used targeted de novo assemblies to control the computational cost. That is, we searched RNA-seq reads against a pseudo-two-copy library using hisat2 v.2.1.0 (ref.¹⁰⁶) given its high speed, extracted reads mapped to this library and performed de novo assembly using Trinity. Similar to the protocol used for flies, we further mapped the contigs against the human genome and examined whether these contigs were compatible with the breakpoints of the duplications. We also confirmed that 52 chimaeras were independently detected in a median of seven tissues (Supplementary Table 9).

Note that the duplication locus Human-LUMPY-281 (Supplementary Table 2) only covered the principal transcripts of *KANSL1*, as judged on the basis of the GTEx calling result. However, our transcript assembly indicated that the range of this duplication was actually larger and both *KANSL1* and the neighbouring *ARL17A/B* were included (Supplementary Table 6). Since numerous studies support the larger duplication^{107,108}, we included this locus in Extended Data Fig. 8. In addition, *ARL17A* is linked with a tandem paralogue (identity >99.8%), that is, *ARL17B*. Given such a high sequence similarity, it is difficult to infer whether fusion occurs between *KANSL1* and *ARL17A* or *KANSL1* and *ARL17B*. Both scenarios are mentioned in the literature^{107,108} and we thus arbitrarily named them *KANSL1-ARL17A*.

Estimation of the expected number of polymorphic duplications. Multiple reasons underlie why the fly and human datasets consist of different numbers of duplicates (270 versus 964). The number of neutral segregating sites in a population could be estimated according to the following classical formula¹⁰⁹:

$$\theta = 4N_e\mu$$

$$E(K_i) \div \theta \sum_{j=1}^{i-1} \frac{1}{j}$$

Herein, N_e and μ refer to effective population size and mutation rate, respectively. $E(K)$ indicates the expected number of segregating sites, while i refers to the number of sampled chromosomes. Thus, for homozygous flies, i is 6 and the cumulative sum is roughly 2.28. For heterozygous humans, i is 290 and the sum is 6.24. Thus, with the sample size alone, the number of polymorphic duplications in flies would be 36.5% (2.28/6.24) of that in humans. However, this

drop would be offset by the difference in N_e between flies and humans (1 million versus 10,000)⁶⁰. As for mutation rates, their values are unknown and the relative frequency of the underlying mutational mechanisms (for example, recombination or repair) is different in the two species¹¹⁰. Actually, it is difficult to estimate μ and two evaluations of μ in *C. elegans*¹¹¹ have 100-fold difference. Since μ is quantified as per gene basis, the different gene number of flies and humans (14,000 versus 20,000) also contribute to between-species difference.

Notably, as shown in Fig. 6a, duplications are generally deleterious and this formula does not apply in this context. Presumably, species with bigger N_e (for example, flies) are subject to more efficient selection and thus deleterious mutations would be more efficiently purged⁴⁵. Thus, the expected numbers of polymorphic duplicates harboured by a population or by an individual (i as 1 or 2) is affected by multiple factors and are not straightforward to estimate.

Shuffling of polymorphic duplicates and HAS distance simulation. For both flies and humans, we used the shuffle function of bedtools v.2.15.0 (ref.¹¹²) to randomly shuffle coordinates of each polymorphic duplicates. In this way, we generated one random dataset while keeping the number and size of duplicates the same as the observed data. We repeated the same procedure and generated 100 random datasets. For each dataset, we calculated the proportions of complete, partial or intronic duplicates, six subtypes of partial duplicates and intronic breakpoints of each subtype of partial duplicates. We then examined how the simulated data differed from the observed.

In flies, we wondered whether X-linked complete duplicates were closer to HASs to better achieve dosage compensation⁶⁵. We downloaded the coordinates of 150 HASs⁶⁴ and lifted the coordinates to the genome assembly version used in this work. The median distance between eight X-linked genes and closest HASs was 28,687 bp. To address whether such a distance could be explained by chance, we again used bedtools to randomly pick up eight genomic regions in X chromosome according to the coordinates of these duplicates. We required that the shuffled locus contained at least one complete gene and recorded the median distance. We repeated sampling 10,000 times and found only 1,260 instances for which the median was shorter than 28,687 bp. Thus, the empirical P value was 0.126 (Extended Data Fig. 6b).

Calculation of nucleotide diversity. Nucleotide diversity was measured in windows with the size as 10 kb and step size as 5 kb. We used vcftools (v.0.1.16)¹¹³ to calculate nucleotide diversity across every two lines (individuals) of fly and human. All 15 line combinations and 100 random human pairs were analysed. The medians across all windows were recorded. For fly, the vcf-format file dgrp2.vcf.gz was downloaded from <http://dgrp2.gnets.ncsu.edu/data/website/dgrp2.vcf>; while for human, the file GTEx_Analysis_20160115_v7_WholeGenomeSeq_635Ind_PASS_AB02_GQ20_HETX_MISS15_PLINKQC.PIR.vcf.gz was used for calculation.

Protein-complex data. We downloaded human protein-complex data from the dedicated database CORUM (a comprehensive resource of mammalian protein complexes)¹¹⁴ but there is no such resource for flies. Therefore, we took advantage of Ensembl BioMart¹¹⁵ and extracted genes annotated with GO:0032991 (protein-containing complex). Electronic (less reliable) annotations with GO tags, including ISS (inferred from sequence or structural similarity), NAS (nontraceable author statement) and IEA (inferred from electronic annotation), were discarded, although the pattern (Fig. 3c) remained robust if such data were included.

Figures and statistical tests. We used violin and box plots to summarize the data distributions. These two types of figures are largely similar, except that the violin plot uses curves to show the probability density. For both plots, the bar indicates the interquartile range (IQR) and the line indicates the median. For panels with ten data points or fewer, we overlaid all specific data points within the plots. To zoom onto the median values and IQR across all panels, we specified a suitable lower and upper bound range for the y axis. Sometimes, a small proportion of outliers (greater than the sum of 75% percentile and 1.5-fold of IQR or smaller than the minus between 25% percentile and 1.5-fold of IQR) were not shown in the figures and this was indicated in the legend.

Unless otherwise specified, the 95% CI calculated via binomial distribution was shown as the error bar.

We estimated significance with Fisher's exact tests, Wilcoxon signed rank tests, random samplings or proportion test depending on the specific contexts. Unless specified, we implemented two-sided tests.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We released the raw and processed data. First, the resequencing data, RNA-seq data and Iso-seq data generated in this study were concurrently submitted to the NCBI BioProject database under accession numbers PRJNA681089, PRJNA681417 and PRJNA693662, and the Genome Sequence Archive in National Genomics Data Center (NGDC, part of the China National Center for Bioinformatics) under accession numbers PRJCA004186, PRJCA001789 and PRJCA004319.

Second, the processed fly data (duplication loci, DNA read alignment depth, RNA-seq alignments, assembled chimaeras, PacBio CCS reads and RT-PCR Sanger sequencing data) and human data (duplication loci and assembled chimaeras) were uploaded to the UCSC genome browser as public sessions '<http://genome.ucsc.edu/cgi-bin/hgPublicSessions>' with names as 'Drosophila Duplication' and 'Human Duplication', respectively. We tried to make these tracks as useful as possible: (1) duplication ID refers to Supplementary Table 2; (2) when users click on each duplication ID, the corresponding lines or individuals harbouring this duplication are shown; and (3) assembled chimaeras were split based on breakpoints and separately aligned to the reference genome as Fig. 4c. Finally, the assembled contigs and the RT-PCR Sanger sequencing files were co-submitted to <https://github.com/Zhanglab-IOZ/Polymorphic-Duplication> and <https://sandbox.zenodo.org/record/946570>.

Code availability

The code for split-read-based duplication calling was submitted to: <https://github.com/Zhanglab-IOZ/Polymorphic-Duplication>. It is also archived at <https://sandbox.zenodo.org/record/946570>.

