SÃO PAULO STATE UNIVERSITY– UNESP

Institute of Biosciences, Humanities and Exact Sciences - São José do Rio Preto Campus

DANIEL SIQUEIRA DE OLIVEIRA

CONTRIBUTION OF TRANSPOSABLE ELEMENTS TO THE HOST SHIFT OF CACTOPHILIC *Drosophila* SPECIES

São José do Rio Preto

2024

DANIEL SIQUEIRA DE OLIVEIRA

CONTRIBUTION OF TRANSPOSABLE ELEMENTS TO THE HOST SHIFT OF CACTOPHILIC *Drosophila* SPECIES

Thesis presented under Joint PhD Program between São Paulo State University (UNESP), Institute of Biosciences, Humanities and Exact Sciences of São José do Rio Preto Campus, and Université Claude Bernard Lyon 1 (UCBL), for graduating as PhD in Biosciences (UNESP) and Biomath, Bioinfo, and Evolutionary Biology (UCBL)

Concentration Area: Genetics and Evolutionary Biology

Supervisor: Prof. Dr Claudia Marcia Aparecida Carareto

Supervisor: Prof. Dr Cristina Vieira

São José Do Rio Preto – Brazil

Lyon - France

2024

O48c

Oliveira, Daniel Siqueira de

Contribution of transposable elements to the host shift of cactophilic Drosophila species/ Daniel Siqueira de Oliveira. -- São José do Rio Preto, 2024

139 p. : il.

Tese (doutorado com dupla titulação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto e Université de Lyon.

Supervisora: Claudia Marcia Aparecida

Supervisora: Cristina Vieira

1. Elementos de Transposição. 2. Mudança de hospedeiro. 3. Adaptação 4. Transcritos quiméricos I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

DANIEL SIQUEIRA DE OLIVEIRA

CONTRIBUTION OF TRANSPOSABLE ELEMENTS TO THE HOST SHIFT OF CACTOPHILIC *Drosophila* SPECIES

Thesis presented under Joint PhD Program between São Paulo State University (UNESP), Institute of Biosciences, Humanities and Exact Sciences of São José do Rio Preto Campus, and Université Claude Bernard Lyon 1 (UCBL), for graduating as PhD in Biosciences (UNESP) and Biomath, Bioinfo, and Evolutionary Biology (UCBL).

Concentration Area: Genetics and Evolutionary Biology

Defense date: 13/12/2024

Jury:

_____
Prof. Dr Claudia Marcia Aparecida Carareto
UNESP – Institute of Biosciences, Humanities and Exact Sciences - São José do Rio Preto

_____
Prof. Dr Cristina VIEIRA
Université Claude Bernard - Lyon 1

_____
Prof. Dr Josefa GONZÁLEZ Pérez
Institut Botànic de Barcelona, CSIC, CMCNB

_____
Prof. Dr Élgion Lúcio da Silva LORETO
Santa Maria federal University

_____
Prof. Dr Lilian MADI-RAVAZZI
UNESP – Institute of Biosciences, Humanities and Exact Sciences - São José do Rio Preto

_____
Prof. Dr Emmanuel DESOUHANT
Université Claude Bernard - Lyon 1

_____
Prof. Dra Patrícia BELDADE
University of Lisbon

_____
Prof. Dr. Gabriel da Luz WALLAU
Oswaldo Cruz Foundation

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée par
l'Université Claude Bernard – Lyon 1

École doctorale 341:
Évolution Écosystèmes, Microbiologie, et Modélisation (E2M2)

Spécialité de doctorat:
Biomath – Bioinfo – Génomique Évolutive

À soutenir publiquement à São José do Rio Preto, São Paulo, Brésil le 13/12/2024, par: **Daniel Siqueira de Oliveira**

Contribution des éléments transposables au changement d'hôte des espèces de *Drosophila* cactophiles

---

Devant le jury compose de:

VIEIRA, Cristina, Professeure à l'Université Claude Bernard - Lyon 1 (France): Directrice de thèse

CARARETO, Claudia Marcia Aparecida, Professeure à Universidade estadual Paulista (Brésil): Directrice de thèse

GONZÁLEZ, Josefa Pérez, Chercheure à Institut Botànic de Barcelona, CSIC, CMCNB. (Espagne): Examinatrice

LORETO, Élgion Lúcio da Silva, Professeur à Universidade federal de Santa Maria (Brésil): Examinateur

MADI-RAVAZZI, Lilian, Professeure à Universidade estadual Paulista (Brésil): Presidente du jury/Examinatrice

DESOUHANT, Emmanuel, Professeur à l'Université Claude Bernard - Lyon 1 (France): Examinateur

BELDADE Patrícia, Professeure à Universidade de Lisboa (Portugal): Rapporteure

WALLAU Gabriel da Luz - Fundação Oswaldo Cruz (Brésil): Rapporteur

AGRADECIMENTOS

Agradeço profundamente à minha mãe, Marinês Siqueira de Oliveira, e ao meu pai, João Alfredo Maciel de Oliveira, que sempre acreditaram no meu potencial e nunca mediram esforços para me apoiar na busca pelos meus objetivos. A integridade que testemunhei em cada um de vocês e a educação que me proporcionaram foram fundamentais para a realização desta tese. Sem o exemplo de perseverança, ética e dedicação que recebi em casa, esta trajetória não teria sido possível. Agradeço também aos meus irmãos, Marielle Oliveira e Émerson Oliveira, que sempre estiveram ao meu lado me apoiando em cada etapa do meu desenvolvimento acadêmico. À família, sou imensamente grato por tudo o que me ensinaram e por terem sido meus grandes incentivadores ao longo deste caminho.

À minha companheira de todos os momentos, Camila Militz Pereira, minha profunda gratidão por sempre estar ao meu lado, me apoiando incondicionalmente sem hesitar em acreditar em mim. Obrigado pelo carinho, amor e incentivo constantes. Mesmo nos momentos mais desafiadores, obrigado por ser minha força e o meu refúgio. O seu suporte incessante foi imprescindível para nesta jornada.

Não há palavras que possam descrever precisamente minha gratidão pelas minhas excelentíssimas orientadoras, professoras Cristina Vieira e Claudia Carareto. Vocês me ensinaram o verdadeiro significado de excelência profissional. Agradeço profundamente pela dedicação do seu tempo e paciência, sempre balanceando uma orientação cuidadosa com estímulo à minha autonomia. A nítida satisfação que ambas possuem em fazer ciência foi o combustível para que eu pudesse desenvolver este trabalho. Agradeço imensamente por todas as oportunidades, ensinamentos, incentivo e confiança desde o começo do desenvolvimento desta tese. Vocês sempre serão uma grande inspiração para mim.

Agradeço também aos tantos amigos que tive o prazer de conhecer no ambiente acadêmico. Obrigado Maryanna Simão por ter facilmente me convencido a trabalhar com o modelo de *D. mojavensis*. Agradeço a amizade, apoio e parceria com os colegas da UNESP, William Nunes, Bianca Manfré e Ana Pires, foi enriquecedor dividir o lab com vocês. Aos amigos da UCBL, Miriam Merenciano, Camille Mayeux e Marius Poulain, obrigado por tornarem minha vida em Lyon muito mais leve. O "*best office*" do edifício Mendel sempre será lembrado com muito carinho. Aos demais amigos e colegas de Lyon, Anaïs Larue, Chloe Garambois, Tomás Carrasco e Alicia Ramirez, muito obrigado pelo convívio e gestos de apoio e encorajamento. Vocês foram um pilar essencial na minha formação, trazendo leveza e alegria ao longo do caminho.

Aos PIs que conheci em Lyon e tive o prazer de dividir o ambiente de trabalho, Rita Rebollo, Marie Fablet e Matthieu Boulesteix, obrigado pela maneira excepcional que vocês possuem em compartilhar o seu valioso conhecimento científico. Além disso, é notável e exemplar como vocês trabalham sinergicamente com estudantes. Aprendi muito sobre *soft skills* de maneira indireta, apenas observando-os, um aprendizado que certamente levarei para o meu futuro profissional.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Agradeço à FAPESP pelo apoio financeiro, concedido por meio do Processo nº 2020/06238-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); ao CNPq, sob o processo 308020/2021-9; à CAPES sob processo 88887.716810/2022-00, e à Agence Nationale de la Recherce (ANR) sob o processo ANR-14-CE19-0016-01. Agradeço a CAPES e ao Programa Eiffel (Campus France) pela concessão de bolsas de estudo para internacionalização, que foram de grande valia para a realização deste trabalho.

ABSTRACT

Transposable elements (TEs) are dynamic components of genomes that drive structural variation and serve as sources of genetic diversity, contributing significantly to evolutionary processes. During transcription, TEs can interact with genes incorporating their sequence into gene's mRNA through transcription initiation/termination and alternative splicing, named as gene-TE chimeric transcripts. In this thesis, we developed ChimeraTE, a novel bioinformatics pipeline designed to detect gene-TE chimeras, using paired-end RNA-seq data through two complementary modes: a reference-guided mode for canonical genome alignment and a reference-free mode that identifies chimeras without the need for a reference genome. Using *Drosophila melanogaster* strains, we validated ChimeraTE and discovered that approximately 1.12% of genes generate chimeric transcripts, with around 23% of these TEs absent from the reference genome, suggesting recent or polymorphic insertions contributing to transcriptome variability. We used this approach to investigate the role of TEs in the adaptive evolution and host-shifting of cactophilic *Drosophila* species, focusing on their interactions with genes associated with host location, acceptance, and usage. Since host shift is a crucial evolutionary process that can lead to reproductive isolation and speciation, we conducted genomic and transcriptomic analyses across seven cactophilic *Drosophila* genomes to explore how TEs contribute to host-shift-related gene regulation and expression. We identified an enrichment of TEs at promoter regions of host shift-associated genes, with *Helitrons* representing about 60% of these cases, indicating a likely *cis*-regulatory role of TEs in influencing gene expression. Differential expression analyses in species with different host preferences revealed notable divergence in gene expression profiles in larvae and head tissues, which are critical to host recognition and adaptation. Although the overall impact of TEs on gene expression levels appears minimal, approximately 1.31% of genes, including those linked to host preference, generate chimeric transcripts, pointing out a direct influence of TEs on transcript diversity. The findings provide compelling evidence that TEs play a role in the adaptive evolution of host shift, fostering genetic and transcriptomic changes that may facilitate ecological specialization and reproductive isolation. This thesis underscores the evolutionary significance of TEs in promoting genetic novelty and highlights their potential to drive adaptive divergence in natural populations of cactophilic *Drosophila*.

