Title: Molecular Basis of Behavioral Diversity in a Sibling Species Trio

Authors: Cynthia M. Chai[1]*

[1] Department of Biological Sciences, Columbia University, 1212 Amsterdam Ave, New York, NY 10027, USA

* Corresponding author: Cynthia M. Chai

Email: cc4960@columbia.edu

Author Contributions: C.M.C. designed the study, performed behavior experiments and molecular biology, analyzed RNA-seq data, acquired resources and funding, prepared the figures and wrote the paper.

Competing Interest Statement: The author declares that there are no competing interests.

Keywords: Brain evolution, Gene expression, Sibling species, Locomotor behavior, Speciation

Abstract

The brain is the main controller of animal behaviors. During the speciation process, divergent behaviors can arise between new sibling species. However, the comprehensive molecular changes in the brain that give rise to these emergent behavioral differences are not well understood. Here, I present a comparative analysis of brain region-specific gene expression between a trio of closely-related *Drosophila* sibling species representing two consecutive speciation events. My analysis revealed strongest patterns of molecular conservation across speciation events in the brain region that executes movement, while the visual sensory system tends to be most divergent in gene expression. I also introduce a novel locomotor activity difference between sibling species, and demonstrate how set theory can be applied in the context of my experimental design to constrain the number of candidate neuronal genes underlying behavioral diversification for further detailed investigation. As genomic resources expand (1), iteratively exploring these patterns across ever larger swaths of animal phylogeny is poised to reveal general principles of brain functional molecular evolution.

Main Text

Results & Discussion

Darwin's observation of the myriad forms of finch beaks in the Galapagos Islands provides a classic example of the rich biodiversity that arises as animals adapt to novel ecological niches (2). Genomic variation impacting essential behaviors, like foraging and mating, can drive reproductive isolation between populations that exhibit differential responses, ultimately resulting in speciation. However, the proximate mechanisms in the nervous system underlying behavioral divergence between new sibling species are unclear. Recent studies using closely-related sibling species have reported either cell-type specific transcriptomic differences between homologous neurons or transcriptomic changes in a subset of brain regions between species or whole-brain bulk changes between species pairs (3-6). Here I take a complementary approach by providing a global view of comparative gene expression changes across the entire nervous system at region-specific resolution in a trio of closely-related sibling species representing two consecutive speciation events.

Like *Drosophila melanogaster*, *Drosophila simulans*, which diverged from *D. melanogaster* around 3 million years ago, is a cosmopolitan species that has colonized diverse habitats (7-9) (Figure 1A). In contrast, *Drosophila mauritiana* is endemic to the volcanic islands of Mauritius and diverged from *D. simulans* around 250,000 years ago (7, 9, 10) (Figure 1A). The *Drosophila* nervous system can be partitioned into three anatomical regions associated with generally distinct hierarchical neurobehavioral functions: visual sensory (optic lobes) (11), higher-order processing (central brain)

(12, 13), and movement (ventral nerve cord) (14) (Figure 1B). By comparing region-specific bulk transcriptomic profiles between all three sibling species (Figure 1C), I first sought to determine the relative strengths of selective pressures acting on neurobehavioral functions as indicated by the extent of inter-specific differential gene expression. As RNA-seq reads were mapped to the *D. melanogaster* reference genome for all sibling species, *D. simulans* and *D. mauritiana* species-specific genes are therefore excluded from this study's analyses (15). Across consecutive speciation events, the ventral nerve cord, which primarily orchestrates movement, exhibited the lowest number of significantly differentially expressed genes between sibling species (Figure 1D and 1E). Conversely, the optic lobes, a major sensory modality in *Drosophila*, was consistently most divergent in gene expression across speciation events (Figure 1E). Collectively, these findings suggest that sensory systems, which interface directly with the changing environment, may experience greater selective pressures to adapt relative to downstream motor circuits in the ventral nerve cord, which are modulated by high-order information processing units in the central brain. In other words, how a fly responds to an external challenge, as opposed to its ability to respond, appears more likely to differ between speciated populations at the molecular level.

