bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GageTracker: a tool for dating gene age by micro- and macro-synteny with high speed and accuracy

Chengchi Fang[1,†], Chuan Dong[2,†], Cheng Wang[1,3], Fan Xiong[1], Suxiang Lu[1], Wenyu Fang[3], Tong Li[1,3], Xiaoni Gan[1], Liandong Yang[1], Honghui Zeng [1,*], Shunping He[1, 4,*]

[1] The Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei, 430072, China

[2] State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Hangzhou, Zhejiang, 311300, China

[3] University of Chinese Academy of Sciences, Beijing, 100039, China

[4] State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei, 430072, China

† The authors would like to make it clear that, in their view, the first two authors should be considered as CO-FIRST AUTHORS.

* To whom correspondence should be addressed. Email: clad@ihb.ac.cn, and zhh@ihb.ac.cn.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

With the advent of the Earth Genome Project, an increasing number of species' genomes presents exciting opportunities for exploring genetic and phenotypic diversity in organisms. Determining the origin time of genes facilitates the elucidation of crucial genetic mechanisms underlying significant biological evolutionary questions such as the transition from aquatic to terrestrial life, the emergence of mammals, the origin of humans, as well as the development of species- or lineage-specific traits. However, accurately determining the origin time of these genes in species separated by long evolutionary distances remains a major challenge in bioinformatics as these genes often undergo significant changes in their genome sequences, making it difficult to trace them back to their origin. Here, we proposed a new approach for dating gene age based on the micro- and macro-synteny algorithms. This approach employs the parallel computation of orthologous genome alignments across multiple species. Our method was integrated into the GageTracker (Gene Age Tracker) software, providing a fast and accurate way to trace gene age with minimal user input, available at https://github.com/RiversDong/GageTracker. Benchmarked against the simMammals dataset (Alignathon), GageTracker achieved the same high-quality genome alignments as the optimized LastZ aligner, but improved operation speed by 1.4-7 times. In a separate analysis of 12 *Drosophila* genomes, GageTracker efficiently assessed the ages of 23,720 genes (including ~13,965 protein-coding genes) in just ~22 hours with default parameters. When comparing with the GenTree database (recognized as the most comprehensive and accurate tool for evaluating gene age), GageTracker achieved an impressive ~94.4% accuracy and ~99% macro consistency in assessing the age of protein-coding genes. Moreover, for the ~5.6% conflicting genes, GageTracker displayed slightly higher support rates than GenTree, as evidenced by data from OrthoDB, FlyBase, and Ensembl ortholog databases. Notably, younger genes identified by GageTracker exhibited a preferential expression pattern in the testis, further reinforcing the reliability of GageTracker in accurately tracing gene age.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Since Darwin's era, biologists have been keenly focused on the question of how organisms evolved from a common ancestor to the diverse species that we see today. Genomic variation has been a major contributor to this vast biological diversity (1). With the advent of the Earth Genome Project, including the Bird 10K Genomes, Fish 10K Genomes, and Protist 10K Genomes, the analysis of an increasing number of species' genomes presents exciting opportunities for exploring genetic and phenotypic diversity in organisms. Determining the origin time of genes facilitates the elucidation of crucial genetic mechanisms underlying the development of species- or lineage-specific traits. Researchers have discovered that even closely related species exhibit a certain degree of variability in genome size and gene number (2). The abundance of available genomes provides a valuable opportunity to answer several central questions in evolution, especially when it comes to comprehending the evolution of new genes. Approximately 380 new functional genes related to brain development evolved in a relatively short period after the divergence of humans and chimpanzees (~6 million years ago). These genes played a crucial role in enhancing human cognition and helping to shed light on what makes humans as human being (3-5). The origin of novel genes has been considered to be a major contributor to adaptive innovation (1,6-8). They often result in the development of new adaptive traits and have a profound impact on resolving sexual conflicts, embryonic development, and reproduction (9-11). In flies, the origin of new genes has provided a new genetic basis to species and they are frequently highly expressed in reproductive organs, such as *jingwei* and *sphinx* (11,12). The roles of new genes that could be involved in the important biological process were also demonstrated by a series of high-throughput RNA silencing (13) and CRISPR-Cas9 experimental technology (14). In plants, new genes have been shown to play important roles in various biological processes, such as male sterility and plant height (15,16). It was also found that fusion genes could regulate phenotype traits including seed germination, shoot length, and root length (17). Furthermore, Zhang et al. discovered that new protein-coding genes could emerge from non-coding DNA sequences, providing a source for additional diversity in protein composition (18). The new genes in bamboo were related to the origin and evolution of its unique character of rapid growth (19). In prokaryotic species, the evolution of new genes, which was shaped by the GC content, expression level, as well as the essentiality of a gene, were important factors that could influence the evolutionary rate (20,21). Together, more and more studies have shown that new genes were one of

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the main driving forces of evolution and innovation, and promote organisms to produce special phenotypic and physiological characteristics. Therefore, the annotation of new genes contributed greatly to the understanding of mechanisms underlying these important biological functions.

However, the progress in comprehending the impact of new genes on evolutionary traits and lineage-specific characteristics has been hindered by the absence of a rapid and precise approach to identifying orthologs and determining the origination age of genes. The annotation of gene age via comparative genomics has been primarily achieved through three strategies: 1) the homologous gene family-level dating strategy (22-24); 2) the synteny-based gene-level dating strategy (22,25-27); 3) the comprehensive orthologous database (4,24,28,29). The first strategy was significantly dependent on the quality of gene annotations in the reference species, thereby leading to potential false positives in the new gene catalog in cases where the annotation was absent in the reference species. The second strategy relied heavily on genome synteny analysis, which allowed for the avoidance of imperfect protein gene annotations and ensured high accuracy in detecting new genes. Nonetheless, this strategy had certain constraints in managing complex sequences such as genome sequencing gaps and low-complexity regions (e.g., telomeres), leading to potential inaccuracies. On the other hand, the third approach, encompassing GenTree, GenOrigin, and OrthoDB, typically employed a combination of methods to construct a searchable database for tracing gene age (4,24,30). However, these databases had limited coverage, given that they only included a few well-studied model species and could not be updated or expanded to include new species of interest, thereby impeding their utility, particularly in an era of rapid growth in genomic data.

The rapid accumulation of genomic data has created a challenge for gene age annotation. Thousands of genome assemblies are publicly available and additional species are being sequenced as part of projects, further exacerbating the challenge (31-33). This dense sampling and sequencing of the tree of life are set to provide a comprehensive understanding of evolution and biodiversity. With the rapid expansion of genomic data to fill out the phylogeny of vertebrates, there is going to be a strong desire to explore genetic innovation and phenotypic diversity of the different species, meanwhile, the growing data also provided a great opportunity for tracing gene origin time in more details. Thus, we proposed a new tool, GageTracker, for tracing gene age in the era of rapid expansion of genomic data. Based on the micro- and macro-synteny algorithm, GageTracker was a one-command running software to search ortholog genome alignments suitable for multiple species and

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

allow a fast and accurate trace gene age with minimal user inputs. It obtained a high alignment quality as the optimized LastZ software but significantly saved the running time as well. Compared with the GenTree database, it achieved ~94.4% accurate consistency and ~99% macro consistency in assessing the age of protein-coding genes. For the ~5.6% conflicting gene age, GageTracker also showed a slightly higher support rate from orthoDB, FlyBase, and Ensembl ortholog database than the Gentree database, suggesting its reliability for dating gene age.

MATERIAL AND METHODS

General Algorithm of GageTracker

The workflows of GageTracker were mainly divided into three parts: data inputs, genome alignment and processing, and age-infer (Figure 1A). The genome alignment behavior of GageTracker was specified by the configuration file and command line parameters. The Last v1229 program (34) was used to conduct genome alignment, and several toolkits from UCSC (mainly including maftochain, chainCleaner, chainSwap, netSyntenic, netChainSubset, chainStitchId, and chainNet) were employed to detect the Reciprocal Best Syntenic Net hits (RBH) alignments according to the description of RBH pipeline (doRecipBest.pl), which can be available at https://genomewiki.ucsc.edu/index.php?title=HowTo:_Syntenic_Net_or_Reciprocal_Best. We used the maf2synteny (35) and the bedtools (36) to construct the block regions and detect sequencing gaps (left panel of Figure 1B). Finally, GageTracker labels gene age according to a voting strategy (left and middle panel of Figure 1B).

Given a genome that has evolved through a steady-state process of gene acquisition and loss, we have employed a voting strategy to assess whether a gene originated on a particular branch (left and middle panel of Figure 1B). First, we conducted a gene presence analysis on each branch. Specifically, we inferred the presence of a gene within a branch if it had orthologous segments in any species within that branch. Subsequently, we systematically evaluated each evolutionary branch in chronological order, from the young to the old, and computed the support rate of the gene's origin on that branch. For example, Gene 1 is found to have orthologs on branches 6, 3, 2, and 0, indicating its presence on branches 6, 3, 2, and 0 but absence on branches 5, 4, and 1 (orange rectangle: presence,

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

blank rectangle: absence, left and middle panel of Figure 1B). Then, we will proceed by sequentially calculating the support rate on branches 6, 3, 2, and 0 (percentage in orange rectangles, left and middle panel of Figure 1B). When focusing on branch 6, Gene 1's support rate is 100%; when focusing on branch 3, its support rate is 50% (2/4, present in 2 branches (6, 3) and absent in 2 branches (5, 4)); when focusing on branch 2, its support rate is 60% (3/5, present in 3 branches (6, 3, 2) and absent in 2 branches (5, 4)); when focusing on branch 0, its support rate reaches 57% (4/7, present in 4 branches (6, 3, 2, 0), absent in 3 branches (5, 4, 1)). Therefore, Gene 1 is assigned to the oldest branch 0, as the support rate is above the voting threshold of 50%. Similarly, Gene 2 is assigned to the oldest branch 1. At the same time, Gene 3 appears on branches 6 and 0. But its support rate of originating on branch 0 (2/7, present in 2 branches, absent in other 5 branches) is only 28.5%, indicating that Gene 3 is more likely to have originated independently rather than being lost in all intermediate species. Thus, Gene 3 independently emerged on branch 6 as a newly originated gene.