Key-words: Transposable elements; Host shift; Adaptation; Chimeric transcripts

RESUMO

Elementos de transposição (TEs) são componentes dinâmicos dos genomas que geram variantes estruturais e servem como fontes de diversidade genética, contribuindo significativamente para processos evolutivos. Durante a transcrição, os TEs podem interagir com genes, incorporando sua sequência no mRNA do gene por meio de iniciação/terminação da transcrição e *splicing* alternativo, formando os chamados transcritos quiméricos gene-TE. Nesta tese, desenvolvemos o ChimeraTE, uma nova *pipeline* de bioinformática projetada para detectar quimeras gene-TE, utilizando dados de RNA-seq *paired-end* por meio de dois modos complementares: um modo guiado por referência para alinhamento genômico canônico e um modo independente de referência, que identifica quimeras sem necessidade de um genoma de referência. Utilizando linhagens de *Drosophila melanogaster*, validamos o ChimeraTE e descobrimos que aproximadamente 1,12% dos genes geram transcritos quiméricos, com cerca de 23% destes TEs ausentes do genoma de referência, sugerindo que inserções recentes ou polimórficas que contribuem para a variabilidade do transcriptoma. Utilizamos essa abordagem para investigar o papel dos TEs na evolução adaptativa e na troca de hospedeiro de espécies cactofílicas de *Drosophila*, com foco em suas interações com genes associados à localização, aceitação e uso de hospedeiros. Considerando que a mudança de hospedeiro é um processo evolutivo que pode levar ao isolamento reprodutivo e à especiação, conduzimos análises genômicas e transcriptômicas em sete genomas de espécies de *Drosophila* cactofílicas para explorar como os TEs contribuem para a regulação e expressão de genes relacionados à mudança de hospedeiro. Identificamos um enriquecimento de TEs em regiões promotoras de genes associados à mudança de hospedeiro, sendo que os *Helitrons* representam cerca de 60% desses casos, indicando um provável papel *cis*-regulador dos TEs na influência sobre a expressão gênica. Análises de expressão diferencial entre espécies com diferentes preferências de hospedeiro revelaram uma notável divergência nos perfis de expressão gênica em tecidos de larvas e cabeça, que são essenciais para o reconhecimento e adaptação ao hospedeiro. Embora o impacto geral dos TEs nos níveis de expressão gênica pareça ser mínimo, aproximadamente 1,31% dos genes, incluindo aqueles ligados à preferência de hospedeiro, geram transcritos quiméricos, apontando para uma influência direta dos TEs na diversidade de transcritos. Os resultados fornecem evidências convincentes de que os TEs desempenham um papel na evolução adaptativa da troca de hospedeiro, promovendo mudanças genéticas e transcriptômicas que podem facilitar a especialização ecológica e o isolamento reprodutivo. Esta tese destaca a importância evolutiva dos TEs na promoção de novidades genéticas e

enfatiza seu potencial em impulsionar a divergência adaptativa em populações naturais de *Drosophila* cactofílicas.

Palavras-chave: Elementos de Transposição; Mudança de hospedeiro; Adaptação; Transcritos quiméricos.

RÉSUMÉ

Les éléments transposables (TEs) sont des composants dynamiques des génomes qui génèrent des variantes structurales et servent de sources de diversité génétique, contribuant de manière significative aux processus évolutifs. Lors de la transcription, les TEs peuvent interagir avec des gènes en intégrant leur séquence dans l'ARNm du gène par des mécanismes d'initiation/terminaison de la transcription et d'épissage alternatif, formant ainsi des transcrits chimériques gene-TE. Dans cette thèse, nous avons développé ChimeraTE, une nouvelle pipeline bioinformatique conçue pour détecter les chimères gene-TE, utilisant des données de RNA-seq paired-end à travers deux modes complémentaires : un mode guidé par référence pour l'alignement génomique canonique et un mode indépendant de référence, qui identifie les chimères sans besoin de génome de référence. En utilisant des lignées de Drosophila melanogaster, nous avons validé ChimeraTE et découvert qu'environ 1,12% des gènes produisent des transcrits chimériques, avec environ 23% de ces TEs absents du génome de référence, suggérant des insertions récentes ou polymorphiques contribuant à la variabilité du transcriptome. Nous avons utilisé cette approche pour étudier le rôle des TEs dans l'évolution adaptative et la transition d'hôte des espèces de Drosophila cactophiles, en nous concentrant sur leurs interactions avec les gènes associés à la localisation, l'acceptation et l'utilisation de l'hôte. Étant donné que le changement d'hôte est un processus évolutif crucial pouvant conduire à l'isolement reproductif et à la spéciation, nous avons mené des analyses génomiques et transcriptomiques sur sept génomes d'espèces de Drosophila cactophiles pour explorer comment les TEs contribuent à la régulation et à l'expression des gènes liés au changement d'hôte. Nous avons identifié un enrichissement de TEs dans les régions promotrices des gènes associés au changement d'hôte, les Helitrons représentant environ 60% de ces cas, ce qui indique un rôle cis-régulateur probable des TEs dans l'influence sur l'expression génique. Des analyses d'expression différentielle entre des espèces avec différentes préférences d'hôte ont révélé une divergence notable dans les profils d'expression génique dans les tissus de larves et de tête, essentiels à la reconnaissance et à l'adaptation à l'hôte. Bien que l'impact global des TEs sur les niveaux d'expression génique semble minime, environ 1,31% des gènes, y compris ceux liés à la préférence d'hôte, génèrent des transcrits chimériques, soulignant une influence directe des TEs sur la diversité des transcrits. Les résultats fournissent des preuves convaincantes que les TEs jouent un rôle dans l'évolution adaptative liée au changement d'hôte, en favorisant des modifications génétiques et transcriptomiques qui peuvent faciliter la spécialisation écologique et l'isolement reproductif. Cette thèse souligne l'importance

évolutive des TEs dans la promotion de nouvelles variations génétiques et met en avant leur potentiel à stimuler la divergence adaptative dans les populations naturelles de Drosophila cactophiles.

Mots clés: Eléments transposable; Changement de hôte ; Adaptation ; Transcripts chimériques.

TABLE OF CONTENTS

INTRODUCTION

Transposable elements as engines of molecular evolution

Transposable elements (TEs) are short DNA sequences capable of moving from one genomic site to another. They have been identified in virtually all eukaryotic genomes studied to date. In insects, their abundance ranges from 0.12% in the Antarctic midge *B. antartica* (KELLEY et al., 2014) to 72% in the weevil *Sitophilus oryzae* (PARISOT et al., 2021). TEs are classified into two major classes based on their transposition mechanism, and further divided into respective families and subfamilies based on their structure and sequence homology (WICKER et al., 2007). Class I elements, or retrotransposons, transpose via an RNA intermediate and the reverse transcriptase enzyme through a process known as "copy and paste". These include long terminal repeat (LTR)/endogenous retrovirus (ERV) elements, non-LTR retrotransposons such as long interspersed nuclear elements (LINEs), and non-autonomous elements like short interspersed nuclear elements (SINEs). Class II elements, or DNA transposons, move by a "cut and paste" mechanism facilitated by the transposase enzyme.

TEs were first discovered by Barbara McClintock in the 1950s (MCCLINTOCK, 1951), and for many years, they were regarded by the scientific community as "junk DNA" (OHNO, 1972). Their ability to move across the genome led to their classification as "genomic parasitic elements". The mutations caused by transposition can have severe impacts on host fitness, particularly if they disrupt genes. The insertions that remain in the genome are likely to be neutral, mildly deleterious and rarely become adaptive (MONTGOMERY; CHARLESWORTH; LANGLEY, 1987). Since mobilization is a mutagenic process producing structural variants, several host mechanisms have evolved to prevent it. These mechanisms include chromatin remodeling factors, DNA methylation, and small RNAs (SLOTKIN; MARTIENSSEN, 2007). For example, in *Drosophila melanogaster* germline, TEs are targeted by piwi-interacting RNAs (piRNAs), which not only mediate the cleavage of TE transcripts but also promote the deposition of repressive chromatin marks at the TE insertion sites, effectively preventing further transcription (FABRY et al., 2021). In *Arabidopsis thaliana*, most TEs are silenced through cytosine methylation, preventing their activation and spread. However, environmental or genomic stress can sometimes reactivate TEs, creating bursts of transposition that introduce genetic variability (ROQUIS et al., 2021).

Over the past two decades, advances in biotechnology, genomic analyses and the extensive sequencing of genomes have accumulated evidence about the role of TEs in genome evolution. Functional genomic studies have demonstrated that TEs not only influence gene

transcription but are also essential for maintaining chromatin structure due to their ability to recruit chromatin proteins (HORVÁTH; MERENCIANO; GONZÁLEZ, 2017; REBOLLO; ROMANISH; MAGER, 2012). Consequently, TEs have been implicated in several key evolutionary processes, such as the structural evolution of genomes through chromosomal rearrangements (HUGHES; COFFIN, 2001; KAZAZIAN, 2004); co-option of regulatory sequences that control nearby gene expression (JORDAN et al., 2003); integration into coding sequences at the transcript level (GOTEA; MAKALOWSKI, 2006; LIPATOV et al., 2005; PIRIYAPONGSA et al., 2007; WANG et al., 2007); alterations in the transcriptome driven by environmental changes (HORVÁTH; MERENCIANO; GONZÁLEZ, 2017; MOSCHETTI et al., 2020; RECH et al., 2019); and the recruitment of TE-derived genes by host genomes for essential functions (VOLFF, 2006).

The recurrent interaction between genes and transposable elements

The specific location of a TE insertion in the genome often determines its evolutionary fate in a population. In response, many TEs have preferential insertional sites to integrate themselves into genomic regions where their propagation is more likely to succeed. It is well-known that TEs are not randomly distributed in the genome, being often found at gene-poor regions such as heterochromatin, pericentromeric, and telomeric regions (BOURQUE et al., 2018). These regions also have low recombination rates, which minimizes the harmful effects that TEs can have by reducing the likelihood of disrupting essential genes (BOURGEOIS; BOISSINOT, 2019). In the arms race against host control, some TEs also show a tendency to insert themselves in euchromatic regions. For instance, the retrotransposons in mold and yeast have evolved to target regions upstream of RNA polymerase III-transcribed genes (CHEUNG; MANHAS; MEASDAY, 2018). This strategy allows them to be transcribed independently without disrupting host gene expression, ensuring their own propagation while minimizing damage to the host genome (SPALLER et al., 2017). The insertion site nearby genes can also persist without a propagation-related cause. In the clonal raider ant, *Ooceraea biroi*, TEs are enriched nearby odorant receptor (OR) genes, possibly facilitating gene duplications via unequal crossing over (MCKENZIE; KRONAUER, 2018). This expansion of OR duplicates in ants is associated to pheromone detection, and it may be crucial for the evolution of complex social behaviors.

Silenced TEs tend to accumulate mutations over time, losing their mobility but potentially keeping functional protein domains/motifs (MILLER et al., 2000). Due to such potential, a TE copy inserted nearby a host gene may cause alterations on gene expression or

alternative splicing (CHUNG et al., 2007; HIRSCH; SPRINGER, 2017; REBOLLO; FARIVAR; MAGER, 2012). The evolutionary outcomes derived from successful gene-TE interactions are diverse and well documented in many taxa (SINZELLE; IZSVÁK; IVICS, 2009; VOLFF, 2006). Long-term stable and adaptive interactions are often termed as domestication or exaptation when become essential for the cell fate or functioning (CAPY, 2021). For instance, in the vertebrate immune system, both genes *Rag1* and *Rag2* have conserved TE-derived protein domains that are crucial to the somatic recombination process known as V(D)J (AGRAWAL; EASTMAN; SCHATZ, 1998). They work together as a transposase, capable of excising a DNA segment containing recombination signals from a donor site and inserting it into a target DNA molecule. In *D. melanogaster*, the telomeric LINE elements TART, Heta-A, and Tahre remain actively transposing specifically into telomeres, being essential to maintain telomere's length and protect against replication-induced erosion (DANILEVSKAYA et al., 1994; GEORGE et al., 2006; PARDUE; DEBARYSHE, 2000). Another example is the Syncytin proteins which are originated from the envelope (*env*) genes of ERVs and have been co-opted for placental development in mammals (LAVIALLE et al., 2013; MI et al., 2000), being essential for nutrient exchange between the mother and the developing embryo.