To probe how neuronal gene expression differences between sibling species could translate into species-specific organismal phenotypes, I next assayed multiple lines of each species for a general locomotor activity behavior (Figure 2A). I found that *D. simulans* and *D. mauritiana* exhibited significantly lower locomotor activity levels relative to *D. melanogaster* (Figure 2A). This sedentary behavior was consistent across all six lines of *D. simulans* from different geographical regions as well as across all four lines of *D. mauritiana* assayed implying that this trait was likely present in the last common ancestor of both sibling species (Figure 2A and 2B). As my comparative RNA-seq experimental design comprises a trio of sibling species mapped to the same reference genome rather than just a species pair, set theory can thus be applied to constrain the number of candidate neuronal genes that could underlie these inter-specific behavioral differences (Figure 2C). As *D. simulans* and *D. mauritiana*, which are members of the *simulans* clade, consistently exhibited the same decreased locomotor activity phenotype relative to *D. melanogaster*, neuronal genes that are differentially expressed in both *D. simulans* (Set A) and *D. mauritiana* (Set B) compared to *D. melanogaster* are more likely to contribute to the behavioral phenotype (Figure 2C). The intersection of both sets ($A \cap B$) results in a 41-52% reduction in candidate gene set size relative to A and B across brain regions (Figure 2D). Neuronal genes that are significantly differentially expressed between *D. mauritiana* and *D. simulans* are not likely to contribute to the shared decreased locomotor activity phenotype. Thus, subsequently obtaining the set difference of $A \cap B$ and the set of differentially expressed genes between *D. mauritiana* and *D. simulans* (Set C) results in a marginal further decrease in candidate neuronal gene set size ($(A \cap B) \setminus C$) (Figure 2D).

The brain region-specific transcriptomics dataset and relevant set operations strategies presented here generalizes and can be applied to all behavioral phenotype combinations examined in this sibling species trio (Figure 2E). Moreover, inclusion of two consecutive speciation events in the experimental design facilitates the inference of corresponding molecular evolution scenarios underlying behavioral divergence (Figure 2F). In the absence of *a priori* knowledge regarding the identity of and changes in neural circuitry underlying behavioral differences between species, spatial analysis of candidate differentially expressed genetic elements reported in this study can provide a path towards uncovering the neuronal sites of evolutionary innovation. Tracing the specific tunable steps leading to the emergence of new species could enable us to better understand and predict Life's trajectory in an ever-changing biosphere.

Materials and Methods

For comparative bulk RNA-seq, total RNA was extracted from brain regions dissected from three to five-day old adult female flies. Paired-end RNA sequencing was performed and reads were mapped to the *D. melanogaster* reference genome. Differential gene expression analysis was performed using DESeq2 in R. For behavior experiments, all flies were maintained under identical growth conditions. Locomotor activity behavior assays were performed using a DAM5H *Drosophila* activity monitor device. See SI Methods for details.

Acknowledgments

I thank the Janelia Quantitative Genomics Team for RNA-seq sample preparation and sequencing. P. Andolfatto provided fly food, some fly species stocks, and equipment (R01 GM115523, National Institutes of Health). This work was supported by a Leon Levy Scholarship in Neuroscience from the Leon Levy Foundation (to C.M.C.) and a Konishi Neuroethology Research Award from the International Society for Neuroethology (to C.M.C.).