Received: 4 February 2021; Accepted: 10 November 2021;

Published online: 30 December 2021

References

- Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
- Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
- VanKuren, N. W. & Long, M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat. Ecol. Evol.* **2**, 705–712 (2018).
- Brooke, N. M., García-Fernández, J. & Holland, P. W. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392**, 920–922 (1998).
- Bridges, C. B. Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. *J. Hered.* **26**, 60–64 (1935).
- Hahn, M. W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* **100**, 605–617 (2009).
- Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
- Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B* <https://doi.org/10.1098/rspb.2012.1108> (2012).
- Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531 (1999).
- Holland, P. W., Marlétaz, F., Maeso, I., Dunwell, T. L. & Paps, J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Phil. Trans. R. Soc. B* **372**, 20150480 (2017).
- Rice, A. M. & McLysaght, A. Dosage-sensitive genes in evolution and disease. *BMC Biol.* **15**, 78 (2017).
- Giorgianni, M. W. et al. The origin and diversification of a novel protein family in venomous snakes. *Proc. Natl Acad. Sci. USA* **117**, 10911–10920 (2020).
- Gururharsha, K. G. et al. A protein complex network of *Drosophila melanogaster*. *Cell* **147**, 690–703 (2011).
- Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl Acad. Sci. USA* **109**, 14746–14753 (2012).
- Qian, W., Liao, B.-Y., Chang, A. Y.-F. & Zhang, J. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* **26**, 425–430 (2010).
- Lan, X. & Pritchard, J. K. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**, 1009–1013 (2016).
- Chang, A. Y.-F. & Liao, B.-Y. Recruitment of histone modifications to assist mRNA dosage maintenance after degeneration of cytosine DNA methylation during animal evolution. *Genome Res* **27**, 1513–1524 (2017).
- Sangrithi, M. N. et al. Non-canonical and sexually dimorphic X dosage compensation states in the mouse and human germline. *Dev. Cell* **40**, 289–301 (2017).
- Lucchini, J. C. & Kuroda, M. I. Dosage compensation in *Drosophila*. *Cold Spring Harb. Perspect. Biol.* **7**, a019398 (2015).
- Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. & Long, M. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* **320**, 1629–1631 (2008).
- Dougherty, M. L. et al. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res* **28**, 1566–1576 (2018).
- Rogers, R. L. & Hartl, D. L. Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. *Mol. Biol. Evol.* **29**, 517–529 (2012).
- Williford, A. & Betrán, E. *Gene Fusion* (eLS, 2013); <https://doi.org/10.1002/9780470015902.a0005099.pub3>
- Kondrashov, F. A. & Koonin, E. V. Origin of alternative splicing by tandem exon duplication. *Hum. Mol. Genet.* **10**, 2661–2669 (2001).
- Letunic, I., Copley, R. R. & Bork, P. Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.* **11**, 1561–1567 (2002).
- Gao, X. & Lynch, M. Ubiquitous internal gene duplication and intron creation in eukaryotes. *Proc. Natl Acad. Sci. USA* **106**, 20818–20823 (2009).
- Gilbert, W. Why genes in pieces. *Nature* **271**, 501 (1978).
- Gilbert, W. & Long, M. *Walter Gilbert: Selected Works* (World Scientific Publishing Company, 2020).
- Irimia, M. & Roy, S. W. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* <https://doi.org/10.1101/cshperspect.a016071> (2014).
- Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11**, 345–355 (2010).
- Smithers, B., Oates, M. & Gough, J. 'Why genes in pieces?'—revisited. *Nucleic Acids Res.* **47**, 4970–4973 (2019).
- Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.* **7**, 211–221 (2006).
- Liu, M. & Grigoriev, A. Protein domains correlate strongly with exons in multiple eukaryotic genomes—evidence of exon shuffling? *Trends Genet.* **20**, 399–403 (2004).
- Pathy, L. Genome evolution and the evolution of exon-shuffling—a review. *Gene* **238**, 103–114 (1999).
- Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
- Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692 (2017).
- Tuke, M. et al. Large copy-number variants in UK Biobank caused by clonal hematopoiesis may confound penetrance estimates. *Am. J. Hum. Genet.* **107**, 325–329 (2020).
- Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224 (2016).
- Sudmant, P. H. et al. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013).
- Schrider, D. R., Hahn, M. W. & Begun, D. J. Parallel evolution of copy-number variation across continents in *Drosophila melanogaster*. *Mol. Biol. Evol.* **33**, 1308–1316 (2016).
- Newman, S., Hermetz, K. E., Weckselblatt, B. & Rudd, M. K. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am. J. Hum. Genet.* **96**, 208–220 (2015).
- Cardoso-Moreira, M. et al. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Res.* **26**, 787–798 (2016).
- Rogers, R. L., Shao, L. & Thornton, K. R. Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLoS Genet.* **13**, e1006795 (2017).
- Konrad, A. et al. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **115**, 7386–7391 (2018).
- Graur, D. & Li, W. H. *Fundamentals of Molecular Evolution* (Sinauer, 2000).
- Adams, M. D. et al. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
- Sakharkar, M. K., Perumal, B. S., Sakharkar, K. R. & Kanguane, P. An analysis on gene architecture in human and mouse genomes. *In Silico Biol.* **5**, 347–365 (2005).
- Deutsch, M. & Long, M. Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**, 3219–3228 (1999).
- Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
- Ranz, J. & Clifton, B. Characterization and evolutionary dynamics of complex regions in eukaryotic genomes. *Sci. China Life Sci.* **62**, 467–488 (2019).
- Mackay, T. F. et al. The *Drosophila melanogaster* genetic reference panel. *Nature* **482**, 173–178 (2012).
- Zichner, T. et al. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res.* <https://doi.org/10.1101/gr.142646.112> (2013).
- Huang, W. et al. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* **24**, 1193–1208 (2014).
- Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677 (2009).
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

56. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
57. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
58. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
59. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
60. Katju, V. In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. *Int. J. Evol. Biol.* **2012**, 341932 (2012).
61. Zhang, W. Y., Landback, P., Gschwend, A. R., Shen, B. R. & Long, M. Y. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* **16**, 202 (2015).
62. Loehlin, D. W. & Carroll, S. B. Expression of tandem gene duplicates is often greater than twofold. *Proc. Natl Acad. Sci. USA* **113**, 5988–5992 (2016).
63. Stark, A. et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
64. Alekseyenko, A. A. et al. A sequence motif within chromatin entry sites directs MSL establishment on the *Drosophila* X chromosome. *Cell* **134**, 599–609 (2008).
65. Bachtrog, D., Toda, N. R. & Lockton, S. Dosage compensation and demasculinization of X chromosomes in *Drosophila*. *Curr. Biol.* **20**, 1476–1481 (2010).
66. Pandey, R. S., Wilson Sayres, M. A. & Azad, R. K. Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. *Genome Biol. Evol.* **5**, 1863–1871 (2013).
67. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404 (2005).
68. Berletch, J. B., Yang, F., Xu, J., Carrel, L. & Distech, C. M. Genes that escape from X inactivation. *Hum. Genet.* **130**, 237–245 (2011).
69. Shvetsova, E. et al. Skewed X-inactivation is common in the general female population. *Eur. J. Hum. Genet.* **27**, 455–465 (2019).
70. Ji, J. et al. Copy number gain of VCX, X-linked multi-copy gene, leads to cell proliferation and apoptosis during spermatogenesis. *Oncotarget* **7**, 78532–78540 (2016).
71. Zhang, Y., Liu, X. S., Liu, Q. R. & Wei, L. P. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.* **34**, 3465–3475 (2006).
72. Clark, M. B. et al. The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625 (2011).
73. Pertea, M. et al. CHES: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
74. Long, M. & Langley, C. H. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
75. Amrani, N., Sachs, M. S. & Jacobson, A. Early nonsense: mRNA decay solves a translational problem. *Nat. Rev. Mol. Cell Biol.* **7**, 415–425 (2006).
76. Baker, E. P. & Hittinger, C. T. Evolution of a novel chimeric maltotriose transporter in *Saccharomyces eubayanus* from parent proteins unable to perform this function. *PLoS Genet.* <https://doi.org/10.1371/journal.pgen.1007786> (2019).
77. Cooper, G. M., Nickerson, D. A. & Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* **39**, S22–S29 (2007).
78. Rigau, M., Juan, D., Valencia, A. & Rico, D. Intronic CNVs and gene expression variation in human populations. *PLoS Genet.* **15**, e1007902 (2019).
79. Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates, 2007).
80. Walsh, J. B. How often do duplicated genes evolve new functions? *Genetics* **139**, 421–428 (1995).
81. Long, M. Y., VanKuren, N. W., Chen, S. D. & Vibranovski, M. D. New gene evolution: little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
82. Rosenberg, S. M. & Queitsch, C. Combating evolution to fight disease. *Science* **343**, 1088–1089 (2014).
83. Richardson, M. F. et al. Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003129 (2012).
84. Mu, J. C. et al. Fast and accurate read alignment for resequencing. *Bioinformatics* **28**, 2366–2373 (2012).
85. Clifton, B. D. et al. Understanding the early evolutionary stages of a tandem *Drosophila melanogaster*-specific gene family: a structural and functional population study. *Mol. Biol. Evol.* <https://doi.org/10.1093/molbev/msaa109> (2020).
86. Rhead, B. et al. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38**, D613–D619 (2010).
87. Cardoso-Moreira, M., Emerson, J. J., Clark, A. G. & Long, M. *Drosophila* duplication hotspots are associated with late-replicating regions of the genome. *PLoS Genet.* **7**, e1002340 (2011).
88. Rogers, R. L. et al. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Mol. Biol. Evol.* **31**, 1750–1766 (2014).
89. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363 (2011).
90. Manuel Rodriguez, J. et al. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–D117 (2013).
91. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
92. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
93. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6**, 80–92 (2012).
94. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
95. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
96. Ma, Y. et al. Genome-wide analysis of pseudogenes reveals HBBP1's human-specific essentiality in erythropoiesis and implication in beta-thalassemia. *Dev. Cell* <https://doi.org/10.1016/j.devcel.2020.12.019> (2021).
97. Shao, Y. et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res.* **29**, 682–696 (2019).
98. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.* **12**, 323 (2011).
99. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
100. Yang, H. et al. Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a *Drosophila melanogaster* population. *Mol. Biol. Evol.* **32**, 216–228 (2015).
101. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644 (2011).
102. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
103. Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G. & de Ridder, D. Making the difference: integrating structural variation detection tools. *Brief. Bioinform.* **16**, 852–864 (2015).
104. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
105. Durbin, R. M. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
106. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
107. Zhou, J. X. et al. Identification of KANSARL as the first cancer predisposition fusion gene specific to the population of European ancestry origin. *Oncotarget* **8**, 50594–50607 (2017).
108. Oliver, G. R., Jenkinson, G. & Klee, E. W. Computational detection of known pathogenic gene fusions in a normal tissue database and implications for genetic disease research. *Front. Genet.* **11**, 173 (2020).
109. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**, 256–276 (1975).
110. Cardoso-Moreira, M., Arguello, J. R. & Clark, A. G. Mutation spectrum of *Drosophila* CNVs revealed by breakpoint sequencing. *Genome Biol.* **13**, R119 (2012).
111. Katju, V. & Bergthorsson, U. Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol. Evol.* **11**, 136–165 (2018).
112. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
113. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
114. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
115. Kinsella, R. J. et al. Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database* **2011**, bar030 (2011).