Autonomous TEs encode genes that enable their mobilization, relying on the host cell's machinery to recognize self *cis*-regulatory sequences that effectively mimic host promoters (reviewed on Chuong, Elde, and Feschotte 2017). Many studies have demonstrated spread TE-derived promoters and enhancers (CAO et al., 2019; CHUONG et al., 2013; YE et al., 2020), transcription factor binding sites (TFBSs) (BOURQUE et al., 2008; THORNBURG; GOTEA; MAKAŁOWSKI, 2006), and insulators (WANG et al., 2015). A remarkable example is the insecticide resistance in *D. melanogaster*, which is caused by a co-option of alternative TFBSs from an Accord TE inserted upstream to the *Cyp6g1* gene, causing its overexpression and higher efficiency to metabolize insecticides (DABORN et al., 2002). In *Caenorhabditis sp*., heat shock response is regulated by a specific transcription factor that binds onto heat shock elements, activating the expression of important genes for cellular protection upon stress (VIHERVAARA; DUARTE; LIS, 2018). Notably, a significant number of these elements in several *Caenorhabditis* species are located within *Helitrons*, contributing to the diversification and rewiring heat shock response in *Caenorhabditis* upon stress (GARRIGUES et al., 2019).

Gene-TE chimeric transcripts as evidence of TE-related novelties

The interactions between genes and TEs can generate mRNA isoforms with both exons and TE-derived sequences, named gene-TE chimeric transcripts (CORONADO-ZAMORA; GONZÁLEZ, 2023; LIPATOV et al., 2005; TREIBER; WADDELL, 2020). Chimeras are produced through molecular mechanisms during transcription or mRNA alternative splicing (reviewed on Ayarpadikannan et al. 2015), and classified into three categories accordingly: 1) TE-initiated transcript: the TE is co-opted as an alternative promoter, providing a novel transcription start site to the gene (LAMPRECHT et al., 2010; MCGINNIS; SHERMOEN; BECKENDORF, 1983); 2) TE-exonized transcript: the TE is located in the intron and retained in the mature mRNA through splicing (ALMEIDA et al., 2008; SELA et al., 2010; SOREK, 2007); 3) TE-terminated transcript: the TE provides an alternative polyadenylation site during RNA maturation (DABORN et al., 2002; FARRÉ; ENGEL; ANGULO, 2016). TE-initiated and TE-terminated transcripts are more likely to be involved in changes of gene expression, whereas the TE-exonized transcripts might alter the protein sequence, having a direct effect on the protein function.

Gene-TE chimeras play significant roles in driving evolutionary change by contributing with genetic novelties in populations, potentially leading to adaptive functions in organisms across diverse taxa. In *D. melanogaster*, the gene *CHKov1* produces a TE-exonized transcript with a TE insertion from the Doc family that has high frequency in South American populations (MAGWIRE et al., 2011). This chimera represents a truncated isoform compared to its ancestral sequence, changing the protein function and resulting in resistance to organophosphates and viral infection (MAGWIRE et al., 2011). Another relevant example is the case of the industrial melanism in British peppered moths (*Biston betularia*), where the black (carbonaria) form became dominant during the Industrial Revolution due to a mutation caused by a TE copy inserted in the first intron of the *cortex* gene. The TE is exonized in the carbonaria form, inducing higher expression of *cortex*, which in turn causes the adaptive dark melanization of the moth (HOF et al., 2016). In the oil palm *Elaeis guineensis*, the hypomethylation state of a LINE copy in the intron of the gene *Deficiens* generates a TE-exonized transcript with premature termination, generating poorly developed fruits with mantled phenotype (ONG-ABDULLAH et al., 2015).

In cancer, TEs become active due to a state of global hypomethylation (EHRLICH, 2009). This activation can lead to the formation of new chimeric transcripts with harmful effects, a process known as onco-exaptation (BABAIAN; MAGER, 2016). A well-

characterized example of this phenomenon occurs in Hodgkin Lymphoma, where the oncogene *CSF1R* has an increase on its transcriptional level caused by an alternative promoter from an ERV copy, acting as a trigger to the development of malignant cells (BABAIAN et al., 2016). In large B-cell lymphoma, the *FABP7* gene utilizes an endogenous retrovirus LTR as a promoter, leading to the production of a novel protein that contributes to abnormal cell proliferation (LOCK et al., 2014). Another study has documented over 4,800 chimeric transcripts distributed among 723 cancer-associated genes, mostly from Alu and L1 elements (CLAYTON et al., 2020). More recently, a comprehensive study has identified 1,068 antigens with TE-derived sequences from cancer samples, named ad TE-chimeric antigens (SHAH et al., 2023). Importantly, these specific chimeras produced functional proteins present on the surface of cancer cells, with appeal for therapeutical exploitation (FUDGE, 2023).

Transcriptome-wide studies revealing gene-TE chimeric transcripts are scarce in insect species, mainly for the non-model ones. Even in the model *D. melanogaster*, comprehensive analysis with tissue specific samples were assessed recently (CORONADO-ZAMORA; GONZÁLEZ, 2023; OLIVEIRA et al., 2023; TREIBER; WADDELL, 2020). The authors found similar quantitative results, showing that ~1.5% of the genes producing chimeric transcripts, which can be considered low compared with up to 14% in human and mice (FAULKNER et al., 2009). Importantly, the overall chimeric isoform contribution to gene expression in *D. melanogaster* was ~27% in head, gut and ovary, although 314 genes are expressed solely with the chimeric isoform and common in wild-type strains (CORONADO-ZAMORA; GONZÁLEZ, 2023). This result demonstrates the extension of TE co-option in *D. melanogaster*, and the possibility to identify them through the detection of chimeras. Additionally, it has been shown that over 80 TE insertions are adaptive across *D. melanogaster* global populations (RECH et al., 2019), impacting the expression of genes associated with ecologically significant traits. Considering the extensive presence of TEs in virtually all eukaryotes, it is plausible that similar adaptive patterns could be observed in other species. Since alternative splicing has a stronger impact on phenotypic changes than gene expression modifications (HEINZEN et al., 2008; YU et al., 2008), the extent to which gene-TE chimeras contribute to environmental adaptability remains an open question. This influence likely affects the evolutionary fitness of insect populations.

The bioinformatic techniques to identify gene-TE chimeric transcripts has evolved following the advances on sequencing technologies. The first studies demonstrating the presence of TE-derived sequence in mRNA were conducted in human, based on expressed sequence tags (EST) evidence (SOREK; AST; GRAUR, 2002). In *D. melanogaster*, ESTs were

also applied to chimeras' identification, demonstrating that only ~1% of genes produce chimeric transcripts in this species (LIPATOV et al., 2005). Importantly, these findings based on ESTs were limited to identify TE-derived sequences at the 5' UTR of genes, so they produced an underestimation of the transcriptome-wide context. Other studies were conducted using Cap Analysis Gene Expression (CAGE) technique, showing a significant proportion of TE-initiated transcripts, ranging from 3% to 14% in human and mice, depending on the tissue (FAULKNER et al., 2009). In the last years, with the popularization of Illumina RNA-seq, a few bionformatic tools to detect gene-TE chimeric transcripts have been published. The pipeline CLIFinder is able to identify gene-TE chimeras derived specifically from long interspersed nuclear elements (LINEs) restricted to the human genome (PINSON et al., 2018). LIONS is not specific to any taxa, but it is designed to identify only TE-initiated transcripts (BABAIAN et al., 2019). FREDDIE is an algorithm able to identify exonization events involving intragenic TEs in mammalian genomes, through transcriptome assembly, quantification, and assessing the coding potential of chimeras (MERCURI et al., 2024). Importantly, all the three methods rely on the reference genome, therefore they only detect chimeras from TE copies present in the reference, losing the relevant information regarding novel and/or polymorphic TE insertions. TEchim was the only pipeline capable of identifying chimeras from polymorphic TEs (TREIBER; WADDELL, 2020), but it was not designed to run automatically with other species than *D. melanogaster*. Although the proven relevance to study gene-TE chimeric transcripts in the evolutionary biology context, their study is constrained by the pervasive repetitive nature of TE copies and the limited availability of methods for their identification. In this thesis, we developed ChimeraTE, a novel pipeline to identify chimeras derived from TEs regardless their presence on the reference genome, from any species, using Illumina paired-end reads (OLIVEIRA et al., 2023).

Host shift as an evolutionary driving force

The idea that adaptation to different environments can drive speciation has been proposed since Darwin's *On the Origin of Species* (1859). Benjamin Dann Walsh, a 19th-century' entomologist, explored such idea hypothesizing that speciation in phytophagous insects could be driven by shifting to new host plants (WALSH, 1864). Walsh's arguments to the process can be summarized in three key points regarding phytophagous insects. Firstly, most of them have preference to feed on or deposit their eggs in one or a few specific host plants. Secondly, if reared on different host plants, they frequently display variations in observable traits, which he referred to as "phytophagic varieties" (WALSH, 1864). Finally, he

hypothesized that if a “phytophagic variety” remains on the same host plant for multiple generations, it might become so well-adapted to that plant that it could no longer feed on an alternative host, or possibly even mate with individuals that utilize the other plant.

In the last decades, many studies have been conducted with phytophagous insects to understand the relationship between host shift (HS) and reduced gene flow between populations (DRÈS; MALLET, 2002; HEARD, 2012; HERNÁNDEZ-ROLDÁN et al., 2016). The conclusions from many species indicate that reproductive isolation emerges due to a series of adaptations driven by HS, which is the result of divergent selection between distinct ecological environments (FORBES et al., 2017; HOWARD; BERLOCHER, 1998; RUNDLE; NOSIL, 2005; WHITEMAN; PIERCE, 2008). Consequently, natural selection increases the prevalence of alleles that provide fitness advantages during HS. Thus, the molecular effects of HS evolution are strongly associated with ecological divergence, specialization, and, ultimately, ecological speciation (WHITEMAN; PIERCE, 2008). The latter is defined as the speciation process by which reduced gene flow occurs as a consequence of ecologically-based divergent selection (MATSUBAYASHI; OHSHIMA; NOSIL, 2010; SCHLUTER, 2009).

Although ecological speciation is a well-established theory, its extent on nature compared to other speciation mechanisms remains under study (HENDRY, 2009). In non-ecological speciation, random chance plays a crucial role, as seen in processes like speciation through genetic drift, founder events, and population bottlenecks (COYNE; ORR, 2004). These events do not necessarily involve environmental differences driving the divergence. Additionally, non-ecological speciation includes models where natural selection is involved, but this selection is either not tied to environmental differences or not divergent across habitats. Examples of these include speciation through sexual selection (LANDE, 1980), where mate choice drives divergence, and sexual conflict (CHAPMAN et al., 2003), where differences between male and female reproductive strategies lead to speciation. Another scenario involves what is known as “mutation-order speciation” (SCHLUTER, 2009), in which populations under similar environmental pressures accumulate different, incompatible genetic changes, resulting in speciation. An important consideration is that ecological speciation can occur in any geographical arrangement, whether the populations are separated (allopatry), partially overlapping (parapatry), or fully overlapping (sympatry), as long as ecologically-based divergent selection is the driving force behind their evolutionary divergence (COYNE; ORR, 2004; RUNDLE; NOSIL, 2005).