References

1. M. Blaxter *et al.*, The Earth BioGenome Project Phase II: illuminating the eukaryotic tree of life. *Frontiers in Science* **Volume 3 - 2025** (2025).
2. C. Darwin, *The voyage of the Beagle*, The Harvard classics, ed by C W Eliot (P.F. Collier & son, New York,, 1909), pp. 1 p. l., 547 p.
3. J. T. Walsh *et al.*, High-resolution single-cell analyses reveal evolutionary constraints and evolvability of sexual circuits in *Drosophila*. *Proceedings of the National Academy of Sciences* **122**, e2516083122 (2025).
4. A. F. Kautt *et al.*, Evolution of gene expression across brain regions in behaviourally divergent deer mice. *Molecular Ecology* **34**, e17270 (2025).
5. D. Lee, M. P. Shahandeh, L. Abuin, R. Benton, Comparative single-cell transcriptomic atlases of drosophilid brains suggest glial evolution during ecological adaptation. *PLoS Biol* **23**, e3003120 (2025).
6. S. H. Montgomery, M. Rossi, W. O. McMillan, R. M. Merrill, Neural divergence and hybrid disruption between ecologically isolated *Heliconius* butterflies. *Proceedings of the National Academy of Sciences* **118**, e2015102118 (2021).
7. D. Garrigan *et al.*, Genome sequencing reveals complex speciation in the Drosophila simulans clade. *Genome Res* **22**, 1499-1511 (2012).
8. M. D. Dean, J. W. O. Ballard, Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. *Molecular Phylogenetics and Evolution* **32**, 998-1009 (2004).
9. J. R. David, F. o. Lemeunier, L. Tsacas, A. Yassin, The Historical Discovery of the Nine Species in the *Drosophila melanogaster* Species Subgroup. *Genetics* **177**, 1969-1973 (2007).
10. S. R. McDermott, R. M. Kliman, Estimation of Isolation Times of the Island Species in the *Drosophila simulans* Complex from Multilocus DNA Sequence Data. *PLOS ONE* **3**, e2442 (2008).
11. N. Nériec, C. Desplan, From the Eye to the Brain: Development of the *Drosophila* Visual System. *Curr Top Dev Biol* **116**, 247-271 (2016).
12. B. K. Hulse *et al.*, A connectome of the *Drosophila* central complex reveals network motifs suitable for flexible navigation and context-dependent action selection. *eLife* **10**, e66039 (2021).
13. Y. Aso *et al.*, The neuronal architecture of the mushroom body provides a logic for associative learning. *eLife* **3**, e04577 (2014).
14. R. Court *et al.*, A Systematic Nomenclature for the *Drosophila* Ventral Nerve Cord. *Neuron* **107**, 1071-1079.e1072 (2020).
15. E. W. Myers *et al.*, A whole-genome assembly of *Drosophila*. *Science* **287**, 2196-2204 (2000).

Figures

Figure 1. Comparative brain transcriptomics between a *Drosophila* sibling species trio representing two consecutive speciation events reveals brain regional patterns of molecular evolution.

(A) Phylogeny of *Drosophila melanogaster*, *Drosophila simulans*, and *Drosophila mauritiana* sibling species (7, 10).

(B) Schematic of dissected *Drosophila* brain regions labeled with their associated general neurobehavioral functions.

(C) Principal component analysis of bulk RNA-seq samples (25 samples). Two pre-identified outliers, one *D. mauritiana* ventral nerve cord sample and one *D. simulans* optic lobe sample, were removed here and for all subsequent analyses.

(D) Hierarchical clustering heatmap of correlation matrix for bulk RNA-seq samples (25 samples).

(E) Number of significantly differentially expressed genes, p-adjusted<0.05 (left) and 0.01 (right), for each species pairwise test for each brain region. Number of genes with non-zero total read count: optic lobes (13,012), central brain (13,325), and ventral nerve cord (13,246).

Figure 2. Set theory further constrains the number of candidate neuronal genes underlying behavioral differences between *Drosophila* sibling species.

(A) Schematic of locomotor activity assay with black arrow indicating one of fifteen infrared beams intersecting each glass assay tube (Top). Locomotor activity behavior for *Drosophila* sibling species lines (Bottom). The total number of infrared beam transitions during the 30-min assay duration is a proxy for locomotor activity levels. n=number of flies. Kruskal-Wallis followed by Dunn's post hoc with Holm correction, n.s.=no significance, ***p<0.001. See SI Methods for species stock details.

(B) Map showing geographical origins of *Drosophila* species lines assayed in (A). Orange – Mauritius, Pink – Madagascar island, Purple – Kenya, Red – Southern California, Black – Northern California, and Green – Oregon. *D. melanogaster* USA Line #1, *D. simulans* USA Line #1, and *D. mauritiana* Line #1 were used to generate the RNA-seq dataset in Figure 1.