In principle, this algorithm could well distinguish independent origin genes to avoid overestimating genes' age. The user could adjust the voting parameter in the control file (voting=0.5) according to the divergence time between species and their own needs.

The masked strategy for genomes

Soft masking of repeat sequences was a common preprocessing step in genome alignment and gene age estimation. GageTracker utilized a *de novo* prediction method for masking genomic repetitive sequences. It utilized WindowMasker to mask interspersed repeats sequences and Tantan to simple-sequence from scratch in genome sequences. In GageTracker, the repeat sequences were first softly masked. Because the protein-coding sequence contained the conserved genetic information between genes. To prevent masking such functional regions, we converted the incorrectly masked sequence to uppercase in the protein-coding regions.

Benchmark genomes alignment

To evaluate the performance of GageTracker, we included two types of benchmark genomes: simulated data and real data, respectively. The simulated mammal’s data was downloaded from the packageMammals in Alignathon database (37). It contained five simulated benchmarks (simHuman, simMouse, simRat, simCow, simDog) that consisted of genomes of different species, sizes, and

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

complexities. The real data sets, including *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster*, as well as their GTF annotation files, were retrieved from Ensembl (version 108). Other 11 *Drosophila* genomes were obtained from FlyBase (version FB2022_06). The detailed version number of each genome were listed in Supplementary Table S1.

The genome alignments were performed by GageTracker, Lastal, and PlastZ aligners respectively on the same computational platform using commands that listed in the following script in the calculation section. Each aligner was controlled in 6 threads on the servers with an E7-4850-CPU (2.5GHZ) array and 1 TB of memory.

- Only running genome alignments in GageTracker command line:

GageTracker example.ctl -step1

- Lastal commands line, which contains two steps: convert repeat sequence into uppercase, and perform genome alignment by Lastal

less mus.fasta|perl -e '{while(<>){if(/^>/){print;}else{$_=uc($_); print}}}'> mus.up.fasta
less human.fasta|perl -e '{while(<>){if(/^>/){print;}else{$_=uc($_); print}}}'> human.up.fasta
lastdb -c -P6 -R10 -uMAM8 humandb ./human.up.fasta
lastal -P6 -C2 -u0 -m50 -p HOXD70 ./humandb ./mus. up.fasta

- Perform genome alignment by Lastz

python3.7 PLastZ.py mus.fasta Human.fasta ./output -lo="O=400 E=30 L=3000 H=2200 T=1 --ambiguous=iupac --format=lav " -p 6

Evaluation of alignments

Genome alignment results often yield many-to-many matches due to the prevalence of gene duplication as a major driver of new gene birth. However, we focused specifically on identifying one-to-one orthologous alignments. To obtain our targeted regions, the many-to-many alignments generated by the aligners were fed into a similar doRecipBest.pl pipeline from UCSC, which mainly included mafToPsl, pslToChain, chainCleaner, chainStitchId, chainSwap, chainNet, netSyntenic, netChainSubset, chainStitchId, chainSwap, chainNet, and netToAxt toolkits. This could create the query-reference reciprocal best chains and nets files (refer to one-to-one orthologous alignments in this work).

Second, we processed the ancestor.noparalogies.maf data in the benchmark alignment file of the Alignathon database by mafTools pipeline (37). It mainly contained: i) The mafFilter to extract alignments that only match the query-reference species; and ii) The mafTransitiveClosure to perform the transitive closure on these query-reference alignments. This could produce four groups (simCow-simHuman, simDog-simHuman, simMouse-simHuman, simRat-simHuman) of query-reference

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

194 genome alignments and the simHuman was the reference genome. These four groups' genome
195 alignments were considered one-to-one truth alignments in our work.
196

$$\text{Precision} = \frac{|A \cap T|}{|A|},\ \text{Recall} = \frac{|A \cap T|}{|T|},\ \text{F} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1)$$

$$\text{OP} = \frac{OG}{Min(TG\ in\ GagTracker,\ TG\ in\ GenTree)} \quad (2)$$

197 Finally, we used the mafComparator to compare the test alignments to these benchmark
198 alignments and calculated the number of identical base pairs. To assess the quality of the three
199 aligners (Lastal, GageTracker, and LastZ), we calculated recall, precision, and balanced *F*-score
200 values in the usual way as equation (1). Let T denote the number of base pairs in one-to-one truth
201 alignments, and A denote the number of base pairs in one-to-one test alignments. Then the precision
202 was the ratio of test positive predictions (A∩T) to the total number of test alignments (A); recall was
203 the ratio of test positive predictions (A∩T) to the total number of truth alignments (T); and *F*-score
204 was the harmonic mean of recall and precision. It is the trade-off between precision and recall and
205 provides a single metric to evaluate the overall performance of a classification model.
206

207 **Age comparison in three different levels**

208 We used the dating gene age of *D. melanogaster* as a case to compare the performance between
209 GageTracker and GenTree at three different levels: the global level, the nearly-detailed level, and the
210 detailed level. More specifically, the first level was used to measure the global consistency with
211 GenTree (only focusing on the proportion of the genes with the same gene age). In the nearly-detailed
212 level, we divided genes into two categories young genes (from branch 1 to branch 6) and old genes
213 (genes in branch 0), and calculated the **o**verlapping **p**roportion (OP) of the two annotated age versions
214 (GenTree and GageTracker). The overlapping proportion was calculated by equation (2), where OG
215 represents the number of **o**verlapping **g**enes in the same comparison level and TG was the **t**otal **g**enes
216 in the current comparison level.

217 In addition, we estimated the macro-consistency between GageTracker and GenTree by
218 comparing the branch distance of the same gene ($\text{branch distance} = |\text{Age in GageTracker} -$
219 $\text{Age in Gentree}|$). There are seven branches in the evolutionary lineage in our case study, and half of
220 them are three. We considered a branch distance of ≤ 3 to indicate a short distance or macro-

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consistency (less than half of the total 7 branches), whereas a branch distance of > 3 was regarded as indicative of inconsistency in the gene age.

The comparison of conflicting genes between GenTree and GageTracker

To compare the accuracy of conflicting genes between GageTracker and GenTree, we used a similar method proposed by a previous study (14). This method involved the support evidence from three ortholog databases: OrthoDB, FlyBase, and Ensembl. First, we retrieved the orthologous gene clusters of *Drosophila* (level 7214 and species 7214 in December 2022) from OrthoDB by our self-written crawler script (https://github.com/RiversDong/tools/blob/main/download_data_orthoDB.py), FlyBase (https://ftp.flybase.org/releases/FB2022_01/precomputed_files/orthologs/), and Ensembl biomart (https://dec2021-metazoa.ensembl.org/biomart/martview/). Subsequently, we re-inferred the age of conflicting genes (between GageTracker and GenTree) using OrthoDB, FlyBase, and Ensembl, respectively with the same voting cutoff that was employed by GageTracker (voting=0.5). Afterward, we counted the number of consistency genes for differential gene classes and calculated the support ratio based on the branch distance to estimate the consistency between GageTracker and GenTree. The difference value here refers to the absolute age difference between genes. For instance, if GageTracker places a gene at branch 6, while GenTree estimates its age as branch 5, then the difference in age between the two is 1 ($|\text{Age in GageTracker} - \text{Age in Gentree}|$). A smaller difference value indicates greater consistency between the annotation results.

Expression data of *D. melanogaster*

FlyAtlas 2.0 (https://motif.mvls.gla.ac.uk/FlyAtlas2/) was an online database that provided expression data for male, female, and larval stages of *D. melanogaster* (38). To investigate the expression patterns of genes with conflicting gene ages, we downloaded (in May 2023) gene expression data for 15 male tissues and 16 female tissues from FlyAtlas 2.0. Specifically, we focused on two groups of genes with conflicting age annotations, which we labeled as group A and group B. Group A comprised genes annotated as belonging to branch 0 by GageTracker but annotated as new genes (ranging from branch 1 to branch 6) by GenTree, while group B included genes annotated as old genes by GenTree but as new genes by GageTracker.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

Features and implementation of GageTracker

GageTracker was an available software that employed a single command line execution to estimate gene age. Figure 1A provided a visual representation of the software's architecture. The input data required by GageTracker was only a GTF annotation of the focal species and the FASTA files for each genome assembly. The primary objective of this algorithm was to determine whether a gene in the focal species had orthologous counterparts in the outgroup species (middle panel of Figure 1B). The algorithm aimed to identify a set of optimal symmetric alignments (micro-synteny) between the focal and query genomes and merge these alignments into larger synteny block regions (macro-synteny). The presence of a gene in the query species was indicated if it overlapped with a micro-synteny region in both the focal and query genomes. Conversely, the absence of the gene in the outgroup species was inferred if it failed to overlap with any micro-synteny region in the outgroup genome. To minimize false positives in complex regions such as sequence gaps, centromeres, and telomeres, the gene age estimation was limited to genes within macro-synteny blocks.

In summary, the GageTracker software consisted of five key processes:

1) Soft masking of repeat sequences in both the reference and query genomes using either the WindowMasker or Tantan software (41,42), then converting the exon regions to uppercase if they have been masked by WindowMasker or Tantan.

2) Identification of a set of local pairwise alignments between the focal and query genomes through the seed-and-extend heuristic with the Last aligner (seeds were excluded from overlapping masked repeat sequences, but the final alignments were allowed to extend into them).

3) Retrieval of Reciprocal Best Syntenic Net hits (RBH) using the UCSC Genome Browser toolkit and conducting micro-synteny analysis between the two genomes.

4) Construction of genome alignment blocks, conducting macro-synteny analysis, and removal of genes in genome alignment gaps with low confidence.

5) Combination of the genes' presence/absence status in all species and tracing of gene age through a voting algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appropriate repeat-masking makes GageTracker more time-efficient

The eukaryotic genome was composed of a vast number of transposable elements (TEs) in various stages of evolution. TEs generated intricate repeat patterns, contributing up to 44% of the human genome (43), 50% of mouse genomes (44), and 64% of the maize genome (45,46). Masked repeats (such as interspersed repeats and simple sequences) were a common pre-alignment step to simplify genome alignment and reduce excessive output. However, excessive repeat-masking posed a risk to ortholog search which would also lead to false positives. If the true ortholog of a genome was masked while a paralog was not, the paralog would be aligned incorrectly, leading to an inaccurate result. To solve this, we integrated an optimized method for filtering repeated sequences. As GageTracker was primarily designed to handle various focal species, it would be beneficial to use a software that could predict repeat sequences from scratch rather than relying solely on the repeatMasker database. We used Windowmasker/Tantan to soft mark the repeat sequence from scratch, revised the wrongly masked sequence in the protein-coding region, and then used the seeds-extend algorithm to search genome alignments. This could minimize aligning dissimilar and unrelated sequences, as if two dissimilar genes had the similar flanking sequences, the overall alignment score was likely to exceed the threshold value and produce a false-positive.