Cactophilic Drosophila species: A model for host shift studies

The *Drosophila* genus has been an important model organism for evolutionary studies regarding the evolution of HS (MARKOW, 2019). The diversification of *Drosophila* species and wide-range of plant hosts provides an ideal opportunity to understand the process of host shifting, and its potential in the plasticity that enables organisms to survive upon climatic and environmental changes. In the subgenus *Sophophora*, species from the *melanogaster* group such as *D. simulans* and *D. melanogaster* are cosmopolitan and have expanded their distribution beyond their ancestral regions through commensal association with humans (LACHAISE; SILVAIN, 2004). *D. sechellia* is a specialist species in *Morinda citrifolia* fruits, which are considered toxic hosts for other species (R'KHA; CAPY; DAVID, 1991). Another specialist, *D. erecta*, has its reproductive cycle in *Pandanus* species (LACHAISE; SILVAIN, 2004). In the *obscura* group, *D. persimilis* and *D. pseudoobscura* have a geographic distribution mainly in the western part of North America, although a small population of *D. pseudoobscura* is found in Bogotá, Colombia (DOBZHANSKY et al., 1964). These species are classified as opportunists, a characteristic of generalist species, as both are abundant at medium and high elevations during the summer, while in winter, these populations migrate to lower elevations near desert habitats throughout their geographic range. These regions are not occupied by these species during the summer. Depending upon availability, *D. pseudoobscura* and *D. persimilis* use mudflows, domestic fruits, cacti, and agave as feeding and breeding sites (POWELL, 1997). *D. willistoni* is considered one of the most numerous and widely distributed *Drosophila* species in the New World (AYALA; POWELL; DOBZHANSKY, 1971), and it uses a wide range of decomposing fruits as hosts. In the subgenus *Drosophila*, *D. virilis* is a species distributed in the Holarctic zone of the globe, which reproduces and feeds in mudflows and decomposing bark (THROCKMORTON, 1982). In the Hawaiian group, *D. grimshawi* is considered a generalist, utilizing decomposing bark from more than seven endemic Hawaiian plant families (MAGNACCA; O'GRADY, 2006).

The evolution of cactophilic *Drosophila* species from the *repleta* group is closely associated with the transition from the use of fermented fruits in humid forests to fleshy-stemmed plants adapted to arid climates, such as cacti, particularly *Opuntia sp*. The vast majority of *repleta* species use fermenting cactus tissues to complete their life cycles in semi-arid or arid environments (RUIZ; HEED; WASSERMAN, 1990). The use of decomposing cacti as hosts in arid areas provides abundant and sustainable resources for *Drosophila* larvae and adults. Nevertheless, specialization in cacti represents a significant ecological challenge,

requiring specific adaptations for cactus-dwelling species. For instance, cacti have lower nitrogen and phosphorus content compared to hosts used by most other *Drosophila* species (OLIVEIRA et al., 2012). Additionally, cacti produce various toxic substances for most insects, an important adaptation that reduces damage caused by herbivores (CLOUDSLEY-THOMPSON, 1996). Finally, cacti are generally found in areas where temperatures are high, imposing desiccation and thermal stress on organisms associated with them (MATZKIN; MARKOW, 2009). It is believed that the ecological shifts towards using cacti as feeding and breeding sites involved numerous adaptive changes in the physiology, reproductive biology, and behavior of cactus-feeding species (MARKOW; O'GRADY, 2007). Therefore, cactophilic species are an excellent system to investigate the genetic basis of HS and adaptation.

A study conducted with 65 species from the *repleta* group showed that the use of cacti from the *Opuntia* genus is common throughout the phylogeny, and the transition to higher complexity columnar cacti occurred independently several times during the evolutionary history of the group (OLIVEIRA et al., 2012). The authors pointed out that it is also common the usage of multiple cacti species, *Opuntia sp.* and columnar. From the phylogenetic context, they demonstrated that the usage of *Opuntia sp*. is a basal condition, while diversification to chemically more complex columnar cacti represent a derived state. In the phylogeny, they reconstructed three scenarios (specific use of *Opuntia*, use of columnar cacti, and polymorphism) that characterize the host shifts can be identified within the subgroups of *repleta*. In the *mojavensis* cluster, for instance, *D. navojoa* is the basal species and it uses exclusively *Opuntia sp*. as its host cactus, while the other two species in the cluster, *D. arizonae* and its sister species *D. mojavensis*, are polymorphic, using both *Opuntia spp*. and columnar cacti (OLIVEIRA et al., 2012). In the *buzzatii* complex, seven out of the ten species have polymorphic usage of cacti. For instance, *D. buzzatii* has preference for *Opuntia sp*., while its sister species *D. koepferae* uses preferentially columnar cacti (Figure 1) (OLIVEIRA et al., 2012). In contrast, species from the *mercatorum* and *hydei* subgroups exclusively use *Opuntia ssp*. as their host (OLIVEIRA et al., 2012).

The pair of species *D. mojavensis–D. arizonae*, which share the use of both types of cacti, differ in many aspects. *D. mojavensis* exhibit genetic, behavioral, ecological, and morphological differences that have led to its classification into four subspecies (*D. m. baja*, *D. m. mojavensis*, *D. m. sonorensis*, and *D. m. wrigleyi*), occurring in different regions in the southwestern United States and western Mexico (Figure 1) (REED; MARKOW, 2004; REED; NYBOER; MARKOW, 2007; ROSS; MARKOW, 2006). Unlike *D. arizonae*, which is polymorphic in its host cactus preference, using both *Opuntia sp.* and columnar cacti throughout

its geographic range, *D. mojavensis* is oligophagous, utilizing *Opuntia sp.* as well as different species of columnar cacti (HOCUTT, 1998; RUIZ; HEED, 1988). *D. m. sonorensis* primarily uses the "organ pipe cactus" (*Stenocereus thurberi*), *D. m. mojavensis* uses the "barrel cactus" (*Ferocactus cylindraceus*), *D. m. wrigleyi* uses various species from the *Opuntia* genus (*Opuntia spp.*), and *D. m. baja* preferentially uses the agria cactus (*Stenocereus gummosus*), despite the presence of all other cacti used by the other subspecies within its distribution range (HOCUTT, 1998; REED; NYBOER; MARKOW, 2006; RUIZ; HEED, 1988).

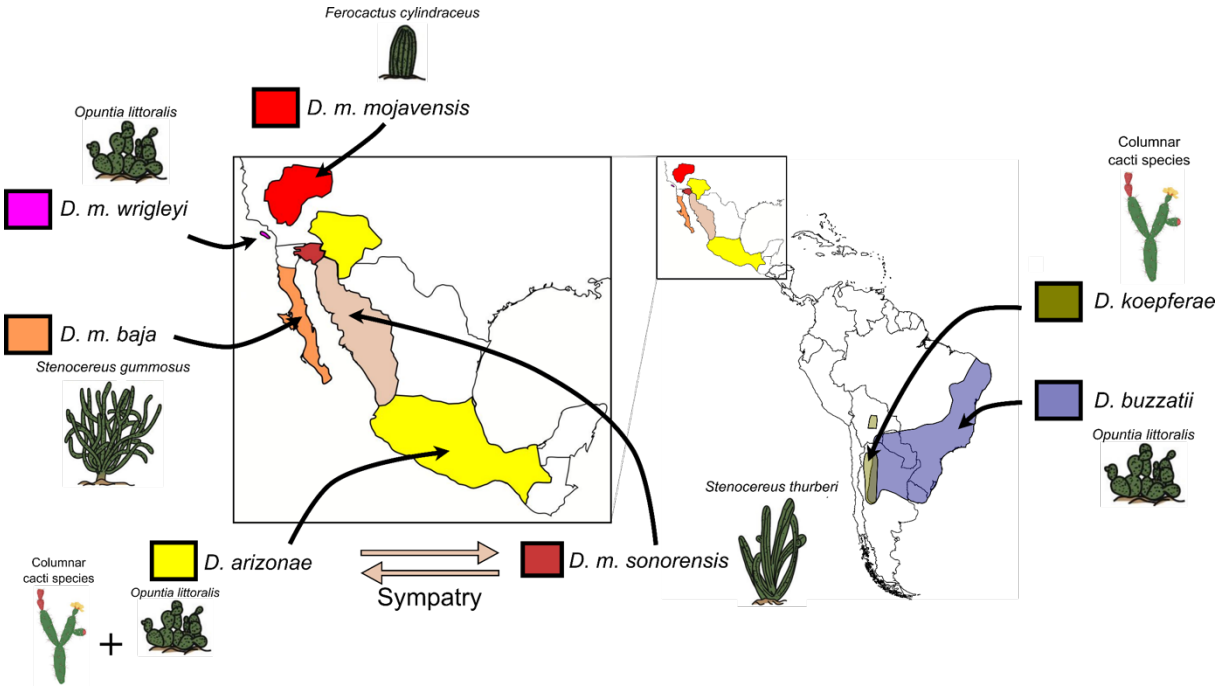


Figure 1: Geographical distribution and cacti preference for *D. buzzatii* and *D. koepferae*, in the South America, and the four *D. mojavensis* subspecies and *D. arizonae*, distributed in the southwestern United States and western Mexico.

These species exhibit incomplete and asymmetric post-zygotic isolation in the hybrid offspring. In crosses from *D. mojavensis* males and *D. arizonae* females, the hybrid males are sterile, whereas in reciprocal crosses, the offspring can be fertile depending on the maternal subspecies (REED; MARKOW, 2004; RUIZ; HEED; WASSERMAN, 1990). In line with the Reinforcement Theory (DOBZHANSKY, 1937), it has been reported that *D. mojavensis* females from sympatric populations exhibit almost complete pre-zygotic reproductive isolation compared to *D. arizonae* males, while *D. mojavensis* females living in allopatry with *D. arizonae* demonstrate lower reproductive isolation (MASSIE; ANN MARKOW, 2006; RUIZ; HEED; WASSERMAN, 1990). Importantly, early stages of reproductive isolation have been observed between *D. m. sonorensis* and *D. m. baja* (MARKOW, 1991; PFEILER et al., 2009; ZOUROS; D'ENTREMONT, 1980). The premating isolation is attributed to differences in

cuticular hydrocarbons, which play a key role in mate recognition signaling pathways (HOWARD; BLOMQUIST, 2005). Notably, experimentally shifting the cactus host was the primary factor leading to changes in hydrocarbon profiles. Therefore, the reproductive barriers are closely tied to the usage of specific hosts (HAVENS; ETGES, 2013; STENNETT; ETGES, 1997).

Factors affecting host shift

The evolution of insects' ability to specialize on particular hosts often requires evolutionary changes in two key aspects: 1) host acceptance behavior: the insect needs to evolve the metabolize and accept the new host as a viable food source; and 2) offspring performance: the offspring must be physiologically capable of surviving and thriving on the new host. In a scenario where oviposition behavior is determined by one genetic locus, and offspring fitness by another, if a new host plant is used for oviposition but is still less suitable for larvae development compared to the ancestral host, then either behavioral avoidance or physiological tolerance to the new host is expected to evolve, but not both simultaneously (CASTILLO-CHAVEZ; LEVIN; GOULD, 1988; GOULD, 1984). Therefore, the evolutionary outcome upon HS is dependent of the genetic background in the population. Behavior and physiology must evolve together, as one alone is not enough for the insect to successfully specialize on a new host (RAUSHER, 1983, 1993). Thus, evolving from a generalist to a specialist is easier than the reverse, because evolving either host acceptance behavior or physiological adaptation alone can allow the insect to specialize (VERTACNIK; LINNEN, 2017).