Acknowledgements

We thank the DGRP team and the GTEx team for generating and releasing the data. We thank M. Cardoso-Moreira for improving the writing. We thank M. Long, B. He, X. Lan, W. Qian, B. Guo and Zhang Laboratory (Institute of Zoology (IOZ), Chinese Academy of Sciences) members for the helpful discussions about the project. We thank Y. Fu, Y.

Ding, G. Gao and Z. Wang for the computational help. We thank J. Wang and X. Huang for experimental help. This research was supported by the National Key R&D Program of China (2018YFC1406902 and 2019YFA0802600), the Chinese Academy of Sciences (ZDBS-LY-SM005, XBZG-ZDSYS-201913 and XDPB17), the National Natural Science Foundation of China (31771410 and 31970565) and the Open Research Program of the Chinese Institute for Brain Research; all were granted to Y.E.Z. 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. conceived and designed the study. D.Z. performed the experimental analyses with the help of J.W.H. and Y.Q.Z. D.Z. and L.L. performed the computational analyses with the help of C.Y.C., H.Y., C.Y.M. and H.C. Y.E.Z., D.Z. and L.L. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41559-021-01614-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41559-021-01614-w>.

Correspondence and requests for materials should be addressed to Yong E. Zhang.

Peer review information *Nature Ecology & Evolution* thanks Ben-Yang Liao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

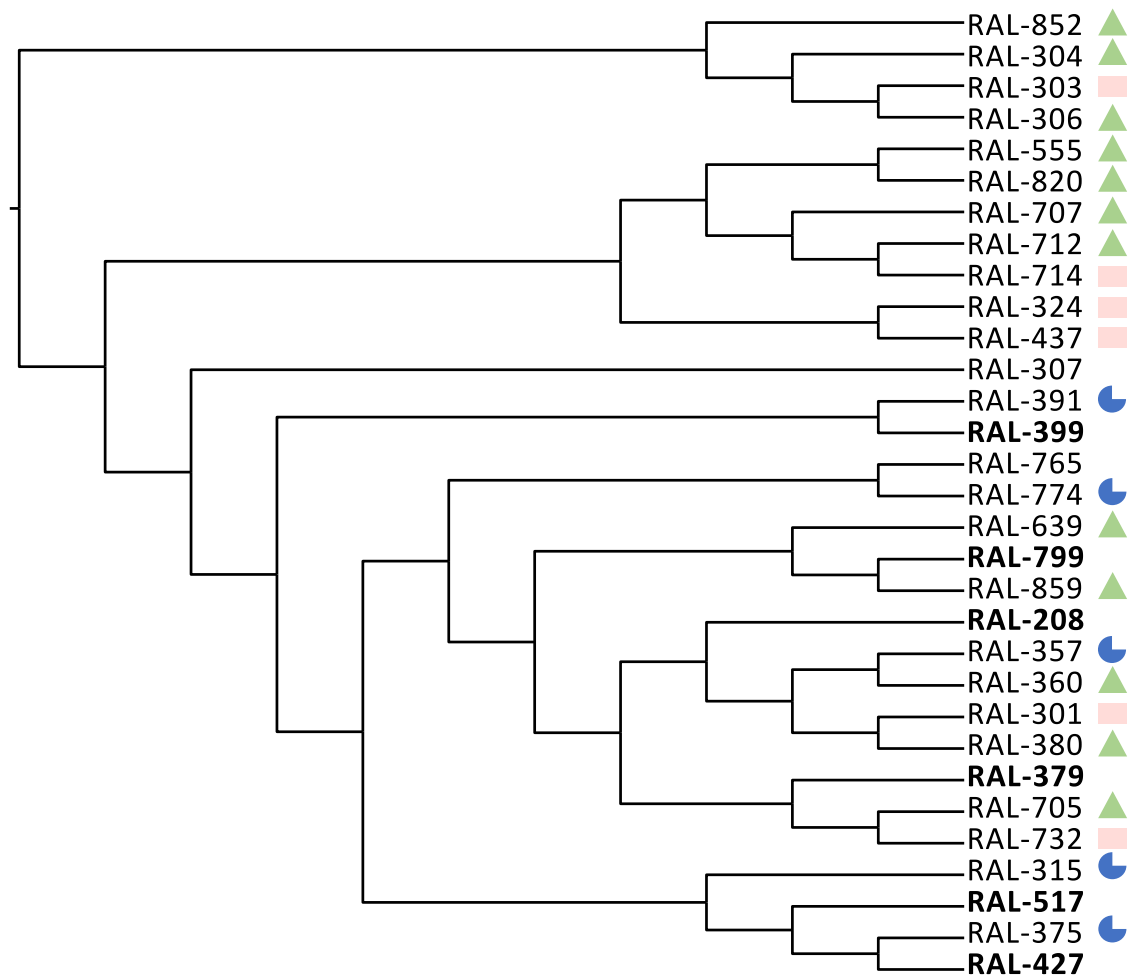
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