To assess the improvement in running time for genome alignments using the new soft-masked strategy, we conducted tests on a larger simulated dataset with distant genomic divergence ("simMammals" dataset in Alignathon (37)). The query-reference pairs (simCow-simHuman, simDog-simHuman, simMouse-simHuman, simRat-simHuman) were aligned using Lastal and GageTracker strategy (Lastal: last+without repeat-masking; GageTracker: last+repeat masking, wrapped in Gagetracker -step1).

GageTracker masked approximately 8.73% of the simHuman genome and 5% of the other four simulated datasets (Table 1). The genome alignment results demonstrated that GageTracker performed outperformed Lastal in terms of speed, with a reduction in running time of 18-22% ($p\text{-}value < 0.01$) by introducing repeat-masking strategy. Furthermore, GageTracker significantly reduced the output alignments (70-75%, $p\text{-}value < 0.01$) compared to Lastal, thereby could result in substantial time savings in subsequent RBH steps.

Moreover, we also evaluated the time cost of GageTracker based on several pairs of real datasets, including *D. melanogaster vs. D. yakuba* and *H. sapiens vs. M. musculus*. As expected, the real

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

datasets had higher levels of repeat sequences. GageTracker masked approximately 22.3% of the repeat sequences in the *D. melanogaster*, leading to a significant reduction of 43.3% in calculations and a 46.7% saving in running time than Lastal. Similarly, in the *H. sapiens vs. M. musculus* genomes, GageTracker masked 36.2% of repeat sequences in the *H. sapiens* genome, reducing redundant calculations by 60.3% and saving 68.7% of the runtime than Lastal. Together, these results indicated that proper repeat-masking in GageTracker could significantly save running time and reduce the output in genome alignment.

Table 1. Genome alignment between Lastal, GageTracker and LastZ.

Genome	Genome size	True repeat	Repeat mask (GageTracker)	MAF file			Runtime (min)		
				Lastal	GageTracker	LastZ	Lastal	GageTracker	LastZ
simHuman	191 Mb	11.09%	8.73%	-	-	-			
simCow	193 Mb	9.88%	4.85%	12.27 Gb	3.61 Gb	2.45 Gb	425	323	1958
simDog	192 Mb	10.18%	5.08%	12.52 Gb	3.71 Gb	2.24 Gb	356	267	2324
simMouse	199 Mb	12.13%	4.84%	8.42 Gb	2.13 Gb	0.64 Gb	363	281	445
simRat	199 Mb	11.47%	5.04%	8.26 Gb	2.05 Gb	0.68 Gb	436	356	541
D. melanogaster	144 Mb	-	22.32%	-	-	-	-	-	-
D. yakuba	166 Mb	-	7.80%	10.06 Gb	5.71 Gb	-	197	105	-
H. sapiens	3.09 Gb	-	36.21%	-	-	-	-	-	-
M. musculus	2.72 Gb	-	7.51%	1068.83 Gb	424.49 Gb	-	5412	1692	-

Note. These query-reference pairs (simCow-simHuman, simDog-simHuman, simMouse-simHuman, simRat-simHuman, *D. melanogaster* - *D. yakuba* and *H. sapiens* - *M. musculus* a) were aligned with Lastal, GageTracker, and LastZ aligner, respectively. The bold species represent the reference genome. Lastal: last + without repeat-masking; GageTracker: last + repeat-masking; LastZ: LastZ + truth repeats. When performing LastZ without repeat-masking, it failed to finish the alignment in 7 days and is excluded from the table. All runs are controlled in 6 threads on the servers with an E7-4850-CPU (2.5GHZ) array and 1 TB of memory.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GageTracker shows high precision and recall in ortholog search

To determine whether repeat-masking would impair the quality of the GageTracker, we compared it with LastZ. LastZ was a widely used and powerful genome alignment software that was relied upon by other software such as UCSC, Cactus, and SegAlign (47-49). The LastZ with suggested parameters (O=400 E=30 L=3000 H=2200 T=1) in UCSC could achieve high sensitivity but also result in much more time-consuming. To address this, we used PLastZ (https://github.com/AntoineHo/PLastZ), a parallel LastZ script, to improve alignment speed. Then, we used LastZ as the benchmark and compared the accuracy from these four strategies: Lastal (last + without repeat-masking), GageTracker (last + repeat-masking), LASTZ (LastZ + without repeat-masking), and LastZ (LastZ + repeat-masking). Generally, assessing the quality of genome aligners was challenging because the true homologous sequences were unknown. The Alignathon simulated datasets provided a solution, as they included a known truth data set for evaluating multiple genome aligners (50-53). To search for one-to-one orthologous alignments, we used the no paralog simulated Mammals data as the truth and the simHuman genome as a reference. We applied chains-and-nets and Syntenic net algorithms to find reciprocal best alignments from test data generated by Lastal, GageTracker, LASTZ, and LastZ strategies and employed mafTransitive and Closure mafComparator to calculate precision, recall, and *F*-score values. All runs were controlled in 6 threads on the servers with E7-4850-CPU (2.5GHZ) array respectively.

Without using repeat-masking strategy, LASTZ (LastZ + without repeat-masking) could not complete genome alignments within 7 days in the simulated data, far exceeding the expected time and making them impractical. But Lastal (last + without repeat-masking) finished genome alignments in 356~436 minutes (Table 1) and reciprocal best alignments in 430~513 minutes (Figure 1C), indicating that Lastal ran much faster than LASTZ. When using repeat-masking process, LastZ employed true repeat sequences and masked a high proportion of repeats (~10%), which should theoretically yield the fastest and most accurate data (Table 1). Nevertheless, the result showed that GageTracker had a faster computation speed and reduced runtime by 1.4-7 times compared to LastZ both in genome alignment and reciprocal best alignments process (Table1 and Figure 1C). Based on our analysis, GageTracker was found to be the fastest in terms of computational speed, followed by Lastal,

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

LastZ, and LASTZ, in descending order.

In addition, GageTracker and LastZ achieved comparable precision, recall, and *F-score* values across four mammal alignment groups (simHuman *vs.* simRat, simHuman *vs.* simMouse, simHuman *vs.* simDog, simHuman *vs.* simCow, Figure 1D). It also slightly improved *F*-scores and runtime compared to Lastal. These findings suggested that, in terms of calculation accuracy, GageTracker performed similarly to LastZ but slightly higher than Lastal. Thus, GageTracker was suitable for searching orthologous fragments, offering high sensitivity, accuracy, and efficient processing.

Gene age exhibits a high level of consistency in the result of GageTracker and GenTree in the real fly data set

We selected *Drosophila melanogaster*, which was an important model organism in genetic research with closely related species of improved sequencing quality (Middle panel of Figure 1B), to compare GageTracker and GenTree consistency in three levels (please refer to our method section: Age comparison in three different levels). GenTree was a reliable gene age database that integrated synteny-based, protein-family-based, and small-scale manual curation methods with functional genomic data to infer gene age (4). GenTree was built as a central portal to catalog lineage-specific genes in several model species. By using GageTracker, it took about 22 hours to estimate 23,720 genes' age (including ncRNA and protein genes) in 12 *Drosophila* sequencing genomes, of which approximately 13,965 were protein genes (Supplementary Table S2). Note that the two methods used different annotations, which would lead to a difference in gene number. We selected the overlapping protein genes in both annotations as our comparison objects (12,888, Figure 1E) to compare GageTracker and GenTree. On the global level, GageTracker exhibited high consistency with GenTree, with ~94.4% of the genes (12,171/12,888) sharing the same gene age (Figure 1F and Supplementary Table S3). Only 5.6% of the genes showed conflicts in exact gene age between the two methods (717, 2.4%+3.2%=5.6%, Figure 1F, and Supplementary Table S3). Additionally, when comparing the old and young gene sets (where branch 0 was considered old and branches 1-6 were considered young), GageTracker showed approximately 98% consistency with GenTree in the old catalog and approximately 80% consistency in the young

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

386 catalog (Figure 1G). At a detailed level, it was found that genes sharing the same age
387 annotation accounted for more than 70% of certain branches (over 98% in branch 0 and
388 almost 80% for genes in branches 5 and 6), with the exception of branches 2 and 3 (Figure
389 1E). However, for branches 2 and 3, the consistency still approached 50%. The consistency of
390 the gene age presented in this work was quite high if we made a parallel comparison. For
391 example, Shao et al. ever compared their results (GenTree) with the results from the other two
392 previous results, and they showed that there were approximately 30% and 40% conflicts
393 between GenTree and the other two related works, respectively. However, the comparison
394 between GageTracker and GenTree on the global level showed a high level of consistency.
395 Additionally, it was important to note that the conflicting gene age in some branches
396 could be caused by differences in genome assembled versions and gene-age dating methods
397 between GageTracker and GenTree. To evaluate the deviation between the two methods, we
398 used the branch distance metric. The results show that more than 97.94% of genes have a
399 branch distance of no more than 2, and more than 99% of genes have a branch distance of no
400 more than 3 (half of the branches are 3), indicating that the age annotated by the two methods
401 was very consistent. Together, these results suggested that GageTracker and GenTree
402 achieved a high level of consistency at global level and macroscopic level.
403
404 **GageTracker avoids overestimating the gene age on conflicting genes**
405 The inability to detect alignment hits in sequencing gaps was addressed by introducing
406 macro-synteny in our study, but GenTree did not consider this issue. To further clarify the
407 accuracy of GageTracker, we checked the ortholog genes of these 717 conflicting genes from
408 three different ortholog databases: OrthoDB, Ensembl, and FlyBase. The corresponding
409 orthologous genes of these 717 genes in 12 *Drosophila* genomes were screened out and the
410 genes' age was also estimated by the same voting cutoff that was used in GageTracker.
411 In OrthoDB, 667 conflicting genes had corresponding orthologs in 12 *Drosophila* species
412 (Supplementary Table S4). Approximately 36.28% (242/667) of conflicting genes had the
413 same gene ages between GageTracker *vs.* OrthoDB, whereas 32.53% (217/667) were
414 consistent between GenTree *vs.* OrthoDB, exhibiting that the consistency between
415 GageTracker *vs.* OrthoDB was slightly higher than that between GenTree *vs.* OrthoDB. This

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

result also suggested that a considerable proportion of the conflicting genes shared the same gene age annotation between either GageTracker and OrthoDB or GenTree and OrthoDB. In addition, we evaluated the conflicting genes by comparing the gene numbers of each branch distance between GageTracker *vs.* OrthoDB, and GenTree *vs.* OrthoDB. A higher proportion of genes in GageTracker (80.06%) had a shorter branch distance with OrthoDB (distance $\leq$2 branches, *p-value*=0.0137, Figure 2A) compared to GenTree (67.31%), which suggested that GageTracker was more consistent with OrthoDB. On the other hand, GenTree had a higher proportion of genes with a larger branch distance (distance $\geq$ 3 branches, *p-value* = 0.0075), indicating a greater propensity for overestimating gene age compared to GageTracker.