The success of the plant-insect relationship is the result of a long and complex evolutionary history marked by the interaction between them. The selection of host plants by phytophagous insects depends on both the plant's characteristics and availability, as well as physiological and behavioral traits of the insect (RUIZ; HEED, 1988). Suitability and location are essential properties for oviposition to occur. The location of the host plant is associated with both the plant's abundance, as well as the insect's sensitivity and behavior. Suitability, in turn, is determined by the plant's physicochemical characteristics, which differ mainly in nitrogen and water content, components that are strictly related to larval growth rates. Additionally, secondary metabolites may assist the insect in locating the host or present high levels of toxicity (RUIZ; HEED, 1988). The adaptations that enabled cactus-feeding species to overcome the physicochemical challenges posed by columnar cacti include three main stages, which summarize the establishment of the relationship between the insect and its host: location,

acceptance, and host usage (MARKOW, 2019). Several morphological and sensory structures are involved in each component, with underlying genetic variability in each.

The first step is to locate the host and distinguish it from many others. For doing so, the insect recognizes a specific odor chemical compound to guide itself, identified by various chemoreceptors distributed across its body (including antennae, wings, and the Johnston's organ). The visual system will also integrate the system to assist in guiding to the host. Olfactory sensilla contain odorant-binding proteins (OBPs) (GALINDO; SMITH, 2001) that bind to different odors and connect them to odorant receptors (ORs), which will send the signal through neurons to the brain (LARTER; SUN; CARLSON, 2016). In *Drosophila sp.*, OBPs correspond to a family of ~50 genes expressed in the sensilla. These are small globular proteins capable of binding odor molecules in the pores of chemosensory sensilla, transporting them to their ligands near the ORs (MARKOW, 2019; SÁNCHEZ-GRACIA; ROZAS, 2008; VIEIRA; SÁNCHEZ-GRACIA; ROZAS, 2007), which are genes expressed in antennae and olfactory sensilla (GOMEZ-DIAZ et al., 2018; HALLEM; HO; CARLSON, 2004). ORs are a family containing more than 60 genes, with the function of receiving external chemical stimuli and transform into electrical pulses (GUO; KIM, 2007; MARKOW, 2019).

Once the insect has recognized its host, the second stage begins, involving the evaluation of the substrate. The host will be assessed not only for its nutritional quality, but also for the presence of competitors, predators, pathogens, and parasites. The substrate's quality is related to two types of receptors: gustatory receptors (GRs), which are primarily expressed in the organs located on the head (proboscis, hypo- and epipharynx, and antennae), as well as in the legs and wings (MONTELL, 2009); and ionotropic receptors (IRs), a gene family expressed in coeloconic sensilla, antennae, and palpal and labial sensilla (GOMEZ-DIAZ et al., 2018; RYTZ; CROSET; BENTON, 2013). GRs compose a family with 60 genes that mediate substrate tasting (MONTELL, 2009) and assess its nutritional quality (DUS et al., 2011; FUJITA; TANIMURA, 2011; STAFFORD et al., 2012). IRs are a group of 55 genes involved in olfaction, taste, oviposition roles in the mouthparts and legs, and are implicated in both oviposition and feeding (CHEN; AMREIN, 2017).

In the third and final stage, the insect must be able to utilize the substrate, in terms of reproduction and development, ensuring the survival of the larvae to become viable adults. Successful larval development depends on factors such as nutritional composition and overcoming the presence of toxic substances, when present. Since some plants, especially cacti, produce a variety of toxins to prevent herbivory, the larvae has a detoxification pathway as a viable strategy to overcome toxicity. Several gene families serve this purpose, including those

from the cytochrome P450 (CYPs), Glutathione S-Transferases (GSTs), UDP-glycosyltransferases (UGTs), Esterases (MARKOW, 2019), and ABC transporters (WU et al., 2019) families. The cytochrome P450 family has approximately 100 genes, highly divergent among *Drosophila* species. These genes are expressed in antennae, maxillary palps, labellum, legs, wings, and larvae (CHUNG et al., 2009; WILLINGHAM; KEIL, 2004), and they encode enzymes that catalyze various molecular reactions, mainly hydroxylations, on endogenous and exogenous proteins, many of which are associated with insecticide detoxification (GOOD et al., 2014). The GST gene family consists of around 30 genes, expressed in antennae, maxillary palps, and larvae (RIVERON; BOTO; ALCORTA, 2013; YOUNUS et al., 2014), with many of these also involved in insecticide resistance. They function by attaching the thiol group of glutathione to an electrophilic center in certain compounds, thereby eliminating toxins from cells by directing them to specific transporters or making them more water-soluble (LOW et al., 2007).

The UGT (UDP-glycosyltransferase) family has its genes expressed in the antennae (FRAICHARD et al., 2020) and is a family with 30 genes that catalyze the transfer of a glycosyl group from a nucleotide sugar, such as UDP-glucuronic acid, UDP-galactose, UDP-glucose, or UDP-xylose, to a variety of small hydrophobic molecules, resulting in more hydrophilic compounds that are efficiently excreted (LUQUE; O'REILLY, 2002). The esterase family consists of approximately 30 genes (OAKESHOTT et al., 1993), expressed in the adult head, larvae, pupae, ovaries, and testes (CAMPBELL et al., 2003). Finally, the ABC (ATP-binding cassette) gene family is a large class of transmembrane proteins, expressed in the brain and eyes of *Drosophila*, playing an important role in transporting a range of substrates, including xenobiotics (WU et al., 2019). Increased expression of ABC transporters has been shown to be associated with insecticide resistance in some insects, such as pyrethroid-resistant *Aedes aegypti* strains (BARIAMI et al., 2012), resistant strains of *Cimex lectularius* (MAMIDALA et al., 2012), as well as in *D. melanogaster* (SUN; BUCHON; SCOTT, 2017).

The genes associated with host location, acceptance, and host usage encode proteins related with feeding, reproduction, and development, hence their evolution is a key process during HS. (MARKOW, 2019). Rane et al. conducted a study with 14 *Drosophila* species and proposed 82 cases of positive selection occurring in the gene families of esterases, GSTs, CYPs, and UDPs, through analyses of selection (based on the ratio between nonsynonymous substitutions (Ka) and synonymous substitutions (Ks) ($\omega = Ka/Ks$)) that have accumulated over time in the divergence of coding sequences.

In addition to the selective pressure acting on the gene families involved in HS, changes in gene expression must also play an important role in the adaptation of *Drosophila sp.* to new environments. Although detoxification occurs predominantly in digestive organs, a study on *D. melanogaster* (CATALÁN; HUTTER; PARSCH, 2012) detected genes differentially expressed in the brains from African and non-African populations. These genes were related to stress response, olfaction, and detoxification. It has also been shown that detoxification genes (CYP and GSTs) and olfaction genes (OBPs) are differentially expressed in larval brains between *D. m. wrigleyi* and *D. m. sonorensis* (BENOWITZ et al., 2020), indicating that gene expression in the nervous system may play a role in detoxification and the olfactory sensory system, thus contributing to HS in *D. mojavensis*. Furthermore, it has been proposed that during larval development in the four subspecies of *D. mojavensis*, *D. navojoa*, and *D. arizonae*, there are behavioral differences related to locomotion as well as pupation site selection (COLEMAN et al., 2018). Considering that the larval stage shows the most conserved gene expression compared to other developmental stages of *Drosophila* (ARTIERI; SINGH, 2010), and that genes related to HS are expressed in this tissue, as well as in organs present in the adult fly's head, the transcriptomes of these two tissues are an excellent subject for evolutionary studies regarding HS in *Drosophila*.

OBJECTIVES

This thesis had two main objectives, which are divided into two chapters. The first objective was to develop a pipeline to identify gene-TE chimeric transcripts with paired-end Illumina reads. The second objective was to analyze the genome and transcriptome of seven cactophilic *Drosophila* species to investigate whether TEs could contribute to the evolution of genes associated with host shift, applying the method developed in the first chapter.

9764–9784 *Nucleic Acids Research, 2023, Vol. 51, No. 18* *Published online 24 August 2023*
https://doi.org/10.1093/nar/gkad671

ChimeraTE: a pipeline to detect chimeric transcripts derived from genes and transposable elements

Daniel S. Oliveira[1,2], Marie Fablet[2,3], Anaïs Larue[2,4], Agnès Vallier[4], Claudia M.A. Carareto[1], Rita Rebollo[4,*] and Cristina Vieira[2,*]

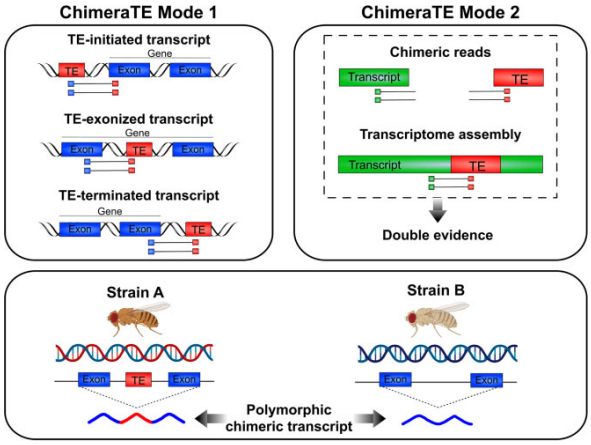
[1]São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, SP, Brazil, [2]Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, UMR5558, Villeurbanne, Rhone-Alpes, 69100, France, [3]Institut Universitaire de France (IUF), Paris, Île-de-FranceF-75231, France and [4]Univ Lyon, INRAE, INSA-Lyon, BF2I, UMR 203, 69621 Villeurbanne, France

Received November 03, 2022; Revised July 25, 2023; Editorial Decision July 27, 2023; Accepted August 09, 2023

ABSTRACT

Transposable elements (TEs) produce structural variants and are considered an important source of genetic diversity. Notably, TE-gene fusion transcripts, i.e. chimeric transcripts, have been associated with adaptation in several species. However, the identification of these chimeras remains hindered due to the lack of detection tools at a transcriptome-wide scale, and to the reliance on a reference genome, even though different individuals/cells/strains have different TE insertions. Therefore, we developed ChimeraTE, a pipeline that uses paired-end RNA-seq reads to identify chimeric transcripts through two different modes. Mode 1 is the reference-guided approach that employs canonical genome alignment, and Mode 2 identifies chimeras derived from fixed or insertionally polymorphic TEs without any reference genome. We have validated both modes using RNA-seq data from four *Drosophila melanogaster* wild-type strains. We found ∼1.12% of all genes generating chimeric transcripts, most of them from TE-exonized sequences. Approximately ∼23% of all detected chimeras were absent from the reference genome, indicating that TEs belonging to chimeric transcripts may be recent, polymorphic insertions. ChimeraTE is the first pipeline able to automatically uncover chimeric transcripts without a reference genome, consisting of two running Modes that can be used as a tool to investigate the contribution of TEs to transcriptome plasticity.