(C) Set operations strategy to constrain the number of candidate neuronal genes underlying inter-specific differences in locomotor activity behavior. The universal set (*U*) is defined as the set of all genes in the *D. melanogaster* reference genome.

(D) Comparison of candidate neuronal gene set sizes between each species pairwise test and after executing set operations strategy in (C) (RNA-seq significance thresholds: p-adjusted<0.05 and $|\log_2 FC|>2.0$).

(E) Generalizable sibling species behavioral phenotype combinations and corresponding set operations strategies to constrain candidate neuronal gene set sizes. *N*(X) denotes the size of set X (i.e. the number of differentially expressed genes in set X). See SI Methods for code implementation.

(F) Molecular evolution scenarios corresponding to sibling species phenotypic outcomes in (E). For each scenario, the circle indicates the relevant speciation event and the arrow indicates the branch along which the causal genomic variation could have occurred. For 3), green indicates an independent change subsequent to the initial change in red that reverts the *D. mauritiana* behavior to the *D. melanogaster* behavior. [#]Working assumption in (E): The subsequent genomic change impacts the same gene(s) impacted by the initial change. If not, then the set operations strategy is A or C.

Figure 1

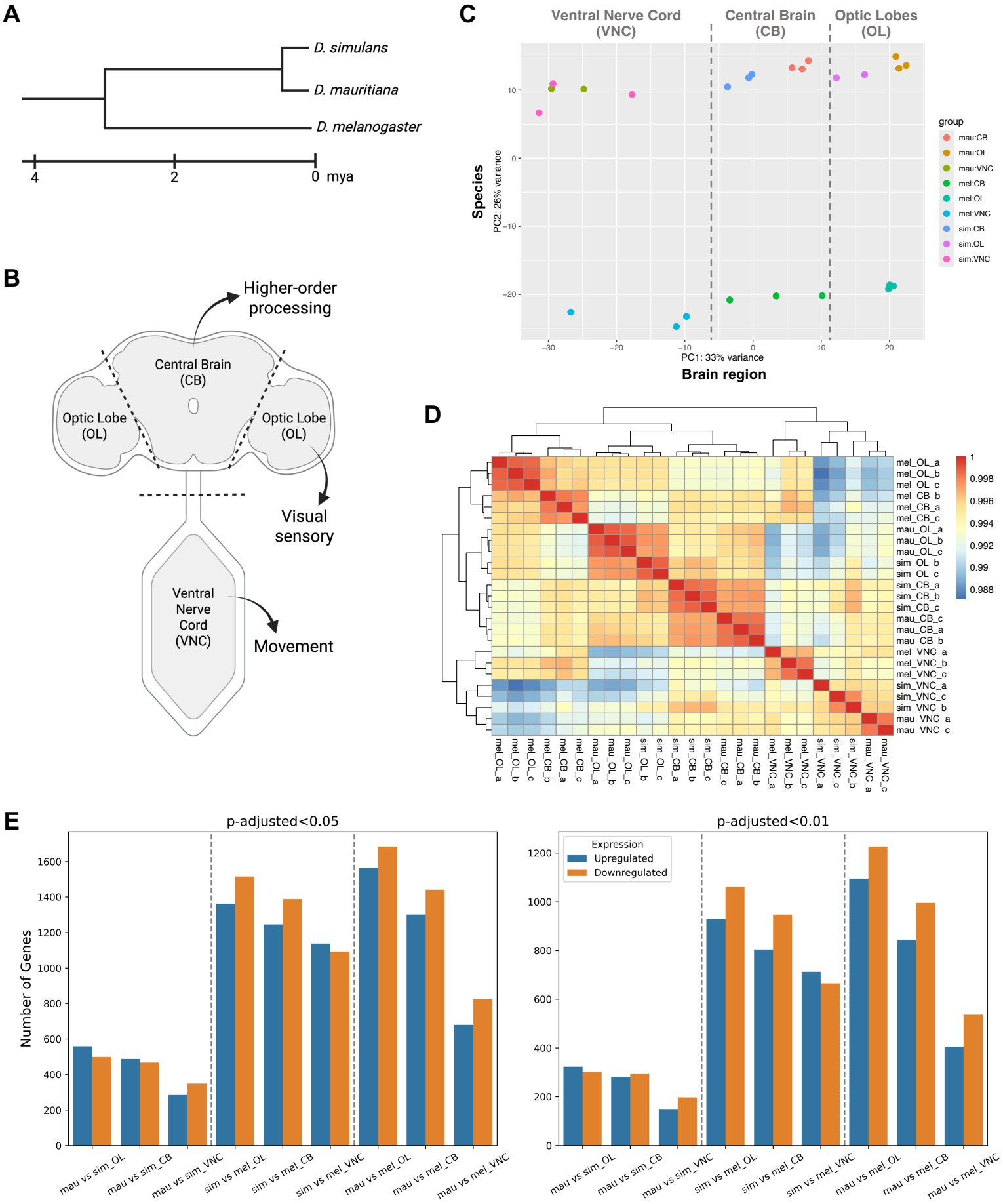
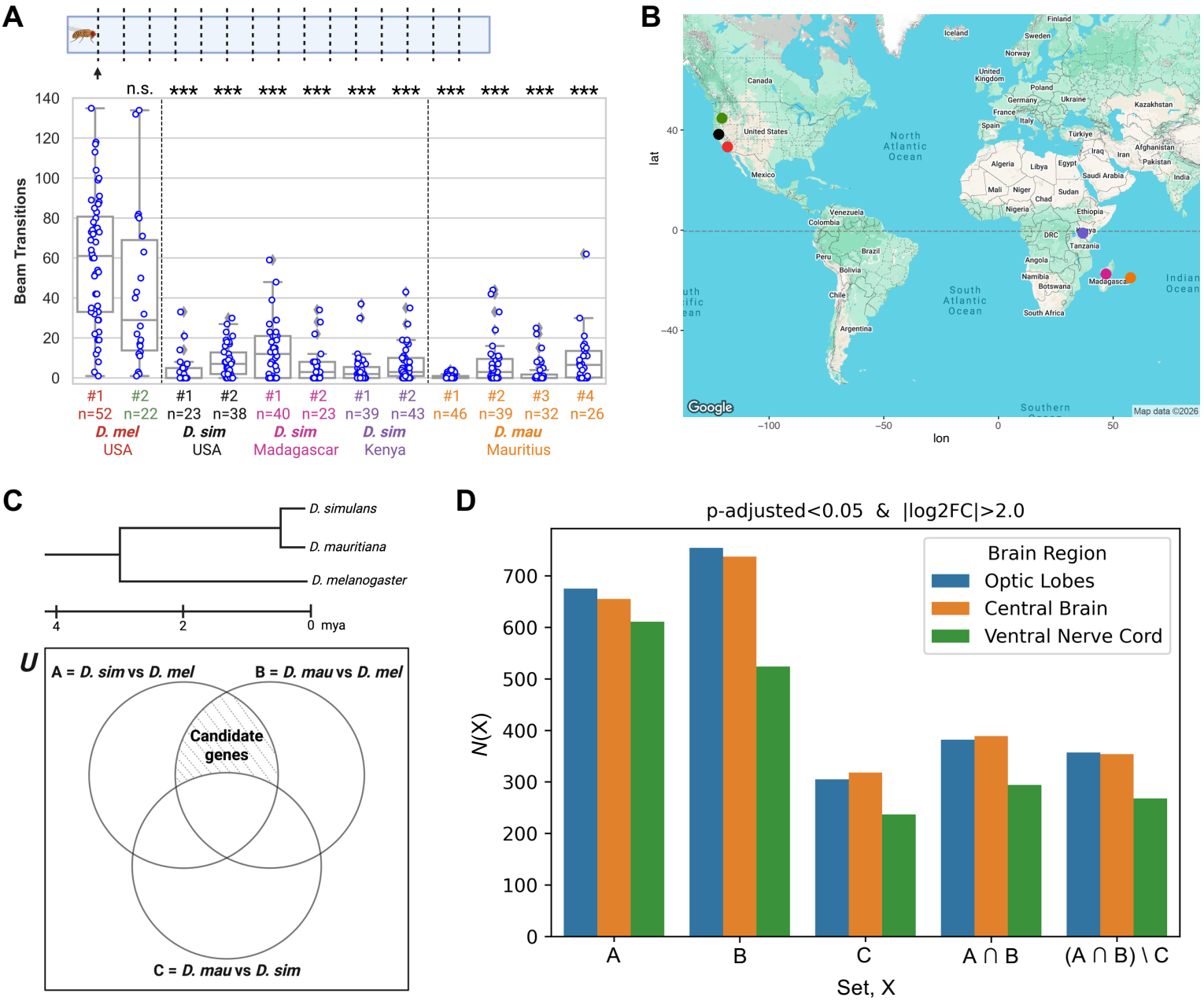