In Ensembl, we observed 717 conflicting genes that had corresponding annotations in our downloaded orthologs (Supplementary Table S5). We compared the gene numbers of each branch distance between GageTracker *vs.* Ensembl, and GenTree *vs.* Ensembl group. GageTracker *vs.* Ensembl group still showed a higher proportion gene with a shorter branch distance (0-1), while GenTree *vs.* Ensembl group had a higher proportion gene with a larger branch distance (Figure 2B, branch distance of 2-6, *p-value* = 0.0021). This result also indicated that, compared to Gentree, Gagetracker showed better consistency with Ensembl.

In FlyBase, 665 conflicting genes traced corresponding orthologs in 12 *Drosophila* species (Supplementary Table S6). GenTree *vs.* FlyBase group also exhibited a higher proportion of genes with larger branch distance (Figure 2C, branch distance of 3-6, *p-value* =0.0266) than the GageTracker *vs.* FlyBase group. To summarize, for the conflicting age, GageTracker avoided overestimation of genes age and obtained a more consistent with three other homologous sources (OrthoDB, FlyBase, and Ensembl), compared to GenTree. Since OrthoDB, Ensembl, and Flybase could not be taken as 'golden age' datasets, these analyses at least suggested that GageTracker's accuracy could be comparable to that of GenTree.

To investigate why GageTracker and GenTree showed age differences compared to three orthologous databases, we examined the copy numbers of these conflicting genes in their corresponding gene family (Supplementary Table S7) based on the paralogs recorded in FlyBase (dmel_paralogs_fb_2022_03.tsv.gz). It was found that approximately 62.12% (443/717) of conflicting genes were multiple copy genes and ~45% had more than two copies (Figure 2D). For conflicting genes with a larger branch distance (>3), around ~72% of them

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were found to be multiple copy genes, indicating that the conflicting genes with larger evolutionary branch bias were likely due to more paralogs (Figure 2D). The high sequence similarity of paralogous sequences increased the complexity of the gene family by adding more copies and/or changing their collinearity, which could result in more difficulty to infer the gene age accurately. Considering the result from the analysis of three ortholog databases, GageTracker demonstrates a higher level of accuracy in estimating gene age.

The expression of conflicting genes indicates that young genes in GageTracker but old in GenTree have a preference expression in testis

Previous research has shown that old genes tended to be expressed more broadly, whereas new genes were more tissue-specific and enriched in testes (3,39,54). Therefore, we conducted an expression pattern analysis on a subset of genes that showed conflicting gene age. First, we divided the genes into two groups (designated A and B) based on their age annotations in GageTracker and GenTree. The genes annotated as belonging to branch 0 in GageTracker and branches 1 to 6 in GenTree formed Group A, while the genes annotated as belonging to branches 1 to 6 in GageTracker and branch 0 in GenTree formed Group B. Using gene expression data from FlyAtlas (38), we compared the expression patterns of these two conflicting age groups. Our analysis revealed that, in male, group A genes demonstrated higher average expression levels than group B genes in 11 out of 15 tissues (Figure 3, Supplementary Table S8, over 73% of tissues). Additionally, group B genes exhibited a significant bias towards expression in the testis (Figure 3, Wilcoxon rank-sum test, *p-value* = 5.48e-11). In females, group A genes were also expressed at higher mean levels than group B genes in 10 different tissues (Figure 4, Supplementary Table S9). Overall, our findings suggested that genes classified as old by GageTracker exhibited higher average expression levels and a wider range of expression, while genes categorized as young by GageTracker but labeled as old by GenTree displayed a significant tendency to be expressed in the testis. To summarize, for genes with conflicting age annotations, those identified as old genes by GageTracker showed expression patterns typical of old genes, while those identified as young genes displayed male-biased expression patterns characteristic of young genes.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

GageTracker was a software tool that has two operation models: wrapped running mode and step-by-step running mode. The wrapped running mode integrated all necessary steps for gene age annotation, such as genome alignment, RBH calculation, and age assignment, into a single process specified in a configuration file. This mode was convenient for users as it separated command parameters and datasets. However, genome alignment and RBH calculation were the most time-consuming stages, and users may want to add new branches and outgroup species to existing genome alignment results. To address this, the step-by-step running mode was introduced, allowing users to add new genome alignments without affecting previous results, thus could save running time. The step-by-step mode also enabled the addition of new branches and focal species. In conclusion, GageTracker facilitated the annotation of new genes through its two operation modes (the wrapped mode integrates all necessary steps into a single process, while the step-by-step mode allows for the addition of new genome alignments without affecting previous results). GageTracker also integrated an RBH detection pipeline that was used by UCSC Genome Browser, which could help GageTracker find reciprocal best regions in a more accurate and more fast manner than other genome-based aligners (34).

Several reasons could affect the estimation of new genes, for example, the quality of genome assembly, genetic relationship between focal and query species, and sequence complexity, among which including high-quality genomes of closely related species help to accurately estimate the number of new genes. However, due to the lack of high-quality assembled genomes for some closely related species, new gene numbers may have been overestimated. Specifically, a gene may have been annotated as new not because it is absent in focal species, but rather due to a sequencing gap. As a result, we suggested that users could select a more complete reference genome to mitigate false positives resulting from genome sequencing gaps. By doing so, a more precise assessment of the number of novel genes could be achieved. Additionally, we recommend selecting 2-3 species from each branch to prevent false positives caused by gene loss. Furthermore, we have incorporated a criterion known as macro-synteny into GageTracker to identify sequencing gaps. Any genes located in such regions will be excluded from the annotation procedure since new genes in such regions may

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be false positives.

It should be noted that GageTracker, in its current version, was unable to differentiate between protein-coding genes and orthologous noncoding sequences. This distinction was a complex process that required comprehensive genomic and expression data to determine whether highly similar orthologous sequences in outgroup species were noncoding sequences. GenTree, on the other hand, utilized both evolutionary and proteomic evidence to assess the protein-coding potential of a sequence and overcome this limitation. Due to its inability to distinguish between protein-coding genes and orthologous noncoding sequences, GageTracker may overestimate the age of some young genes, leading to a slight overestimation of their true age (the estimated age is older than the actual age). However, for young genes, GageTracker was still able to provide reliable new gene candidates and their estimated origin age (i.e., the age when homologous sequences first emerged). This allowed researchers to trace back from the estimated ancient origin age and identify evidence of the gradual evolution of orthologous noncoding sequences into *de novo* genes. We have recently developed a random forest-based model that can detect coding capacity (55), which provided a means for estimating protein-coding potential. In the upcoming version of GageTracker, we plan to integrate the algorithm with transcriptome data to enable more accurate identification of gene structures. This will allow us to perform a more detailed analysis of the origin and evolutionary process of young genes, and generate structural diagrams for better visualization. This integration will improve the accuracy of our predictions and provide a more comprehensive understanding of the evolutionary history of young genes.

SUPPLEMENTARY DATA

Table S1. The genomes used in this study.

Table S2. Protein-coding genes' age of *Drosophila melanogaster* dated by GageTracker. Seven genes are located outside the block region, with low reliability (NCON)

Table S3. The estimation of gene age between GageTracker and GenTree.

Table S4. The gene age inferred by orthoDB for the conflicting genes between GageTracker

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

535 and GenTree.

536 Table S5. The conflicting gene age inferred by Ensembl orthologs database between

537 GageTracker and GenTree.

538 Table S6. The gene age inferred by the Flybase orthologs database for the conflicting genes

539 between GageTracker and GenTree.

540 Table S7. Paralogous of conflicting genes from the Flybase homologs database.

541 Table S8. Expression pattern of conflicting genes in male *Drosophila melanogaster*.

542 Table S9. Expression pattern of conflicting genes in female *Drosophila melanogaster*.

543

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AUTHOR CONTRIBUTIONS

SH, HZ, CF, and CD conceived and designed the study. CF and CD collected the data and performed the analysis. CW, FX, HZ, LY, XG, SL, WF and TL provided technical support. CF and CD wrote the manuscript. SH and XG revised the paper. All authors have read and accepted the final version of the manuscript.

Acknowledgments

We would like to thank Dr. Suxiang Lu, Ning Sun, and Han Xu for testing the GageTracker and providing some valuable suggestions.

Funding

This project is supported by the “Strategic Priority Research Program” of the Chinese Academy of Science (Grant No. XDB31000000), the National Natural Science Foundation of China (Grant, No. 32170438).