GRAPHICAL ABSTRACT



INTRODUCTION

Transposable elements (TEs) are mobile DNA sequences that comprise a large fraction of eukaryotic genomes, from 15% in *Drosophila melanogaster* (1), 45% in humans (2), to 85% in maize (3). Many TE copies have lost their ability to transpose as a result of accumulated mutations and recombination throughout evolution (4). Despite their lack of mobility, such ancient TE insertions may still harbor functional protein domains, alternative splice sites, and *cis*-acting regulatory sequences, as transcription factor binding sites (TFBSs) and polyadenylation (PolyA) sites. Therefore, TEs are a major source of genetic diversity, not only due to their mobilization, but also because they donate protein domains to genes (5–8), and regulatory sequences that modify the expression of nearby genes (9–13). The participation of TE-derived sequences in the host biology is a process called domestication or exaptation (14). The ancestral role of the TE sequence can be domesticated into an essential host

*To whom correspondence should be addressed. Tel: +33 4 72 44 82 16; Email: cristina.vieira@univ-lyon1.fr
Correspondence may also be addressed to Rita Rebollo. Tel: +33 4 72 43 62 46; Email: rita.rebollo@inrae.fr

function, but it can also be modified, and adapted, into a new exapted function that may also be essential for the host species (14).

Chimeric transcripts are RNAs stemming from two sequences of different origins (15). Hereafter we define chimeric transcripts as mature transcripts that have both gene and TE-derived sequences. These transcripts can be divided into three types: (1) TE-initiated transcripts: chimeric transcripts with a TE transcription start site (TSS) (16,17); (2) TE-exonized transcripts: TE sequences are incorporated into the transcript either partially or as full-length exons (18–20); and (3) TE-terminated transcripts: chimeric transcripts with a TE transcription termination site (21,22). TE-initiated and TE-terminated transcripts might modulate gene expression levels either by the presence of TFBSs, PolyA sites, or chromatin changes; while TE-exonized transcripts may alter the protein sequence of coding genes and have a direct effect on the protein function. Regardless of the TE position, such events of TE exaptation and domestication have been associated with many biological roles and are widespread among eukaryotic species (14). In *D. melanogaster*, the *CHKov1* gene produces a chimeric transcript with a truncated mRNA resulting in resistance to insecticide and viral infection (23). In humans, the *SETMAR* gene produces a chimeric transcript containing a *Hsmar1* copy, involved in non-homologous end-joining DNA repair (24). In cancer, TEs become active due to a global hypomethylation state (25) and such activation may generate new chimeric transcripts with detrimental outcomes (26), a process called onco-exaptation (9). For example, in large B-cell lymphoma, the *FABP7* gene has an endogenous retrovirus LTR co-opted as a promoter, generating a novel protein involved in abnormal cell proliferation (27). Recently, such phenomenon has been observed in a larger scale. A pan-cancer study revealed 1,068 tumor-specific TE-initiated transcripts from genes coding antigens, showing the high prevalence of TE exaptation in cancer (28). Therefore, chimeric transcripts have a large impact on host biology, but their study remains hindered by the ubiquitous repetitive nature of TE copies, and lack of methods to identify chimeric transcripts.

Previous studies with Cap Analysis Gene Expression (CAGE) revealed a significant percentage of genes producing TE-initiated transcripts, ranging from 3–14% in humans and mice, depending on the tissue (29). More specifically, in human pluripotent stem cells, chimeric transcripts comprise 26% of coding and 65% of noncoding transcripts (30). In *D. melanogaster*, a study using expressed sequence tags (ESTs) showed that the proportion of genes with chimeric transcripts is only ∼1% (31). Another study used an approach based on cap-enriched mRNA sequencing but with a preliminary transcript elongation step allowing not only to detect TSSs but also to estimate mRNA expression rates (RAMPAGE) (32). This study detected TE-initiated chimeric transcripts in 36 stages of *D. melanogaster* life cycle, and observed that up to 1.6% of all transcripts were chimeras, representing up to ∼1% of all genes (33). More recently, a tissue-specific study has shown that 264 genes produce chimeric transcripts in the midbrain of *D. melanogaster*, corresponding to ∼1.5% of all genes (34).

Several bioinformatic methods have been developed to take advantage of RNA sequencing (RNA-seq) to identify chimeric transcripts, such as CLIFinder (35) and LIONS (36). The former is designed to identify chimeric transcripts derived from long interspersed nuclear elements (LINEs) in the human genome, whereas the latter identifies only TE-initiated transcripts. Both methods need a reference genome and they only detect chimeric transcripts derived from TE insertions present in the reference genome. Therefore, it is impossible to identify chimeric transcripts derived from polymorphic TE insertions that may exist in other populations, strains, or individuals. Finally, the latest addition to chimeric transcript detection, TEchim (34), can detect chimeras with TEs that are polymorphic and absent from the reference genome of *D. melanogaster*, but it is not a pipeline designed to run automatically with any other genome.

In this study, we have developed ChimeraTE, a pipeline that uses paired-end RNA-seq reads to identify chimeric transcripts. The pipeline has two Modes: Mode 1 can predict chimeric transcripts through genome alignment, whereas Mode 2 performs chimeric transcript searches without a reference genome, being able to identify chimeras derived from fixed or polymorphic TE insertions. To benchmark the pipeline, we have used RNA-seq from ovaries of four *D. melanogaster* wild-type strains, for which we have assembled and annotated genomes. We found that ∼1.12% of genes have chimeric transcripts in the ovarian transcriptome, of which ∼88.97% are TE-exonized transcripts. Our results also reveal that the retrotransposon *roo* is the most frequent exonized TE family. In addition, with Mode 2, we found 11 polymorphic chimeric transcripts deriving from TE insertions that are absent from the *D. melanogaster* reference genome. Therefore, this work provides a new strategy to identify chimeric transcripts with or without the reference genome, in a transcriptome-wide manner.

MATERIALS AND METHODS

ChimeraTE: the pipeline

ChimeraTE was developed to detect chimeric transcripts with paired-end RNA-seq reads. It is developed in *python3 v.3.6.15* language, and it is able to fully automate the process in only one command line. The pipeline has two detection Modes: (1) genome-guided, the reference genome is provided and chimeric transcripts are detected aligning reads against it; and (2) genome-blind, the reference genome is not provided and chimeric transcripts are predicted for fixed or polymorphic TEs. These Modes have distinct approaches that may be used for different purposes. In Mode 1, chimeric transcripts will be detected considering the genomic location of TE insertions and exons. Chimeras from this Mode can be classified as TE-initiated transcripts (TE located upstream of the gene), TE-exonized transcripts (TE within introns or embedded in gene exons), and TE-terminated transcripts (TE located downstream of the gene). In addition, results from Mode 1 can be visualized in a genome browser, which allows a manual curation of chimeric transcripts in the reference genome. Mode 1 does not detect chimeric transcripts derived from TE insertions absent from the provided reference genome. Mode 2 predicts chimeric

transcripts considering the alignment of reads against transcripts and TE insertions, in addition to a transcriptome assembly (user optional). Hence, Mode 2 detects chimeric transcripts from *de novo* TE insertions and an assembled genome is not necessary. In this Mode, two alignments are performed: (i) transcript alignment and (ii) TE alignment. Then, based on both alignments, the pipeline identifies chimeric reads that support chimeric transcripts, regardless of the TE genomic location. In Mode 2, since there is no alignment against an annotated genome, it is not possible to classify chimeric transcripts considering the TE position as in Mode 1.

Both ChimeraTE Modes use chimeric reads, which are defined by paired-end reads that contain both TE and exon sequences, as evidence of chimeric transcripts. This method has been widely demonstrated by other authors as a potential source of artifactual reads, mainly due to the occurrence of mixed clusters (*index hopping*) on the Illumina's flow cell that may be too close to each other, as well as jumping PCR (*in vitro* crossover artifact), generating read pairs that are connecting two cDNA portions that are not joined in the sample (37–40). Indeed, it has been shown that up to 1.56% of all reads produced by Illumina multiplexed approaches might generate chimeric reads (41), including cases that may support chimeric transcripts derived from different genes. These artifactual reads originate more likely from highly expressed genes since there are more molecules on the Illumina's cell. Conversely, because TE-derived sequences might comprise a low proportion of the transcriptome, artifactual reads from TEs should be produced at a low frequency. Furthermore, it is unlikely that artifactual reads from the same gene and TE family among RNA-seq replicates would be produced. Nonetheless, to avoid including false chimeric reads, both Modes of ChimeraTE only call chimeric transcripts that are detected in at least two RNA-seq replicates (user optional for more).

ChimeraTE mode 1: genome-guided approach

In ChimeraTE Mode 1, paired-end RNA-seq reads, a whole genome assembly, and its respective gene/TE annotation are used to predict chimeric transcripts (Figure 1A). We highlight the need for robust gene and TE annotations to submit an analysis to Mode 1, since each TE insertion and exon will be used to identify chimeric reads. Firstly, RNA-seq alignment to the genome is performed with *STAR v.2.7.6a* with default parameters (Figure 1B), and transcript expression is assessed with *Cufflinks v.2.2.1* (42) for each RNA-seq replicate. We consider FPKM $\geq$1 as an expressed gene by default, but it can be changed by the user with the *–fpkm* parameter. Only concordant reads (both reads from the pair are aligned) with unique alignment (*samtools -q 255*) are selected and converted to BED format with *samtools v.1.10* (43) and *bedtools v.2.30.0* (44), implemented in *python3* with *pybedtools v.0.9.0* (45). Split reads are also considered as chimeric reads when one mate of the pair stems from the TE to the exon, or vice-versa (Supplementary Figure 1). The alignment position of these reads are considered independently with *bedtools bamtobed -split*. Finally, reads aligned into the forward and reverse strands are separated with *samtools* (43).

The IDs from reads that have aligned against exons are identified with *bedtools intersect* (Figure 1B). Then, reads with at least 50% of their length (*–overlap 0.50*) aligned against reference TE copies have their IDs selected (Figure 1B), and TE copies without aligned reads are removed from the downstream analysis. These two lists of IDs are intersected between each other and reads that have aligned to both exons and TEs are identified, generating a raw list of chimeric reads (Figure 1C). TE-initiated transcripts are the first to be searched by ChimeraTE Mode 1. Expressed genes that have TE copies within their 3 kb upstream region (default but adjustable with *–window* parameter) are identified with *bedtools* (44). For each gene, the pipeline will check whether the TE located upstream has chimeric reads shared with the gene exons (Figure 1D). If there are no chimeric reads, both gene and TE are discarded. The same method is applied to identify TE-terminated transcripts, but with genes harbouring downstream TEs up to 3 kb (user adjustable value) (Figure 11D). Finally, TE-exonized transcripts are also identified thanks to chimeric reads aligned to exons and TEs located within genes. Based on the TE position towards exons, they are classified as embedded, intronic and overlapped. Embedded TEs are located entirely within an exon. In these cases, reads aligned only to the TE sequence are not considered chimeric reads (Supplementary Figure 1). Indeed, the exon can be artificially divided into two portions, and chimeric reads must have a read (or a split read) aligned to the TE, whereas its mate (or the extension from the split read) aligned to at least one exon portion (Supplementary Figure 1). This method avoids autonomous TE expression as evidence of chimeric transcript expression. The same is performed to identify TE-exonized transcripts derived from TEs that are partially overlapping exons but are also located in introns (Supplementary Figure 1). Finally, TE-exonized transcripts with TEs located in introns are identified with *bedtools intersect*, and chimeric reads are detected to support chimeric transcripts.