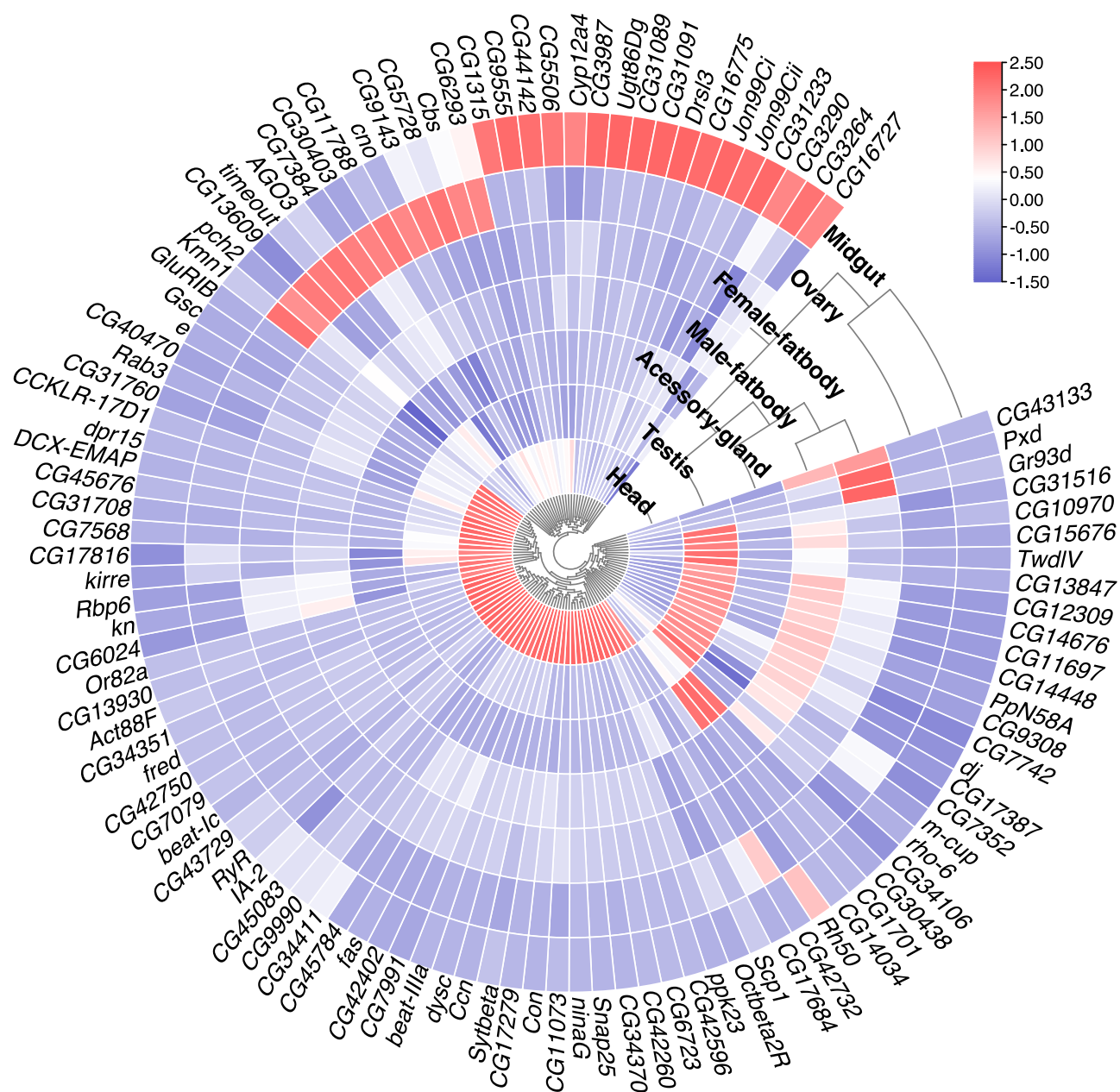
▲ Lines infected by *Wolbachia*

■ Lines harboring inversion

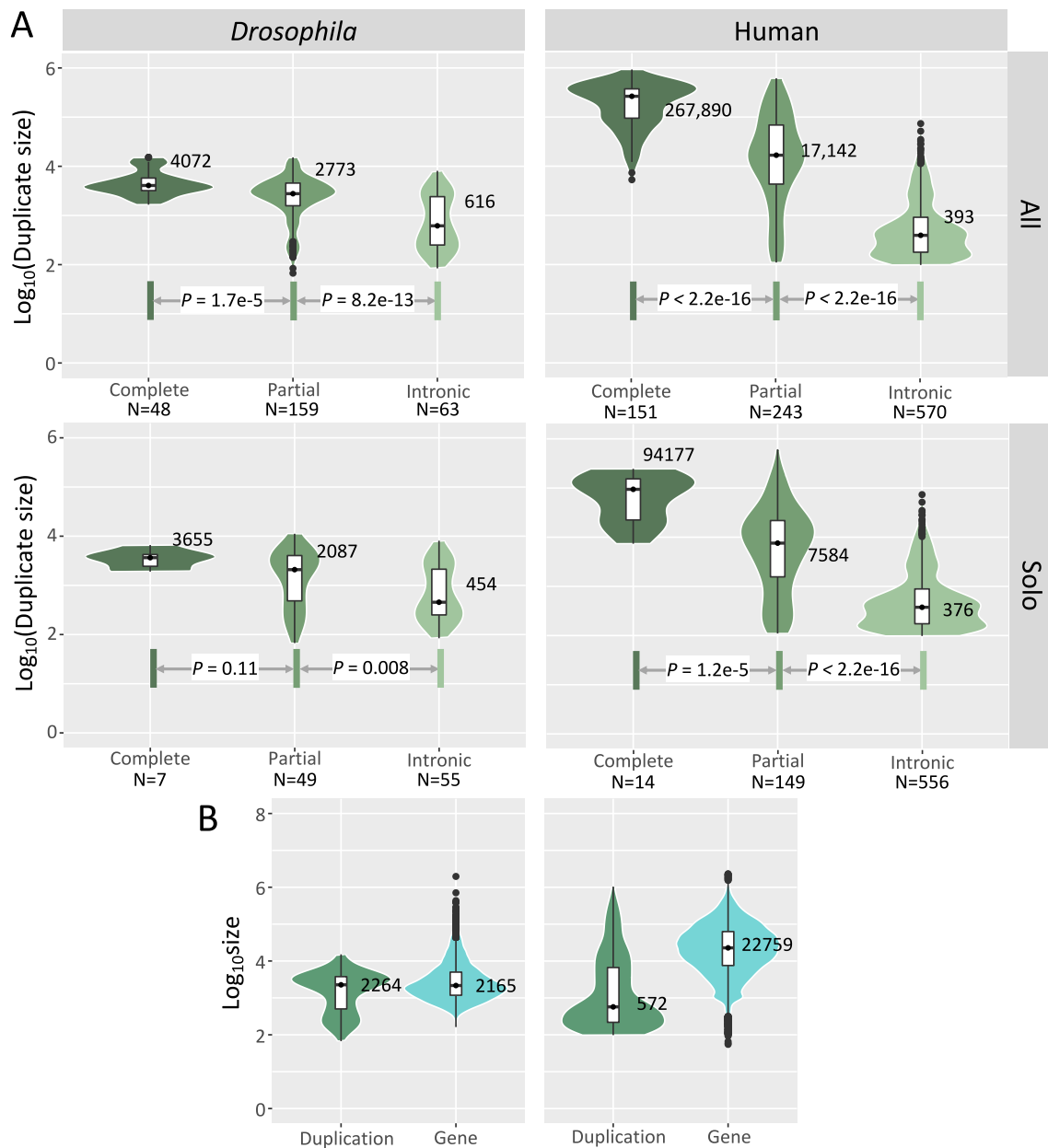
● Lines with the fraction of segregating variants in at least one chromosome arm higher than 2%



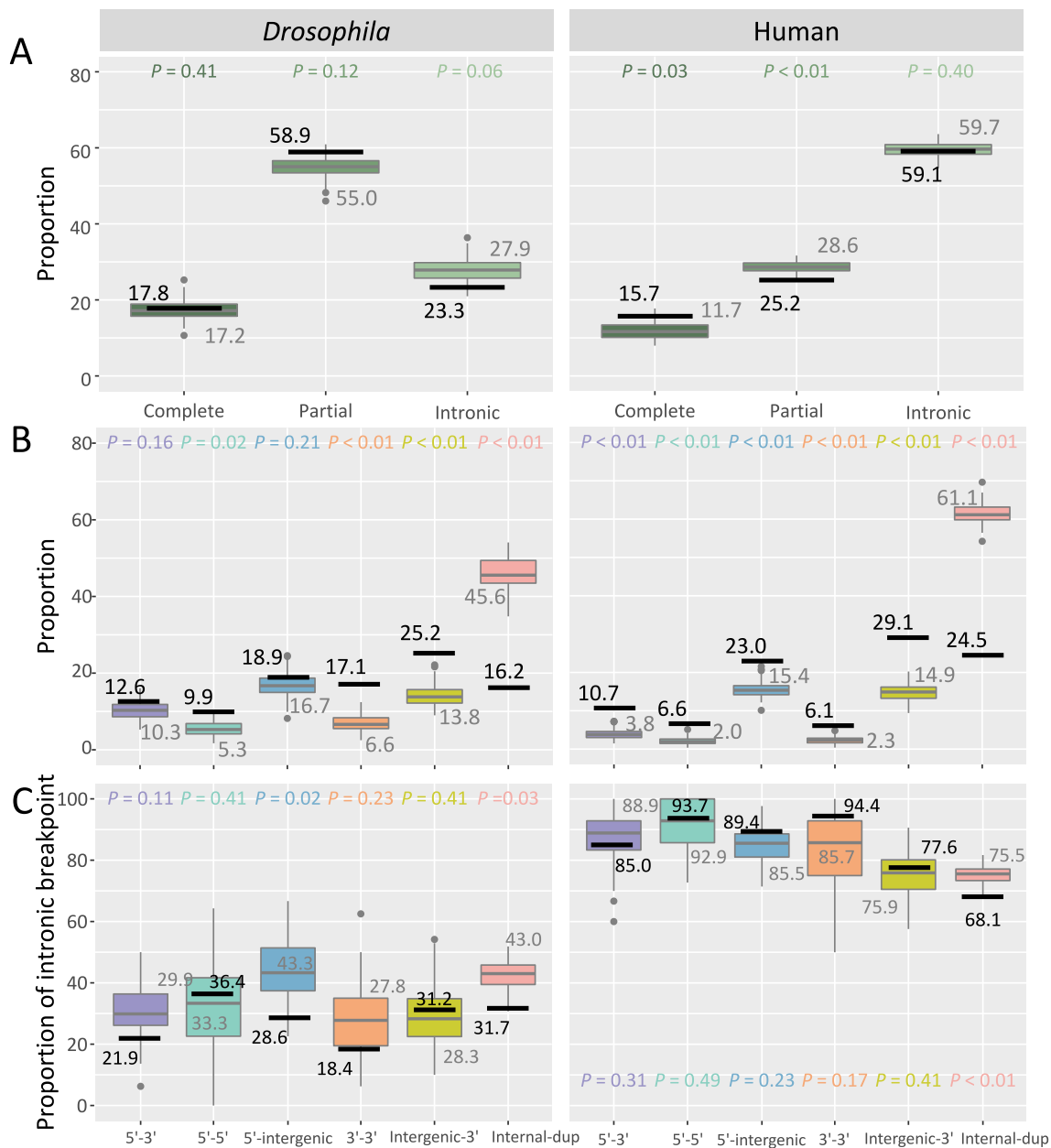
Extended Data Fig. 1 | Phylogeny and features of 31 DGRP core lines. We ordered 40 DGRP core lines from the Bloomington Stock Center but could only maintain 31 of these lines at the moment of sequencing. Lines infected with *Wolbachia*, lines harbouring inversions and lines with a fraction of segregating variants in at least one chromosome arm higher than 2% are marked in green, pink and blue, respectively. Among the remaining lines, six less related lines (marked in bold black) were subsequently used.