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Chen, S., Krinsky, B.H. and Long, M. (2013) New genes as drivers of phenotypic evolution. *Nat Rev Genet*, **14**, 645-660.
2. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Farrell, C.M., Feldgarden, M., Fine, A.M., Funk, K. *et al.* (2023) Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res*, **51**, D29-D38.
3. Zhang, W., Landback, P., Gschwend, A.R., Shen, B. and Long, M. (2015) New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol*, **16**, 202.
4. Shao, Y., Chen, C., Shen, H., He, B.Z., Yu, D., Jiang, S., Zhao, S., Gao, Z., Zhu, Z., Chen, X. *et al.* (2019) GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res*, **29**, 682-696.
5. Zhang, Y.E., Landback, P., Vibranovski, M.D. and Long, M. (2011) Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol*, **9**, e1001179.
6. Long, M.Y., VanKuren, N.W., Chen, S.D. and Vibranovski, M.D. (2013) New Gene Evolution: Little Did We Know. *Annu Rev Genet*, **47**, 307-333.
7. Levy, A. (2019) Genes from the Junkyard. *Nature*, **574**, 314-316.
8. Wei, W., Jin, Y.T., Du, M.Z., Wang, J., Rao, N. and Guo, F.B. (2016) Genomic Complexity Places Less Restrictions on the Evolution of Young Coexpression Networks than Protein-Protein Interactions. *Genome Biol Evol*, **8**, 2624-2631.
9. Fang, C., Gan, X., Zhang, C. and He, S. (2021) The new chimeric chiron genes evolved essential roles in zebrafish embryonic development by regulating NAD(+) levels. *Sci China Life Sci*, **64**, 1929-1948.
10. VanKuren, N.W. and Long, M. (2018) Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat Ecol Evol*, **2**, 705-712.
11. Long, M. and Langley, C.H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science*, **260**, 91-95.
12. Dai, H., Chen, Y., Chen, S., Mao, Q., Kennedy, D., Landback, P., Eyre-Walker, A., Du, W. and Long, M. (2008) The evolution of courtship behaviors through the origination of a new gene in Drosophila. *Proc Natl Acad Sci U S A*, **105**, 7478-7483.
13. Chen, S., Zhang, Y.E. and Long, M. (2010) New genes in Drosophila quickly become essential. *Science*, **330**, 1682-1685.
14. Xia, S.Q., VanKuren, N.W., Chen, C.Y., Zhang, L., Kemkemer, C., Shao, Y., Jia, H.X., Lee, U., Advani, A.S., Gschwend, A. *et al.* (2021) Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in Drosophila development. *Plos Genet*, **17**.
15. Xia, S., Wang, Z., Zhang, H., Hu, K., Zhang, Z., Qin, M., Dun, X., Yi, B., Wen, J., Ma, C. *et al.* (2016) Altered Transcription and Neofunctionalization of Duplicated Genes Rescue the Harmful Effects of a Chimeric Gene in Brassica napus. *Plant Cell*, **28**, 2060-2078.
16. Huang, Y., Chen, J., Dong, C., Sosa, D., Xia, S., Ouyang, Y., Fan, C., Li, D., Mortola, E., Long, M. *et al.* (2022) Species-specific partial gene duplication in Arabidopsis thaliana evolved novel phenotypic effects on morphological traits under strong positive selection. *Plant Cell*, **34**, 802-817.
17. Zhou, Y., Zhang, C., Zhang, L., Ye, Q., Liu, N., Wang, M., Long, G., Fan, W., Long, M. and

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Wing, R.A. (2022) Gene fusion as an important mechanism to generate new genes in the genus Oryza. *Genome Biol*, **23**, 130.

18. Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., Zhou, R. *et al.* (2019) Rapid evolution of protein diversity by de novo origination in Oryza. *Nat Ecol Evol*, **3**, 679-690.
19. Jin, G., Ma, P.F., Wu, X., Gu, L., Long, M., Zhang, C. and Li, D.Z. (2021) New Genes Interacted With Recent Whole-Genome Duplicates in the Fast Stem Growth of Bamboos. *Mol Biol Evol*, **38**, 5752-5768.
20. Wei, W., Zhang, T., Lin, D., Yang, Z.J. and Guo, F.B. (2013) Transcriptional abundance is not the single force driving the evolution of bacterial proteins. *BMC Evol Biol*, **13**, 162.
21. Wei, W., Ning, L.W., Ye, Y.N. and Guo, F.B. (2013) Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One*, **8**, e72343.
22. Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M.L. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet*, **23**, 5866-5878.
23. Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. and Alba, M.M. (2009) Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol*, **26**, 603-612.
24. Tong, Y.B., Shi, M.W., Qian, S.H., Chen, Y.J., Luo, Z.H., Tu, Y.X., Xiong, Y.L., Geng, Y.J., Chen, C. and Chen, Z.X. (2021) GenOrigin: A comprehensive protein-coding gene origination database on the evolutionary timescale of life. *J Genet Genomics*, **48**, 1122-1129.
25. Tay, S.K., Blythe, J. and Lipovich, L. (2009) Global discovery of primate-specific genes in the human genome. *Proc Natl Acad Sci U S A*, **106**, 12019-12024.
26. Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A. and Long, M. (2010) Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol*, **8**, e1000494.
27. Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J.L. *et al.* (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat Genet*, **50**, 285-296.
28. Kabza, M., Ciomborowska, J. and Makalowska, I. (2014) RetrogeneDB--a database of animal retrogenes. *Mol Biol Evol*, **31**, 1646-1648.
29. Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simao, F.A. and Zdobnov, E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*, **47**, D807-D811.
30. Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E.V. and Zdobnov, E.M. (2023) OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res*, **51**, D445-D451.
31. Fan, G., Song, Y., Yang, L., Huang, X., Zhang, S., Zhang, M., Yang, X., Chang, Y., Zhang, H., Li, Y. *et al.* (2020) Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *Gigascien*ce, **9**, giaa080.
32. Zhang, G., Rahbek, C., Graves, G.R., Lei, F., Jarvis, E.D. and Gilbert, M.T. (2015) Genomics: Bird sequencing project takes off. *Nature*, **522**, 34.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

33. Miao, W., Song, L., Ba, S., Zhang, L., Guan, G., Zhang, Z. and Ning, K. (2020) Protist 10,000 Genomes Project. *Innovation (Camb)*, **1**, 100058.
34. Frith, M.C. and Kawaguchi, R. (2015) Split-alignment of genomes finds orthologies more accurately. *Genome Biol*, **16**, 106.
35. Kolmogorov, M., Armstrong, J., Raney, B.J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T.M., Thybert, D. *et al.* (2018) Chromosome assembly of large and complex genomes using multiple references. *Genome Res*, **28**, 1720-1732.
36. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
37. Earl, D., Nguyen, N., Hickey, G., Harris, R.S., Fitzgerald, S., Beal, K., Seledtsov, I., Molodtsov, V., Raney, B.J., Clawson, H. *et al.* (2014) Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res*, **24**, 2077-2089.
38. Leader, D.P., Krause, S.A., Pandit, A., Davies, S.A. and Dow, J.A.T. (2018) FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res*, **46**, D809-D815.
39. Zhang, Y.E., Vibranovski, M.D., Krinsky, B.H. and Long, M. (2010) Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res*, **20**, 1526-1533.
40. Fang, C., Zou, C., Fu, Y., Li, J., Li, Y., Ma, Y., Zhao, S. and Li, C. (2018) DNA methylation changes and evolution of RNA-based duplication in Sus scrofa: based on a two-step strategy. *Epigenomics*, **10**, 199-218.
41. Morgulis, A., Gertz, E.M., Schaffer, A.A. and Agarwala, R. (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, **22**, 134-141.
42. Frith, M.C. (2011) A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res*, **39**, e23.
43. Mills, R.E., Bennett, E.A., Iskow, R.C. and Devine, S.E. (2007) Which transposable elements are active in the human genome? *Trends Genet*, **23**, 183-191.
44. Bruno, M., Mahgoub, M. and Macfarlan, T.S. (2019) The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. *Annu Rev Genet*, **53**, 393-416.
45. SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765-768.
46. Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S. *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524-527.
47. Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D. and Haussler, D. (2011) Cactus: Algorithms for genome multiple sequence alignment. *Genome Res*, **21**, 1512-1528.
48. Goenka, S.D., Turakhia, Y., Paten, B. and Horowitz, M. (2020) SegAlign: A Scalable GPU-Based Whole Genome Aligner. *Proceedings of Sc20: The International Conference for High Performance Computing, Networking, Storage and Analysis (Sc20)*.
49. Lee, B.T., Barber, G.P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M. *et al.* (2022) The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res*, **50**, D1115-D1122.
50. Kille, B., Balaji, A., Sedlazeck, F.J., Nute, M. and Treangen, T.J. (2022) Multiple genome

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

alignment in the telomere-to-telomere assembly era. *Genome Biol*, **23**, 182.

51. Minkin, I. and Medvedev, P. (2020) Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nat Commun*, **11**, 6327.
52. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J. *et al.* (2020) Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**, 246-251.
53. Armstrong, J., Fiddes, I.T., Diekhans, M. and Paten, B. (2019) Whole-Genome Alignment and Comparative Annotation. *Annu Rev Anim Biosci*, **7**, 41-64.
54. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650-659.
55. Yu, J., Jiang, W., Zhu, S.B., Liao, Z., Dou, X., Liu, J., Guo, F.B. and Dong, C. (2023) Prediction of protein-coding small ORFs in multi-species using integrated sequence-derived features and the random forest model. *Methods*, **210**, 10-19.

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE AND FIGURES LEGENDS

Table 1. Genome alignment between Lastal, GageTracker and LastZ.

These query-reference pairs (simCow-simHuman, simDog-simHuman, simMouse-simHuman, simRat-simHuman, *D. melanogaster* - *D. yakuba* and *H. sapiens* - *M. musculus* a) were aligned with Lastal, GageTracker, and LastZ aligner. The bold species represent the reference genome. Lastal: last + without repeat-masking; GageTracker: last + repeat-masking; LastZ: LastZ + truth repeats. When performing LastZ without repeat-masking, it failed to finish the alignment in 7 days and is excluded from the table. All runs are controlled in 6 threads on the servers with an E7-4850-CPU (2.5GHZ) array and 1 TB of memory.

Figure 1. The pipeline and performance of GageTracker.