These steps are repeated for all RNA-seq replicates provided in the input. Then, the raw results from replicates are compared, and all chimeric transcripts that were found in $\geq$2 replicates and have been identified with at least $\geq$2 chimeric reads on average between replicates are considered true chimeric transcripts (Figure 1C). These thresholds may be changed by the user with *–cutoff* and *–replicate* parameters. Mode 1 output is a table with a list of predicted chimeric transcripts categorized by TE position, with gene ID, TE family, chimeric read coverage, TE location, gene location, and gene expression (Figure 1D).

ChimeraTE mode 2: a genome-blind approach to uncover chimeric transcripts

ChimeraTE Mode 2 is the genome-blind approach of the pipeline. The input data are stranded RNA-seq reads, gene transcripts, and reference TE insertions. TE consensus can be used, but it will decrease the alignment rate, and therefore it should be used only when a fasta file with TE insertions is not available (Figure 2A). The data will be used to perform two alignments with *bowtie2 v.2.4.5* (46), one against all transcripts and another against all TE sequences, both with parameters: *-D 20 -R 3 -N 1 -L 20 -i S,1,0.50* (Figure 2A). To

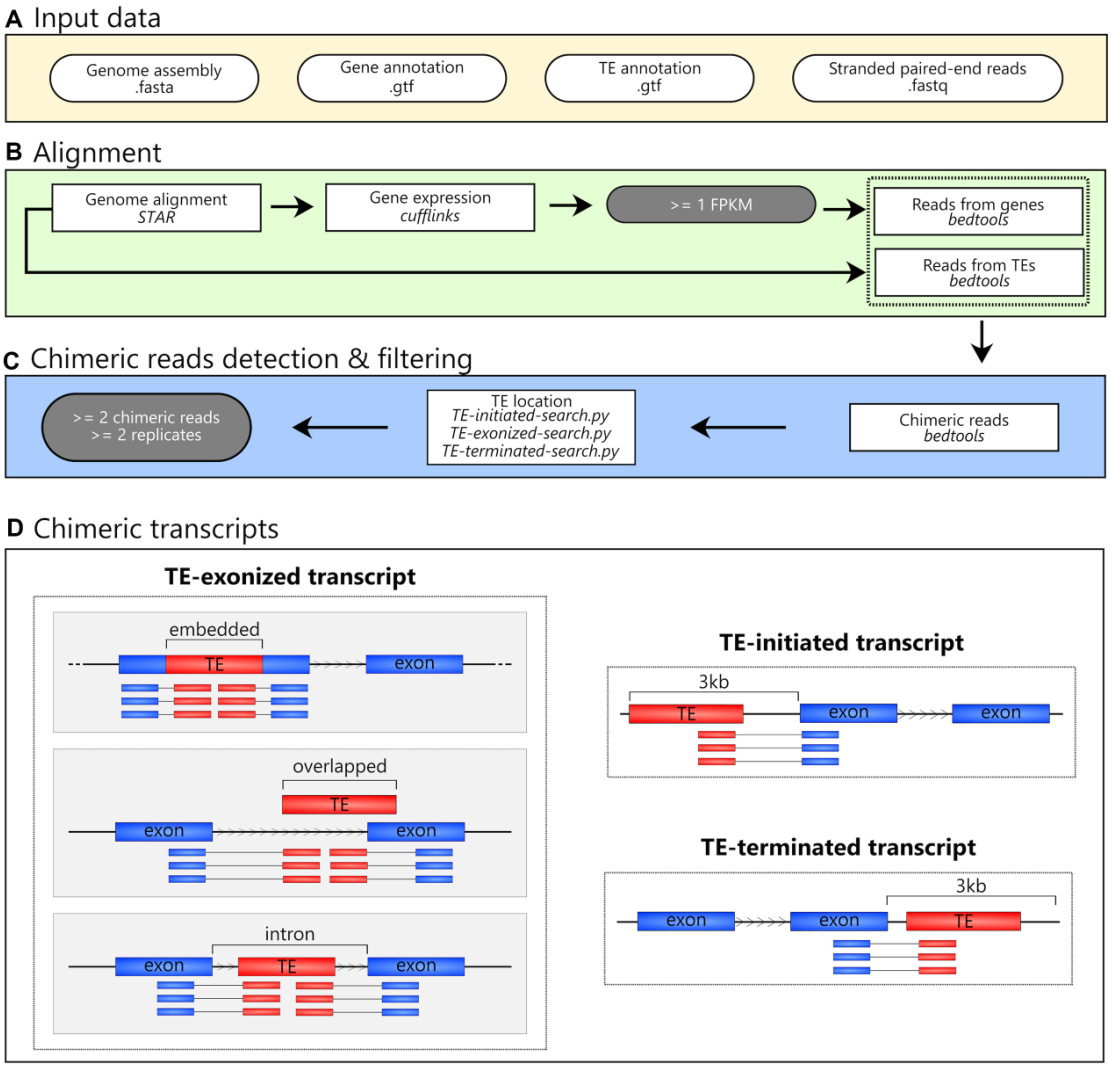


Figure 1. ChimeraTE Mode 1 (genome-guided) workflow. Round white boxes: input data; square boxes: pipeline step; round gray boxes: thresholds that can be modified. (**A**) Input data: fasta file with the genome assembly, gtf files with gene and TE annotations, as well as stranded paired-end reads from RNA-seq (fastq). (**B**) Alignment: RNA-seq alignment to the genome is used to calculate gene expression levels. Genes with FPKM < 1 are removed from downstream analyses. A subsequent list of reads that have aligned against genes or TE insertions is created. (**C**) Chimeric read detection & filtering: both read lists are then compared and read pairs that have common reads between the two lists are named chimeric reads, *i.e.* paired-end reads mapping to a gene and a TE copy. The average of these reads between replicates is used as chimeric read coverage for each putative chimeric transcript. All putative chimeras are then processed with three ChimeraTE scripts to categorize them into TE-initiated, TE-exonized, and TE-terminated transcripts. These steps are performed for all RNA-seq replicates. Finally, all chimeric transcripts present in at least 2 replicates and with at least 2 chimeric reads on average between replicates are maintained. (**D**) Chimeric transcripts: Three predictions obtained from Mode 1. Blue boxes: exons; red boxes: TEs; arrowhead in between TE and exon boxes: transcription sense; blue and red boxes linked by a line: chimeric reads. The ChimeraTE mode 1 output is divided into three predictions: (i) TE-initiated transcript: the TE insertion is located upstream of the gene region; (ii) TE-exonized transcript: the TE insertion is present within exons (embedded), overlapping exons (overlapped), or introns (intronic); (iii) TE-terminated transcript: the TE insertion is located downstream of the gene region.

avoid very low-expressed transcripts predicted as chimeras and to ensure reasonable processing time, the SAM alignment is converted to BAM with *samtools v.1.10* (43), and FPKMs are computed for the reference transcripts provided in the input using *eXpress v.1.5.1* (47). Then, all genes with average FPKM <1 are removed from the downstream analysis (Figure 2B). To identify chimeric reads between TEs and gene transcripts, both alignments are converted to BED with *bedtools v.2.30.0* (44). Among all aligned paired-end reads, the pipeline considers as chimeric transcripts the ones that have at least one read aligned to the TE sequence (singleton mapping) and its mate to the gene transcript, or when both reads have aligned (concordant mapping) to the TE and gene transcript. In order to identify these reads, the TE alignment output is used to create a list with all read 1 IDs that have aligned against TEs, and another list with all

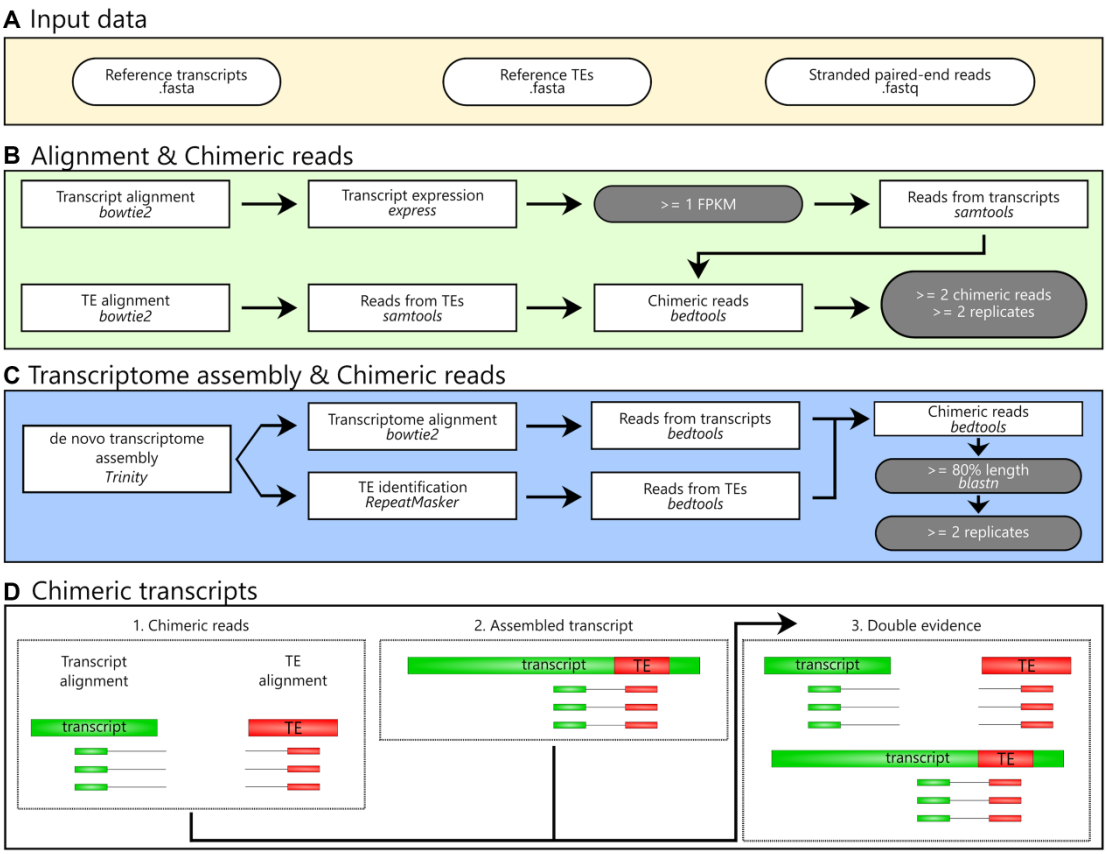


Figure 2. ChimeraTE Mode 2 (genome-blind)workflow. Round white boxes: input data; square boxes: pipeline step; round gray boxes: thresholds that can be modified. (**A**) Input data: two fasta files containing reference transcripts and TE insertions, as well as stranded paired-end reads from RNA-seq (fastq). (**B**) Alignment and chimeric reads: The alignment against transcripts is performed and their expression is calculated. Transcripts with FPKM < 1 are removed from the downstream analysis. Next, a list of reads aligned against transcripts is created. Through the alignment of reads against TE insertions, a second list with reads stemming from TEs is also created. Then, mapped paired-end reads and singletons are identified, generating the list of chimeric reads, for all replicates. All chimeric transcripts that have an average of chimeric reads $>= 2$ and are present in $>= 2$ replicates are maintained as true chimeras. (**C**) Transcriptome assembly and chimeric reads: The de novo transcriptome assembly is a non-default option of ChimeraTE Mode 2. It performs a transcriptome assembly and aligns reads against the assembled transcripts. Then, TE insertions in the assembled transcripts are identified with RepeatMasker and the TE reads are recovered. Using the two lists of reads (transcripts and TEs), the chimeric read list is generated and the putative assembled chimeric transcripts are predicted. Next, a blastn is performed between these transcripts and the reference transcripts provided in the input. All transcripts with length $>= 80\%$ are selected. The process is repeated for all RNA-seq replicates and chimeric transcripts assembled $>= 2$ replicates are maintained as true chimeras. (**D**) Chimeric transcripts: if the assembly is activated, ChimeraTE mode 2 provides three outputs: (1) Chimeric reads: predicted only based on the method depicted in B; (2) Assembled transcripts: predicted only based on the transcriptome assembly method depicted in C; and (3) Double evidence: predicted by both methods -B and C-.