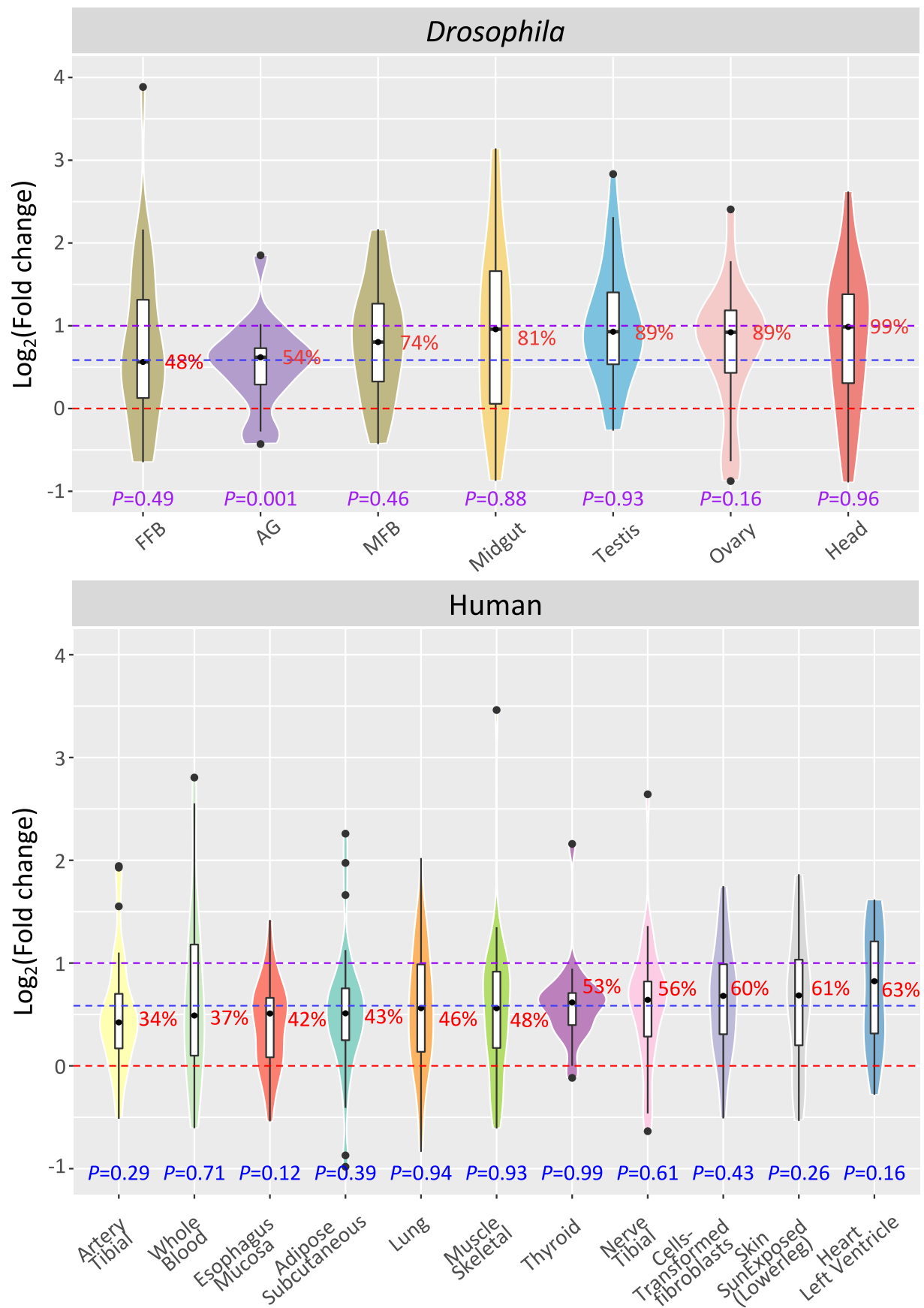
Extended Data Fig. 2 | Heatmap of 108 tissue-specific genes. The pheatmap package was used to generate this figure, in which expression was rescaled for each gene to highlight the tissue-specificity.



Extended Data Fig. 3 | Size distribution of different types of duplicates and protein-coding genes. a) All (top) duplications and the duplications spanning a single gene (bottom). **b)** Length distribution of all duplications relative to the length of all protein-coding genes. The median values are shown with each group. P values were calculated with Wilcoxon signed rank tests.