(A) The overview of inputting data (orange), UCSC Genome Browser toolkit (green), post-processing (purple), and output (red). (B) A diagram about alignment block (macro-synteny) and micro-synteny. The rectangles with the same colors mean the homologous genes (left panel in B). The left panel of B also illustrates the Voting strategy for assessing the support ratio of a gene's origin on a branch. The orange rectangle represents a gene that has an ortholog on the corresponding branch, while the white rectangle means absence. The “br” represents the short name of a branch. The number in the orange rectangle represents the support ratio of the gene originating in that branch. For example, Gene 1 in the dotted box has corresponding ortholog on branches 6, 3, 2, and 0. The support ratio of Gene 1 's origin on branch 3 is 50% (2/4, present on 3 out of 4 branches), on branch 2 is 60% (3/5, present on 3 out of 5 branches), and on branch 0, its support rate reaches 57% (4/7). Then, this gene age was placed on its oldest branch 0 (red arrow). Gene 3 independently emerges on branch 6 as its support rate of originating on branch 0 is only 28.5%. The detailed calculation strategy could be seen in the Method. The middle panel of B is a phylogeny tree of *Drosophila* species and the blue lines indicated the evolution direction leading to *D. melanogaster*. The topology of the tree and the divergent time of each branch point were retrieved from the previous study (39). The right panel of B is the gene number comparison between GenTree and GagaTracker. (C) the total running time of three aligners for producing reciprocal best alignments: Lastal

bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(last+without repeat-masking), GageTracker (last+repeat-masking/GageTracker -step1), and Lastz (LastZ + true repeats). LASTZ (LastZ+without repeat-masking) was not displayed in this figure due to its failure to finish the genome alignments in 7 days in the simulated data. We applied chains-and-nets and syntenic net algorithms to find reciprocal best alignments and Closure mafComparator to calculate precision, recall, and *F-score* values. (D) Precision, recall as well as *F*-score values of three aligners. (E) The comparison of Gene annotations employed by GenTree and GagaTracker. (F) Comparison on the global level. (G) Comparison on the gene catalogs of young and old genes (nearly-detailed level). (H) Comparison on the gene catalogs of each branch (detailed level). (I) Cumulative curve of gene number with branch distance between the GageTracker and Gentree database.

Figure 2. Comparison of conflicting genes with branch distance between GageTracker and GenTree. (A) Cumulative curve of conflicting ages with branch distance between GageTracker *vs.* orthoDB and Gentree *vs.* orthoDB. *p-value* (0-2) represents the significance test in a shorter branch distance (0-2) group and *p-value* (3-6) represents the significance test in a larger branch distance (3-6) group. (B) Cumulative curve of conflicting ages with branch distance between GageTracker *vs*. Ensembl and Gentree *vs.* Ensembl. (C) Cumulative curve of conflicting ages with branch distance between GageTracker *vs.* FlyBase and Gentree *vs.* FlyBase. (D) The percentage of conflicting genes with multiple copies.

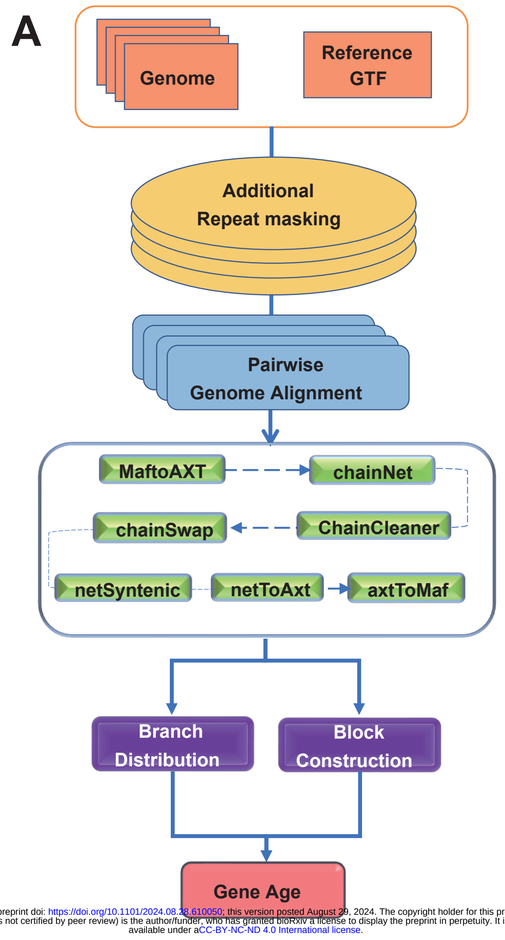
Figure 3. Expression pattern of conflicting genes in male of *D. melanogaster*. The tissue was marked in the top-left corner of each subfigure. The genes annotated as old gene in GageTracker （branch 0） but young gene in GenTree （from branches 1 to 6） formed Group A, while the genes annotated as young gene in GageTracker but old gene in GenTree formed Group B. The expression value was transformed by log function (log10). The red and green dash lines represent the mean expressions in group A and group B.

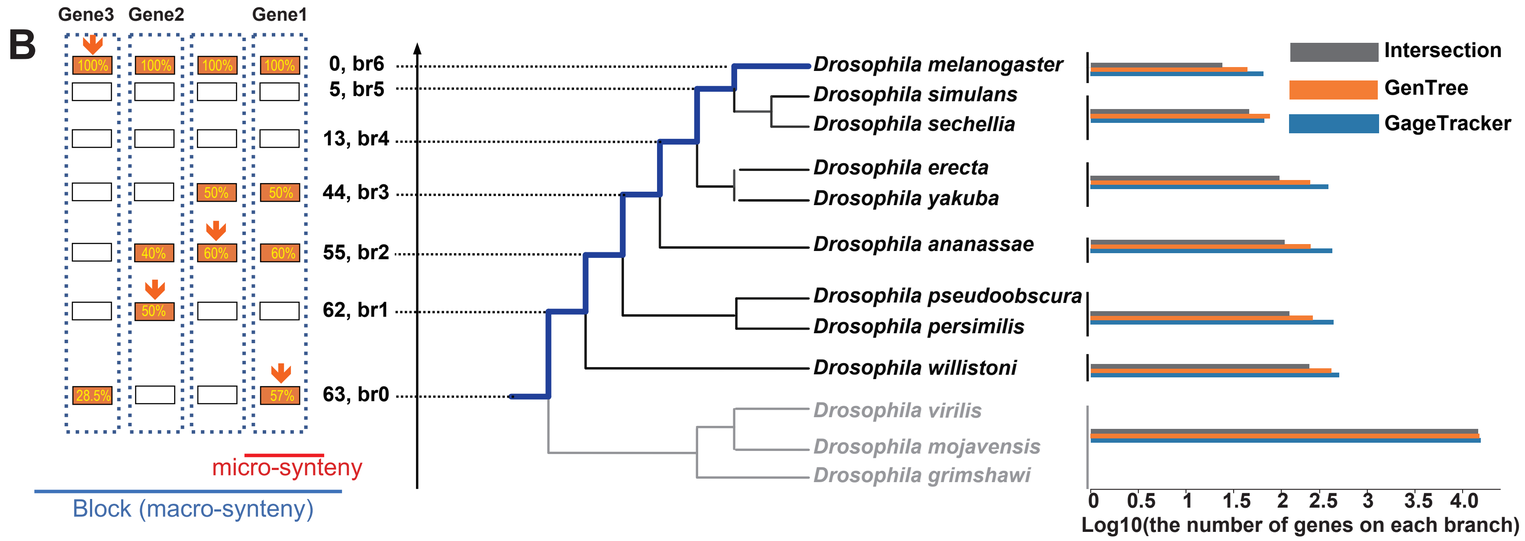
Figure 4. Expression pattern of conflicting genes in female of *D. melanogaster*. The tissue was marked in the top-left corner of each subfigure. The genes annotated as old gene in

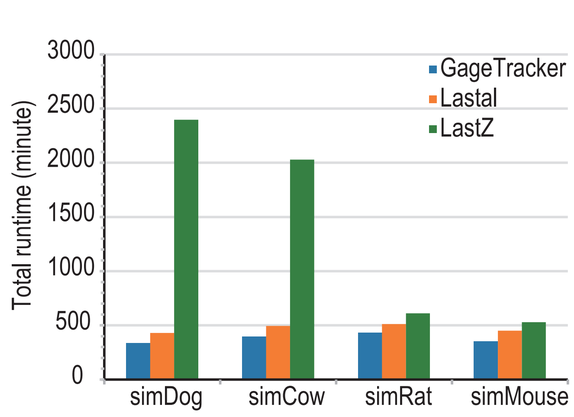
bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

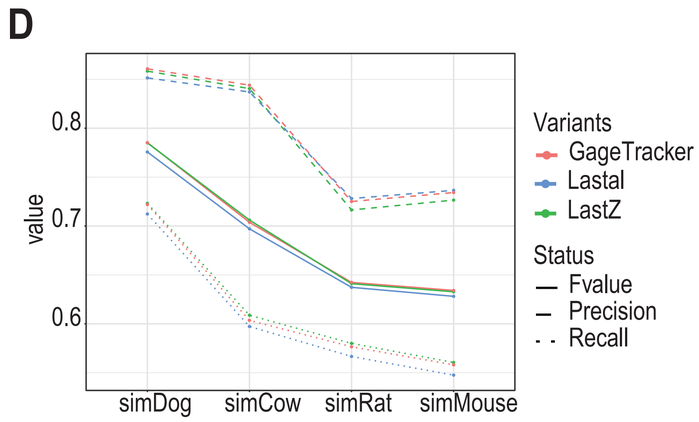
GageTracker（branch 0）but young gene in GenTree（from branches 1 to 6）formed Group A, while the genes annotated as young gene in GageTracker but old gene in GenTree formed Group B. The expression value was transformed by log function (log10). The red and green dash lines represent the mean expressions in group A and group B.