A
D. simulans
D. mauritiana
D. melanogaster
4
2
0 mya
B
Higher-order processing
Central Brain (CB)
Optic Lobe (OL)
Optic Lobe (OL)
Visual sensory
Ventral Nerve Cord (VNC)
Movement
C
Ventral Nerve Cord (VNC)
Central Brain (CB)
Optic Lobes (OL)
Species
PC2: 26% variance
PC1: 33% variance
Brain region
group
mau:CB
mau:OL
mau:VNC
mel:CB
mel:OL
mel:VNC
sim:CB
sim:OL
sim:VNC
D
E
p-adjusted<0.05
p-adjusted<0.01
Number of Genes
Expression
Upregulated
Downregulated
mau vs sim_OL
mau vs sim_CB
mau vs sim_VNC
sim vs mel_OL
sim vs mel_CB
sim vs mel_VNC
mau vs mel_OL
mau vs mel_CB
mau vs mel_VNC

Figure 2

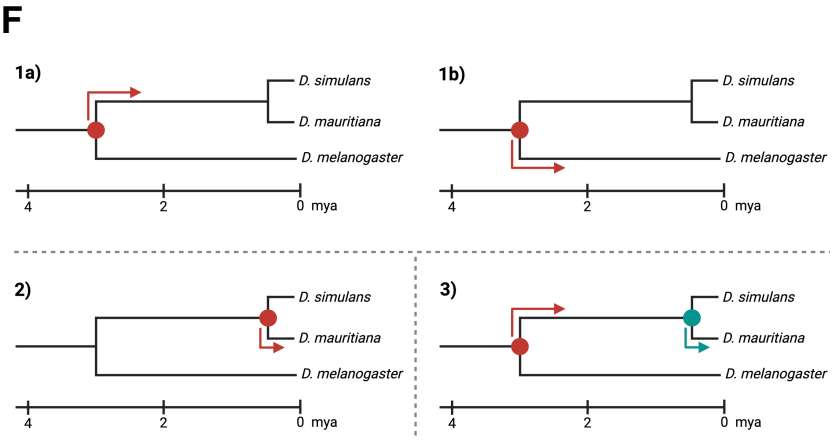


E

	D. mel	*D. sim*	*D. mau*	Set operations strategy
1)	Ref.	✓	✓	(A ∩ B) \ C
2)	Ref.	x	✓	B ∩ C
3)	Ref.	✓	x	A ∩ C #
4)	Ref.	x	x	-
5)	Ref.	✓		A
6)	Ref.		✓	B

✓ has phenotype relative to *D. mel* reference

x no phenotype relative to *D. mel* reference



Supporting Information Text
Extended Methods
Animal stocks and maintenance. *Drosophila* fruit flies were raised on molasses-based fly food Recipe J in narrow vials (Lab Express, Michigan, USA). The following *Drosophila* lines were used in this study:

Species	Stock ID	Origin	Source	Figure 2A
Drosophila melanogaster	Dickinson Lab (DL)	Pasadena, California, USA	Collected by M.H. Dickinson	D. mel USA #1 ***
Drosophila melanogaster	Oregon-R	Oregon, USA	Bloomington *Drosophila* Stock Center, Indiana University Bloomington	D. mel USA #2
Drosophila simulans	D. sim 14021-0251.194	Winters, California, USA	National *Drosophila* Species Stock Center, Cornell University	D. sim USA #1 *** #
Drosophila simulans	D. sim 14021-0251.216	Winters, California, USA	National *Drosophila* Species Stock Center, Cornell University	D. sim USA #2
Drosophila simulans	MD06	Madagascar	Collected by B. Ballard in 2002	D. sim Madagascar #1
Drosophila simulans	MD221	Madagascar	Collected by B. Ballard in 2002	D. sim Madagascar #2
Drosophila simulans	NS37	Nairobi, Kenya	Collected by P. Andolfatto in 2006	D. sim Kenya #1
Drosophila simulans	NS67	Nairobi, Kenya	Collected by P. Andolfatto in 2006	D. sim Kenya #2
Drosophila mauritiana	D. mau(1) SD 14021-0241.01	Mauritius	National *Drosophila* Species Stock Center, Cornell University	D. mau Mauritius #1 ***
Drosophila mauritiana	R10	Mauritius	-	D. mau Mauritius #2
Drosophila mauritiana	R13	Mauritius	-	D. mau Mauritius #3
Drosophila mauritiana	R27	Mauritius	-	D. mau Mauritius #4

*** Lines used to generate bulk RNA-seq dataset.

I incidentally discovered that the *D. simulans* 14021-0251.194 line (*D. simulans* USA Line #1) harbors a 4.6 kb retrotransposon insertion within the first exon of *Irk3* (inwardly rectifying potassium channel), likely resulting in a null allele. Targeted PCR-based genotyping of the *Irk3* locus indicates that the insertion is a Long Terminal Repeat (LTR) retrotransposon belonging to the Ty1-*copia*/Pseudoviridae family (classification by InterPro (1), EMBL-EBI). Correspondingly, no *Irk3* transcript expression was detected for *D. simulans* in the RNA-seq dataset. Targeted PCR-based genotyping of the *Irk3* locus confirmed that this structural variant was not present in the second *D. simulans* 14021-0251.216 line isolated from Winters, California that was also assayed for behavior in this study. Of relevance, the retrotransposon-mediated *Irk3* knockout in *D. simulans* USA Line #1 does not result in a locomotor activity behavior difference relative to other *D. simulans* lines (Figure 2A).

Locomotor activity behavior assay. Three to five-day old adult female flies raised at 22.5-23.0°C and 30-35% humidity on a 12-hr light/12-hr dark cycle were assayed. Locomotor activity behavior assays were performed using a DAM5H *Drosophila* activity monitor device (TriKinetics Inc., Waltham MA) connected to a computer running the DAMSystem3 data collection program. Each monitor device can assay up to 32 flies in parallel. Fly movement within each glass assay tube was recorded as the interruption of fifteen equally spaced infrared beams. Each assay tube was sealed on each end using small plugs fashioned from Kimwipes. The number of times each fly moved from one beam to another (i.e. beam transitions) was tallied across the entire trial duration for each animal.

Flies were cold anesthetized on ice prior to loading into glass assay tubes and were given 15 minutes of recovery time in the device followed by 30 minutes of locomotor activity recording time. All assays were performed during the four-hour duration before incubator lights out. Data collection was performed at 24-25°C and 35% humidity. Representatives of each *Drosophila* sibling species were always assayed on each day of experiments. Data across all trials were pooled for statistical analysis. Statistical analyses were performed using the SciPy and scikit-posthocs packages in Python.

Cross-species brain transcriptomics. Brains were dissected from three to five-day old adult female flies raised at 22°C and 50% humidity on a 16-hr light/8-hr dark cycle. For each sibling species, ten whole brains were dissected per sample (a, b, or c) and their optic lobes, central brain, and ventral nerve cords were sub-dissected and pooled in Trizol (Invitrogen #15596020) for lysis followed by total RNA extraction. cDNA was generated using the same buffer and enzymes as described previously (2).