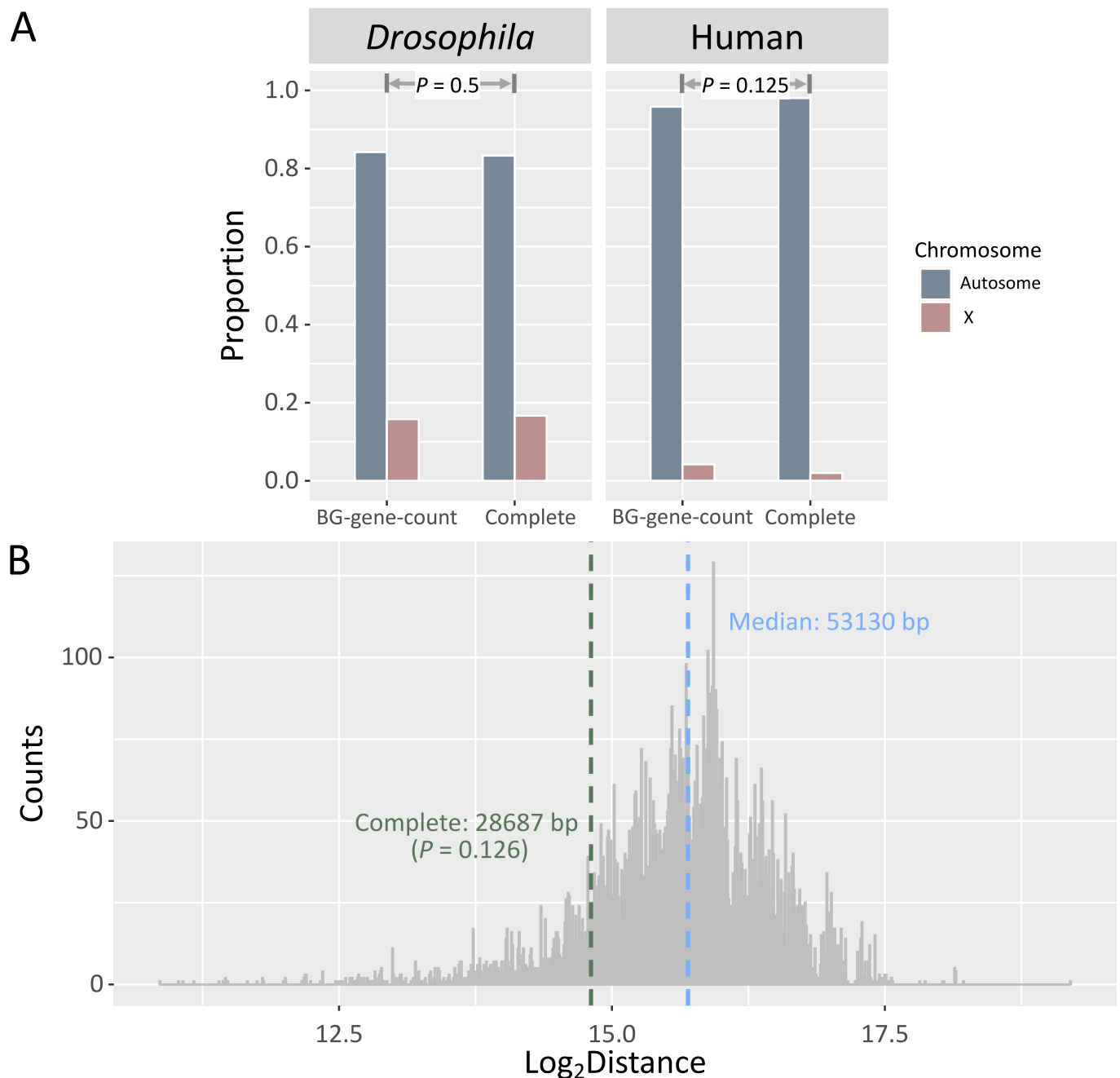


Extended Data Fig. 4 | Proportion distribution in the shuffling dataset. a) Proportion of complete/partial/intronic duplicates. **b)** Proportion of six arrangements of partial duplicates. **c)** Proportion of intronic breakpoints. We performed 100 simulations. For each simulation, we calculated the proportion of complete, partial and intronic duplicates. For partial duplicates, we further calculated the proportion of six subtypes. For each subtype, we also calculated the proportion of intronic breakpoints. We summarized 100 values as a boxplot. The median simulated values and observed values are shown along each group. For each plot, we marked the observed value as a dark black line and showed the empirical P value as the times out of 100 replicates where the observed value is more extreme. For example, in the case of partial duplicates, the observed proportion is 58.9%, which is only lower than that in 12 out of 100 random replicates (top left panel). The corresponding P value would be 0.12.

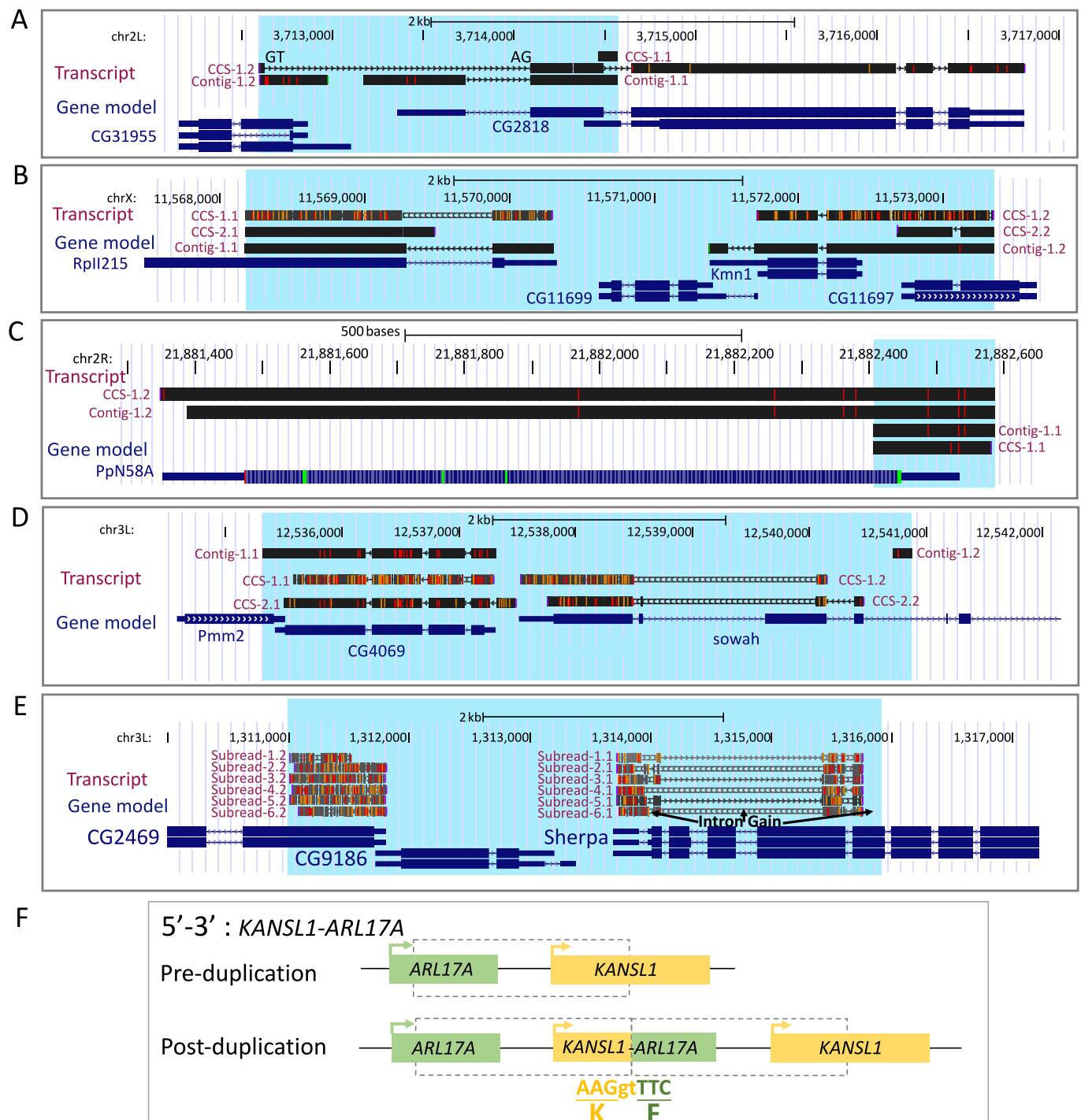


Extended Data Fig. 5 | See next page for caption.






















Extended Data Fig. 5 | Tissue-level transcriptional fold-change distribution for complete duplicates. The convention of this figure follows Fig. 3a. For each gene in each tissue, we calculated the fold-change with the median expression in duplication-present and duplication-absent lines or individuals. In flies, each tissue at least covers 19 transcribed (TPM > 1) genes. In humans, only 11 tissues or cell lines covering at least 15 transcribed genes were shown. Wilcoxon signed rank test was used to estimate the statistical significance relative to the expected 100% upregulation (top panel) and 50% upregulation (bottom panel). Despite some moderate fluctuation, the extent of upregulation generally does not deviate from the expectation (for 6 out of 7 tissues in flies, 11 out of 11 in humans). FFB, MFB and AG refer to female fat bodies, male fat bodies and accessory glands. Note that for AG, the deviation is significantly different from 100% simply because of the distribution of data most of which are smaller than 100% (purple curve). By contrast, FFB is not significant despite the smaller median (48% versus 54%).