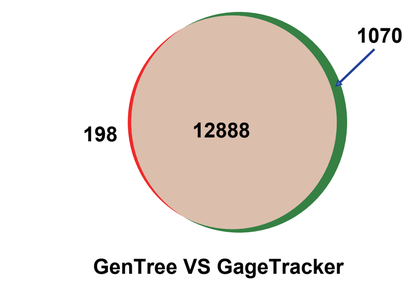
bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

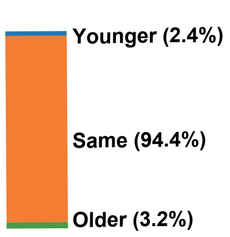


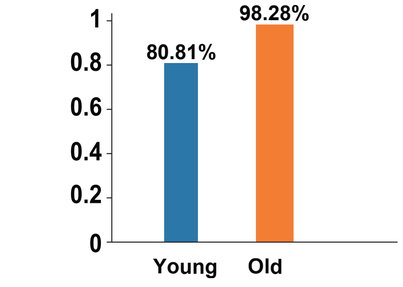


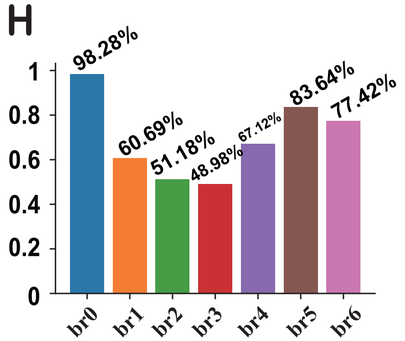


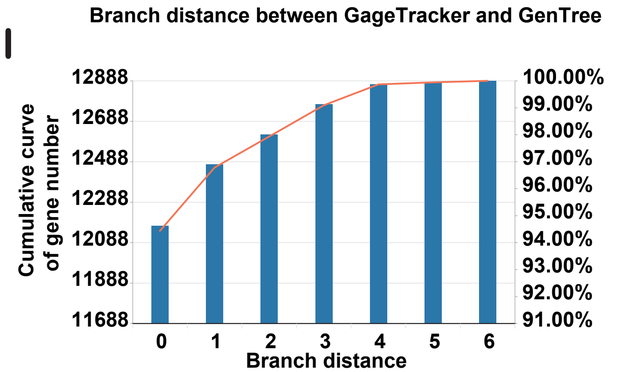






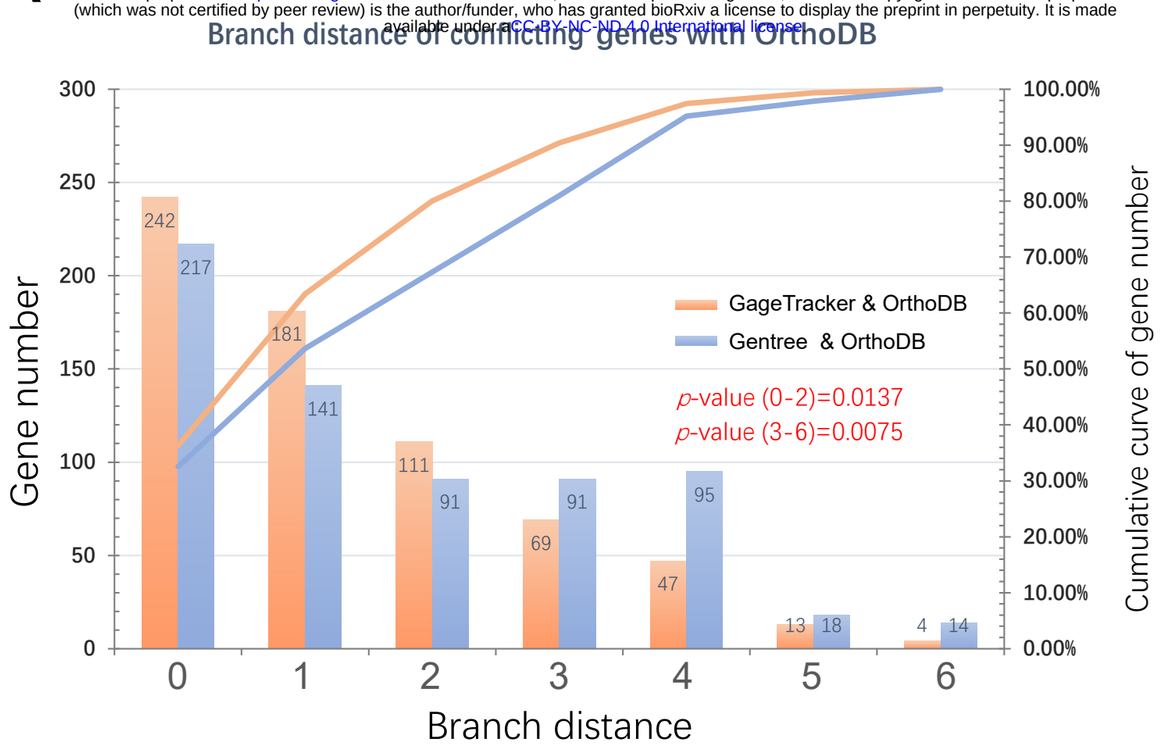




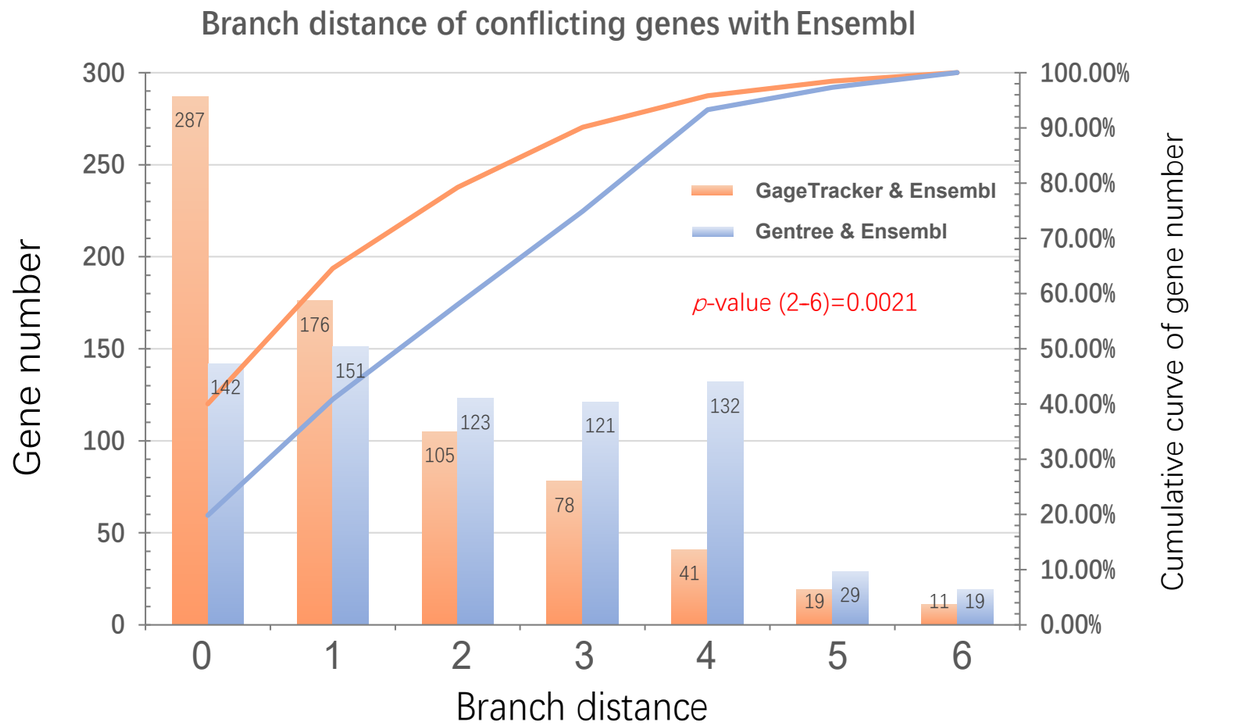


bioRxiv preprint doi: https://doi.org/10.1101/2024.08.28.610050; this version posted August 29, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