For each sample, 1 ng of purified total RNA was reverse transcribed using:

- RT primer (5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT VN) and
- TSO primer (/5'Biosg/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACArGrGrG)

PCR amplification was performed using primers:

- R1: 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG
- R2: 5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G

cDNA was purified with Ampure XP Beads (0.6x ratio; Beckman Coulter) followed by quantification using Qubit and bioanalyzer. Sequencing libraries were generated using a Nextera XT kit (Illumina #15027987) and sequenced on an Illumina Nextseq 2000 platform, with read lengths of 75 bp for read 1 and read 2; 10 bp for i7 and i5. Sequencing adapters were trimmed from reads using Cutadapt prior to alignment the *D. melanogaster* reference genome with STAR (3). The resulting transcript alignments were then passed to RSEM to generate gene expression counts (4). Differential gene expression analysis was performed using DESeq2 in R. Two pre-identified outlier samples, 'mau_VNC_b' and 'sim_OL_a', are removed for all differential gene expression analyses.

Targeted PCR-based genotyping. Flies were homogenized with a pipette tip and genomic DNA was extracted using a QIAwave DNA Blood & Tissue Kit (Qiagen #69554). The *Irk3* locus was PCR amplified using genomic DNA as template, and the PCR product was purified using a DNA Clean & Concentrator-5 Kit (Zymo #D4003T). The following genotyping primers (Integrated DNA Technologies, 25 nmole DNA oligonucleotides) were used: ACATATACGAGCGTGACATGAC (for) and GATAGCTTCATGCGCGAGTTC (rev). Amplicon sequencing was performed by Plasmidsaurus, Inc.

Dataset S1. Raw read counts for all 27 bulk RNA-seq samples. Related to Figures 1C and 1D.

Dataset S2. Number of differentially expressed genes for each pairwise species comparison per brain region, p-adjusted<0.05. Related to Figure 1E.

Dataset S3. Number of differentially expressed genes for each pairwise species comparison per brain region, p-adjusted<0.01. Related to Figure 1E.

Dataset S4. Beam transition counts for locomotor activity behavior assays. Related to Figure 2A.

Dataset S5. Differentially expressed genes in Optic Lobes for *D. simulans* vs *D. melanogaster*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S6. Differentially expressed genes in Optic Lobes for *D. mauritiana* vs *D. melanogaster*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S7. Differentially expressed genes in Optic Lobes for *D. mauritiana* vs *D. simulans*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S8. Differentially expressed genes in Central Brain for *D. simulans* vs *D. melanogaster*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S9. Differentially expressed genes in Central Brain for *D. mauritiana* vs *D. melanogaster*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S10. Differentially expressed genes in Central Brain for *D. mauritiana* vs *D. simulans*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S11. Differentially expressed genes in Ventral Nerve Cord for *D. simulans* vs *D. melanogaster*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S12. Differentially expressed genes in Ventral Nerve Cord for *D. mauritiana* vs *D. melanogaster*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S13. Differentially expressed genes in Ventral Nerve Cord for *D. mauritiana* vs *D. simulans*, p-adjusted<0.05 and $|\log_2 FC|>2.0$. Related to Figure 2D.

Dataset S14. Concatenated lists of differentially expressed gene names from Datasets S5 to S13. Related to Figure 2D.

Dataset S15. Number of candidate neuronal genes per brain region for each set and after performing set operations strategy on Dataset S14. Related to Figure 2D.

Software S1. R code for DESeq2 analysis and visualization for all bulk RNA-seq samples. Related to Figure 1C and 1D.

Software S2. R code for DESeq2 analysis of region-specific bulk RNA-seq samples. Related to Figure 1E and 2D.

Software S3. Python script for plotting behavior data and statistical analysis. Related to Figure 2A.

Software S4. R script for map visualization of *Drosophila* species line geographical origins. Related to Figure 2B.

Software S5. Python script for implementing set operations strategies. Related to Figure 2D and 2E.

SI References

1. M. Blum *et al.*, InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* **53**, D444-D456 (2024).
2. M. S. Cembrowski *et al.*, Dissociable Structural and Functional Hippocampal Outputs via Distinct Subiculum Cell Classes. *Cell* **173**, 1280-1292.e1218 (2018).
3. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
4. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).