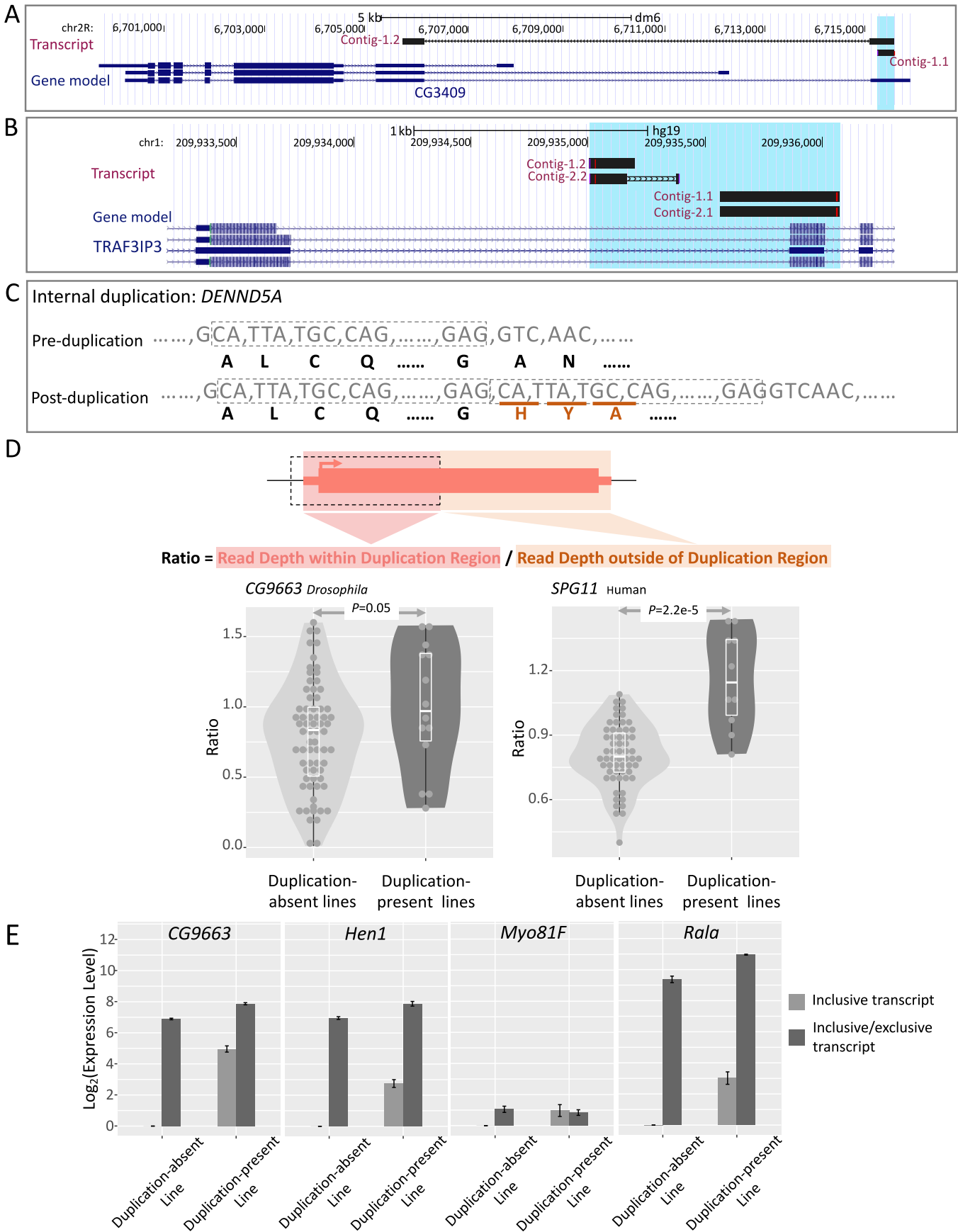
Extended Data Fig. 6 | The distribution of duplicates across different chromosomes and distribution of distances to HASs. a) Chromosomal distribution of duplicates. We only included the major autosomes (*Drosophila*: 2L, 2R, 3L, 3R; Human: 1-22) and the X chromosome in the calculation. The proportion of annotated coding genes on each chromosome in the reference genome is used as the control (BG or background). The P was calculated with the one-sided proportion test. **b)** distances to HASs. The median distance (28,687 bp) of 8 complete duplicates relative to the closest HAS was highlighted with a green line. We generated 10,000 random samples and calculated the empirical P as the percentage of how many times random samples showed a shorter median distance. The blue line indicates the median value of random samples. The two median values are also shown.



Extended Data Fig. 7 | Five fusion transcripts supported by PacBio reads and example of frameshift in *KANSL1-ARL17A*. **a)** *CG2818-CG31955* fusion. **b)** *Rpl1215-CG11697-Kmn1* fusion. **c)** 5'-Intergenic duplication of *PpN58A*. **d)** *CG4069-Pmm2-sowah* fusion. **e)** Subreads supporting three new introns of *Sherpa-CG2469*. **f)** *KANSL1-ARL17A* fusion. For Panels A to E, five out of 11 chimeric genes assembled in the testis samples of RAL-379 were confirmed with PacBio data. For A) to D), only the longest assembled contig was shown for simplicity. In A) to E), the conventions follow those in Fig. 4c. Similar to Fig. 4c, CCS reads sometimes encode novel exon/intron structures. Only the intron in the *CG2818-CG31955* fusion transcript harbours the standard splicing signal (GT-AG, Panel A). Six subreads in Panel E consistently exhibit three novel introns, as shown in Fig. 4c. Note that different CCS reads have different sequence qualities, as determined by the number of low-quality supporting subreads (for example, Panel E). For example, CCS-2 has a higher quality than CCS-1 in Panel B. In Panel F, the duplicated region is framed with dashed lines. The fusion transcript involves a 2-bp intronic sequence (the splicing site, 'gt') causing the frameshift. Two neighbouring codons are also shown.

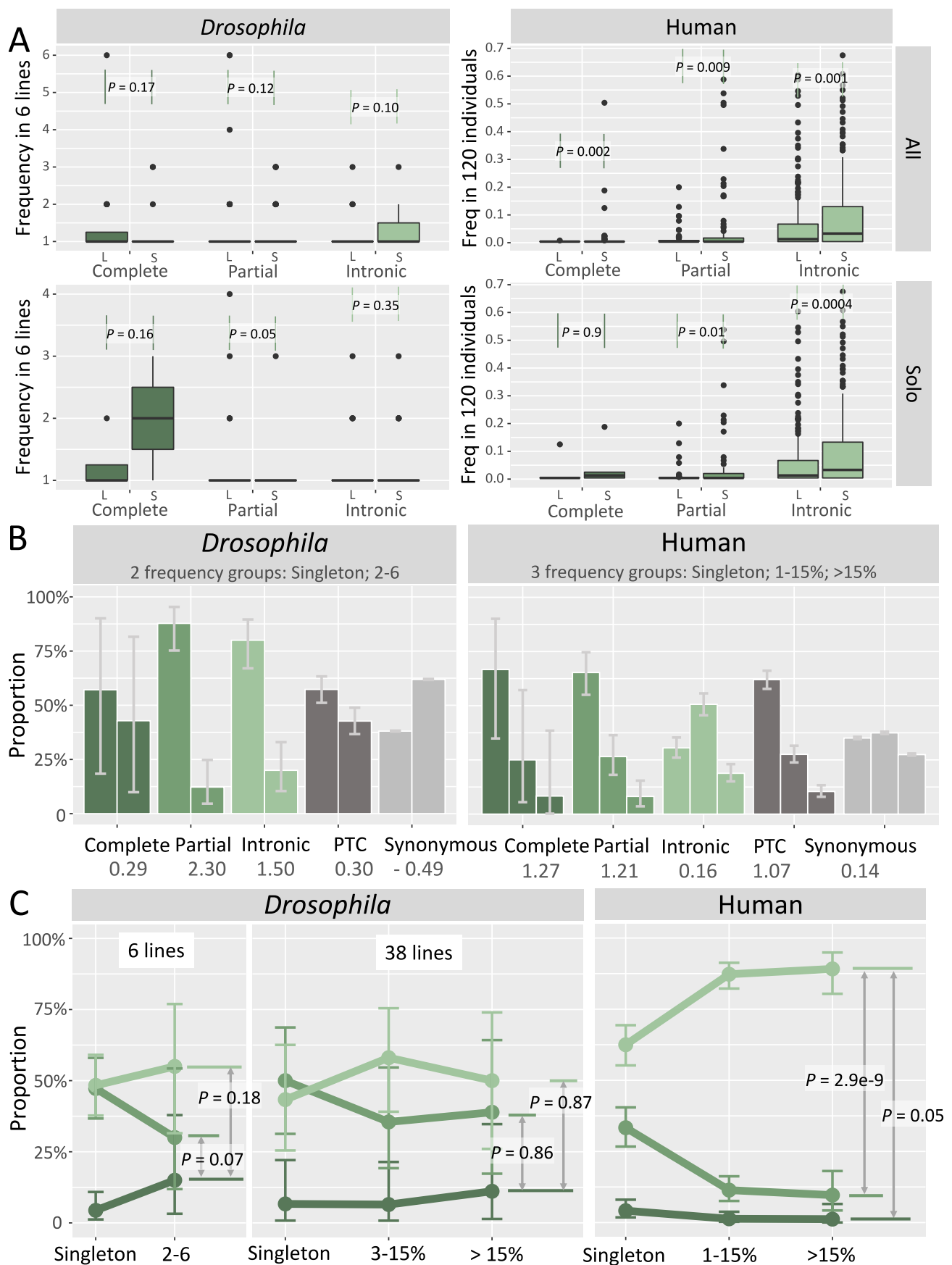
Species	Gene	Breakpoint	Combination	Frame	Domain rearrangement
<i>Drosophila</i>	<i>CG17387-ctrip</i>	Exon, exon	CDS-CDS	In-frame	
	<i>Sherpa-CG2469</i>	Exon, exon	CDS-CDS	In-frame	
	<i>Ugt86Dg-Ugt86De</i>	Exon, exon	CDS-CDS	In-frame	
	<i>CG14749-Vps25</i>	Exon, exon	CDS-CDS	Out-of-frame	
	<i>CG17322-CG17323</i>	Exon, exon	CDS-CDS	Out-of-frame	
	<i>Or82a-TwdIV</i>	Exon, exon	CDS-CDS	Out-of-frame	
	<i>CG7352-DppIII</i>	Exon, intron	CDS-CDS	Out-of-frame	
	<i>pch2-mRpS18A</i>	Exon, intron	CDS-CDS	Out-of-frame	
	<i>CG7384-Fatp1</i>	Exon, 3'UTR	CDS-3'UTR	NA	
	<i>CG7402-CG7408</i>	Intron ^U , intron	5'UTR-CDS	NA	
	<i>CG42321-cnn</i>	5'UTR, exon	5'UTR-CDS	NA	
	<i>Prosbeta5R2-CG5681*</i>	3'UTR, exon	Gene-CDS	NA	
Human	<i>COMMD10-AP3S1</i>	Intron, intron	CDS-CDS	In-frame	
	<i>HIP1-CCL26</i>	Intron, intron	CDS-CDS	In-frame	
	<i>P4HTM-ARIH2</i>	Intron, intron	CDS-CDS	In-frame	
	<i>TFG-ADGRG7</i>	Intron, intron	CDS-CDS	In-frame	
	<i>TFDP2-XRN1</i>	Intron, intron	CDS-CDS	In-frame	
	<i>ACSS1-NINL</i>	Intron, intron	CDS-CDS	Out-of-frame	
	<i>ARRB1-GDPD5</i>	Intron, intron	CDS-CDS	Out-of-frame	
	<i>KANSL1-ARL17A</i>	Intron, intron	CDS-CDS	Out-of-frame	
	<i>ADNP2-RBFADN</i>	Intron ^U , intron ^U	5'UTR-3'UTR	NA	