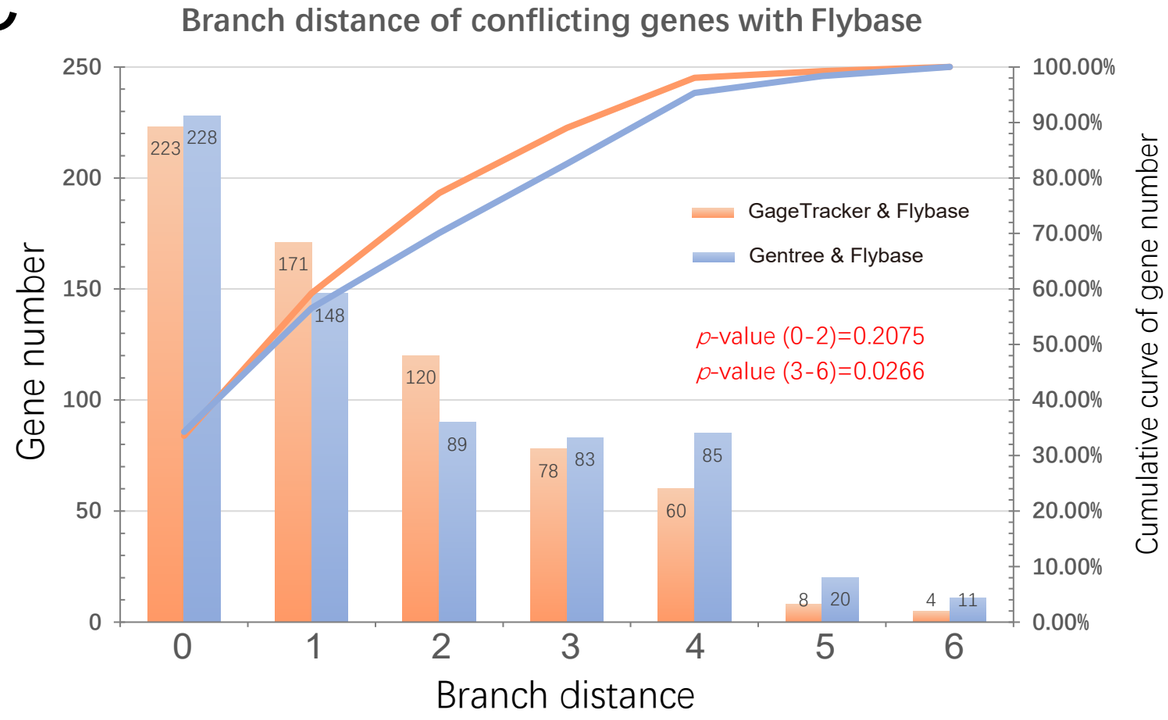
A



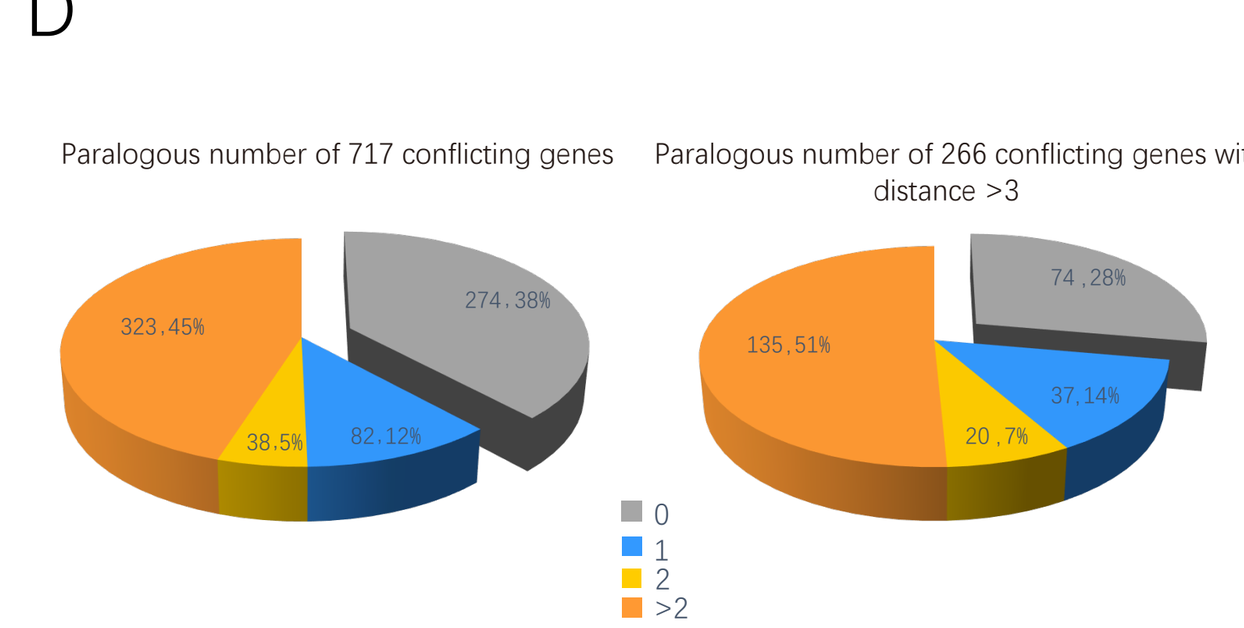
B

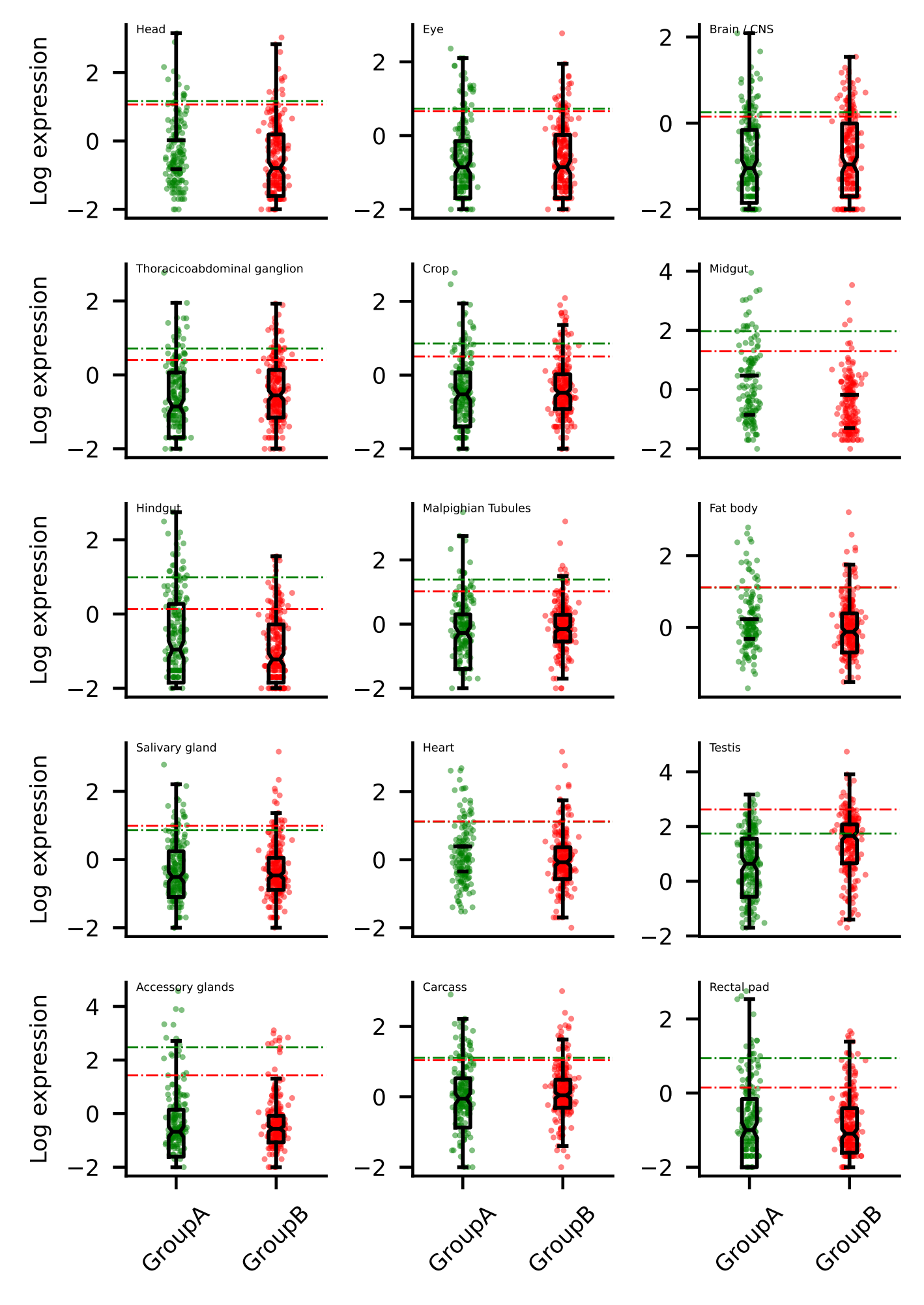


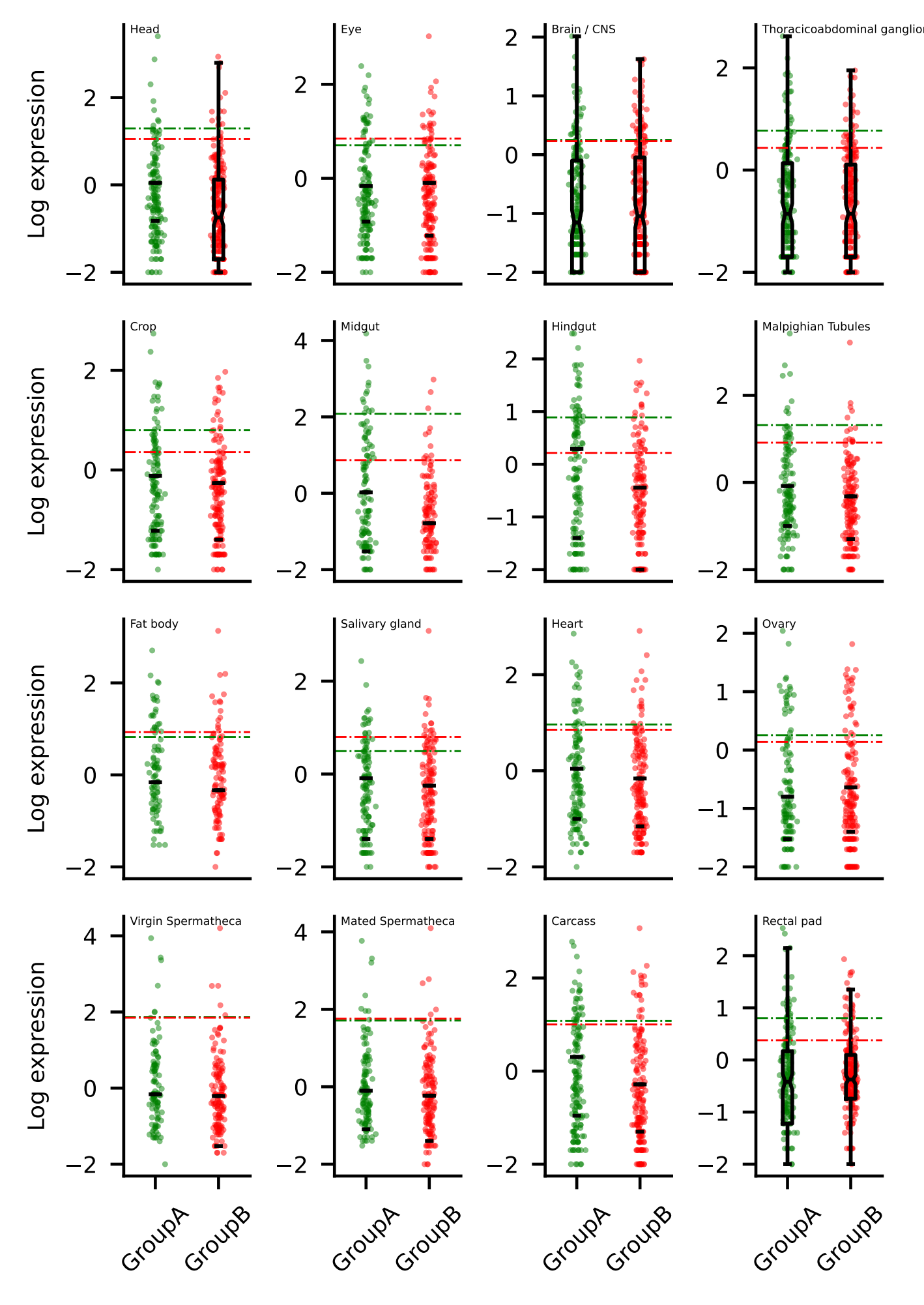
C



D




Head
Eye
Brain / CNS
Thoracicoabdominal ganglion
Crop
Midgut
Hindgut
Malpighian Tubules
Fat body
Salivary gland
Heart
Testis
Accessory glands
Carcass
Rectal pad
Log expression
GroupA
GroupB


Head
Eye
Brain / CNS
Thoracicoabdominal ganglion
Crop
Midgut
Hindgut
Malpighian Tubules
Fat body
Salivary gland
Heart
Ovary
Virgin Spermatheca
Mated Spermatheca
Carcass
Rectal pad
Log expression
GroupA
GroupB