Extended Data Fig. 8 | Sequence features of 5'-3' fusion genes. CDS refers to coding sequence. Whether a frameshift occurs was inferred based on the codons adjacent to the breakpoints. Domains are shown by blue (the leading gene) or red (the lagging gene) rectangles, and zigzags show mid-domain breaks. Note: the fused region of the leading gene for *CG17387-ctrip*, that of the lagging gene for *Sherpa-CG2469* and both regions for *TFDP2-XRN1* do not harbour annotated domains. The lengths of the rectangles are roughly proportional to the relative domain size within each gene. 'Intron^U' and 'intron' refer to introns located within UTRs and introns within coding regions, respectively. For *Prosbeta5R2-CG5681*, *Prosbeta5R2* was completely duplicated, whereas the 3' part of *CG5681* was duplicated. After fusion, *Prosbeta5R2* harbours a longer 3' UTR by incorporating a *CG5681*-derived sequence.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Sequence and transcription features of internally duplicated genes. **a)** Partial duplication of the 5' UTR in *CG3409*. *CG3409* encodes different 5' UTRs, one of which is duplicated and a longer 5' UTR was generated. Because this 5' exon is alternative the original coding sequence is presumably not affected. **b)** Post-duplication intron retention of *TRAF3IP3*. According to the duplication boundaries (highlighted in light blue), two contigs indicate inclusive transcripts where exons and flanking introns are simultaneously transcribed. **c)** Out-of-frame transcript of *DENND5A*. The duplicated region is framed with dashed line boxes. Codons are separated with commas, and amino acids are shown accordingly. The incompatibility of codon phases causes the frameshift. **d)** Read depth-based quantification of isoforms (see also Methods). Two examples (*CG9663*, *SPG11*) are shown. Given only 10 data points for the duplication-present lines of *SPG11*, we showed all specific values as grey dots. For each gene, we calculated the ratio of the read depth (duplicated region versus flanking region) in lines with or without duplication. For these two examples, the ratio found for lines with duplications was significantly (one-sided Wilcoxon signed rank test $P \leq 0.05$) higher than that in lines without duplications, supporting the presence of inclusive isoforms. According to the median ratios, we further calculated the relative expression ratio of inclusive and exclusive transcripts (Fig. 5f). **e)** qRT-PCR-based quantification of isoforms in flies. Whole-body samples of duplication-absent lines were used as controls. Primers targeting inclusive transcripts did not show signals in the duplication-absent lines, and primers targeting both inclusive and exclusive transcripts generally generated weaker signals than their counterparts in the duplication-present lines. The error bar represents the standard deviation based on three technical replicates.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | The frequency distribution of duplicates classified by the median size and proportions of duplicates across different allele frequency groups. **a)** The upper and lower panels show all duplications and the duplications spanning a single gene, respectively. For each type of duplicates, we divided them into large (L) and small (S) groups according to the median size. The Wilcoxon signed rank test was used to compare the median frequencies in these two groups. Small duplicates generally show higher frequencies than large duplicates in humans. Moreover, despite the small sample size of fly dataset, larger duplications also seem to show lower allele frequencies. **b)** Allele frequency spectra of complete, partial and intronic duplicates in addition to PTCs and synonymous substitutions in *Drosophila* (the left panel) and humans (the right panel). Only duplications spanning a single gene were plotted. **c)** Proportion of complete, partial and intronic duplicates within different frequency groups. In the left panel, all duplicate loci were plotted with six-line allele frequency data. In the middle and right panels, only duplications spanning a single gene were plotted. The figure conventions of Panel B and C follow those of Fig. 6a and b, respectively.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Mx3000p was used to process qRT-PCR data.

Data analysis We used public softwares including NovoAlign v3.03, Delly v0.7.2, CNVnator v0.3, Breakdancer v1.2.6, Pindel v0.2.5.a8, BWA mem v0.7.10, Picard v1.119, GATK v3.5, SnpEff v4.3, STAR v2.5, RSEM v1.3, Trinity v2.6.5, BLAT v35, SAMtools v1.3, Hisat2 v2.1.0, vcftools v0.1.16, bedtools v2.15.0 and NCBI ORFfinder web server. Additionally, in-house codes developed for this study are available at both <https://github.com/Zhanglab-IOZ/Polymorphic-Duplication/tree/main/Code> and <https://sandbox.zenodo.org/record/946570>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The resequencing data, RNA-seq data and Iso-seq data generated in this study were concurrently submitted to the NCBI BioProject database under accession numbers PRJNA681089, PRJNA681417 and PRJNA693662, and the Genome Sequence Archive under accession numbers PRJCA001789, PRJCA004186 and PRJCA004319, respectively. The processed fly data (duplication loci, DNA read alignment depth, RNA-seq alignments, assembled chimeras, PacBio CCS reads and RT-PCR Sanger sequencing data) and human data (duplication loci and assembled chimeras) were uploaded to the UCSC genome browser as public sessions "http://

genome.ucsc.edu/cgi-bin/hgPublicSessions" with names as "Drosophila Duplication" and "Human Duplication", respectively. The assembled contigs and the RT-PCR Sanger sequencing files were co-submitted to <https://github.com/Zhanglab-IOZ/Polymorphic-Duplication/tree/main/Sequence> and <https://sandbox.zenodo.org/record/946570>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In flies, we resequenced 6 phylogenetically representative DGRP lines and identified 270 polymorphic duplicate genes. In humans, we downloaded resequencing data from 145 individuals and identified 964 polymorphic duplicate genes. For flies, we generated RNA-sequencing (RNA-seq) data for 7 somatic or germline tissues for each line. For humans, we downloaded RNA-seq data for a median of 9 tissues in each individual. The reasonably big sample size of duplicate genes and high diversity of transcriptome data enable our downstream analyses.
Data exclusions	The rationale for our data processing has been described. For example, we confirmed 52 out of 66 (78.8%) cases for which RT-PCR primers flanking breakpoints could be designed (Supplementary Table 10, Methods).
Replication	Two biological replicates were used in RNA-seq. Three technical replicates were used for qPCR experiments.
Randomization	Not applicable.
Blinding	The staffs were not blinded during data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems	Methods
n/a	n/a
<input checked="" type="checkbox"/> <input type="checkbox"/> Involved in the study	<input checked="" type="checkbox"/> <input type="checkbox"/> Involved in the study
<input checked="" type="checkbox"/> <input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/> <input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/> <input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/> <input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/> <input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/> <input type="checkbox"/> MRI-based neuroimaging
<input type="checkbox"/> <input checked="" type="checkbox"/> Animals and other organisms	
<input checked="" type="checkbox"/> <input type="checkbox"/> Human research participants	
<input checked="" type="checkbox"/> <input type="checkbox"/> Clinical data	
<input checked="" type="checkbox"/> <input type="checkbox"/> Dual use research of concern	

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	D. melanogaster Genetic Reference Panel (DGRP) were obtained from the Bloomington Drosophila Stock Center. All of these strains were maintained on standard food at 25°C and 60% humidity in a 12h:12h light/dark cycle.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	No ethical approval or guidance was required since we did not perform relevant experiments.

Note that full information on the approval of the study protocol must also be provided in the manuscript.