read 2 IDs, regardless if their mates have also aligned (concordant mappings or singleton mappings). The same lists are created by using the transcript BED file: (1) read 1 IDs of transcript mapping reads and (2) mate 2 IDs of all mate 2 reads, regardless of mate mapping. All mate 2 IDs that have a TE-aligned read 1 are searched in the list of transcript-aligned mate 2. The same is performed in the opposite direction (TE-aligned read 2, transcript-aligned mate 1). These read pairs will therefore be comprised of two mates from the same pair that were singletons in the alignments, *i.e.* pairs comprised of one read that has aligned against a TE, and its mate against a gene transcript. The cases in which the chimeric transcript does not have the TE insertion in the reference transcript will be supported only by these singleton chimeric reads (Figure 2D). For cases in which the TE insertion is present inside the reference transcript, chimeric reads supporting it may either be singleton or concordant reads. Therefore, chimeric reads can be concordant reads in both alignments (TEs and genes), or they may be concordant only in the gene transcript alignment and singleton in the TE alignment. Due to the repetitive nature of TEs, short-read alignment methods provide very few unique aligned reads against loci-specific TE copies as most reads align ambiguously between similar TE insertions. Therefore, when a chimeric transcript has been identified involving more than one TE family, the TE family with the highest coverage of chimeric reads is maintained. Subsequently, ChimeraTE uses two chimeric reads as a threshold for calling a chimeric transcript, that can be modified by the user with the *–cutoff* parameter. Such value does not represent

transcript expression nor TE expression, but it represents the coverage supporting the junction between a gene transcript (CDSs/UTRs) and a TE sequence. Finally, the output tables show the list of genes and the respective TE families detected as chimeras, reference transcript ID and the total coverage of chimeric reads supporting it.

The support of chimeric transcripts performed by ChimeraTE Mode 2 is from chimeric reads aligned by an *end-to-end* approach. Such alignment may reduce alignment sensitivity, since exon/TE junctions may be covered by split reads. To mitigate the loss of detection power due to the alignment method with Mode 2, alongside with chimeric read detection using alignments against transcripts and TEs, there is an option to run Mode 2 with a transcriptome assembly approach, which can be activated with *–assembly* parameter (Figure 2C). This approach will use RNA-seq reads to perform a *de novo* transcriptome assembly with *Trinity v2.9.1* (48). *RepeatMasker v4.1.2* (49) is used to identify assembled transcripts that may have TE-derived sequences providing *–ref_TEs*, a custom TE library, or predefined TE consensus sequences from *Dfam v3.3* (50), according to the taxonomy level, i.e.: flies, mouse, humans. Then, RNA-seq reads are aligned with *bowtie2 v.2.4.5* (46) against the assembled transcripts. Subsequently, the alignment is used to identify whether transcripts containing TE-derived sequences have chimeric reads, including split reads. All assembled transcripts with chimeric transcripts are selected as candidate chimeric transcripts. Next, these candidates are submitted to a homology analysis with reference transcripts using *blastn v2.11.0* (51). Finally, all assembled transcripts with masked TEs that have at least 80% of similarity with reference transcripts across 80% of their length (can be modified with *–min_length* parameter) are considered chimeric transcripts. All these steps are repeated for all RNA-seq replicates provided in the input. Finally, the list of chimeric transcripts obtained from all replicates with the transcriptome assembly approach is compared, and all chimeras that have been identified with at least $\geq$2 chimeric reads and were found in $\geq$2 replicates, are considered as true chimeric transcripts. By activating the *–assembly* option in Mode 2, the output table will provide chimeric transcripts that have been predicted based on different evidences (Figure 2D): (1) only based on chimeric reads; (2) only based on transcript assembly; (3) based on chimeric reads and transcript assembly.

D. melanogaster wild-type strains: genome assemblies and RNA-seq

To assess ChimeraTE's performance, as well as the efficiency in the identification of chimeric transcripts derived from polymorphic TE insertions, we have mined previously available RNA-seq data from ovaries of four *D. melanogaster* wild-type strains (52), two from Gotheron, France (44_56′0″N 04_53′30″E), named dmgoth101 and dmgoth63; and two from São José do Rio Preto, Brazil (20_41′04.3″S 49_21′26.1″W), named dmsj23 and dmsj7. Such RNA-seq data was produced with mRNA libraries, which is strongly advised to use with ChimeraTE, since total RNA libraries contain pre-mRNA sequences and might introduce many false positives of TE-exonized transcripts. Sequencing was performed on an Illumina Hiseq (125 bp paired-end reads), with two biological replicates. All RNA-seq libraries were trimmed by quality, and adapters were removed with Trimmomatic (53). Each strain had also its genome previously sequenced by Nanopore long reads and assembled (54). The high-quality assemblies allowed us to manually check whether chimeric transcripts predicted by both ChimeraTE Modes have the predicted TE insertions inside/near genes, as well as manually curate the presence of chimeric reads.

Running ChimeraTE with *D. melanogaster* data

To run ChimeraTE Mode 1 on the available RNA-seq data, we performed gene annotation in the four *D. melanogaster* genomes with *Liftoff v.1.6.3* (55) using *-s 0.5 -s 0.75 -exclude_partial* parameters and the r6.49 *D. melanogaster* genome (*dm6* strain) available in Flybase as reference. TE annotation was performed with *RepeatMasker v4.1.2* (49), with parameters: *-nolow*; *-norna*; *-s;* and *-lib* with the TE sequence library for *D. melanogaster* provided by Bergman's lab *v.10.2* (https://github.com/bergmanlab/drosophila-transposons/blob/master/current/D_mel_transposon_sequence_set.fa). Small TE insertions with length < 80 bp were discarded due to the lack of robustness to define them as TEs, following the 80–80–80 rule (56). In addition, due to the presence of simple sequence repeats (SSRs) in many TEs from *D. melanogaster*, such as *roo*, *412*, *tirant* (57), several insertions from *RepeatMasker* might be indistinguishable from SSRs if the repeat region comprises a large proportion of the TE sequence (58,59). Therefore, to increase the robustness of the TE annotation, we identified the presence of SSRs with *TRF v.4.09* (60) across all insertions annotated on each strain, with parameters *2*, *5*, *6*, *75*, *20*, *50*, *500*, *-m*, *-h* and removed TEs that presented SSRs over >50% of their sequences. We used ChimeraTE Mode 1 with default parameters, setting *–strand rf-stranded*, since our paired-end reads have the orientation as reverse and forward for the first and second reads, respectively. In case the stranded paired-end reads were produced as forward and reserve (first and second reads, respectively), the parameter would be *–strand fwd-stranded*. For ChimeraTE Mode 2, we demonstrate its potential in detecting chimeric transcripts derived from TE insertions that are not present in a reference genome, even though the transcript sequences and TE copies provided as input stemmed from the *dm6* reference genome. Therefore, we have used ChimeraTE mode 2 with RNA-seq from the four wild-type strains listed previously, with reference transcripts from *D. melanogaster* (*dm6* strain) available in Flybase r6.49 (http://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta/dmel-all-transcript-r6.49.fasta.gz) (61). The TE annotation for the *dm6* genome was assessed with the same protocol used on the four wild-type strains, with *RepeatMasker* (49). We have used default parameters, except by *–assembly*.

Additional species data

In order to demonstrate that ChimeraTE can be used with other species than *Drosophila* spp., we applied it to

Arabidopsis thaliana, *Homo sapiens* and the fish *Poecilia reticulata* (guppy) RNA-seq datasets. The latter was selected as an example of non-model species. The genome and gene annotation of *A. thaliana* was obtained from NCBI (GCF_000001735.14), version TAIR10.1. The TE annotation was assessed with *RepeatMasker v.4.1.2* (49), with *parameters -species arabidopsis -cutoff 225 -nolow -norna -a -s*. We downloaded two replicates of paired-end RNA-seq from *A. thaliana* leaf, available on NCBI (SRR21230172, SRR21230173). Regarding human data, the genome, gene and TE annotations were retrieved from NCBI (62), version GRCh38.p14. Two paired-end RNA-seq replicates from the human leukemic cell-line K562 were downloaded from NCBI (SRR521457, SRR521461). Finally, *P. reticulata*'s genome and gene annotation were also retrieved from NCBI (GCF_000633615.1), and TE annotation was assessed with *RepeatMasker v.4.1.2*, parameters*: -species Actinopteri -cutoff 225 -nolow -norna -a -s*. Two RNA-seq replicates from ovary follicular tissue were downloaded from NCBI (SRR17332506, SRR17332508). Both ChimeraTE Mode 1 and Mode 2 were used with default parameters, except by *–threads 32*, and *–ram 64*.

Benchmarking polymorphic chimeric transcripts with nanopore genomes

Once chimeric transcripts were identified, we used the high-quality Nanopore assemblies for dmgoth101, dmgoth63, dmsj23 and dmsj7 previously published (54) to confirm whether genes predicted by Mode 1 as chimeric transcripts have indeed the respective TE insertion located near or within them. To do so, we have used an *ad-hoc* bash script to create three bed files from genes: 3 kb upstream; 3 kb downstream, and gene region. Then, we used *bedtools intersect* (44) to identify genes with TEs located in the three regions. For Mode 1 we have randomly sampled 100 chimeric transcripts of each wild-type strain to visualize the alignments performed by Mode 1 on the *IGV* genome browser (63). For Mode 2, all genes not found by *bedtools intersect* with the predicted TE insertion were visualized in *IGV*. In both manual curations, we considered false positives those cases in which we did not find the TE insertion, or we found the TE insertion, but without chimeric reads.

In order to assess the number of chimeric transcripts found by Mode 2 in wild-type strains derived from TE insertions absent in the *dm6* genome, we also used the *ad-hoc* bash script to create the bed files with 3 kb ± and the gene regions for *dm6*. Then, we used *bedtools intersect* (44) with TE annotation and the gene regions. By using this method, we generated a list of genes with TEs located 3 kb upstream, inside genes (introns/exons), and 3 kb downstream for the *dm6* genome. Then, the polymorphic chimeric transcripts were identified with the comparison of genes with TEs inside/nearby in *dm6* and the list of chimeric transcripts in the wild-type strains. In addition, all chimeric transcripts derived from TEs that were not found in *dm6* were manually curated with the IGV genome browser (63).

Validation of chimeric transcripts by RT-PCR

D. melanogaster strains were kept at 24°C in standard laboratory conditions on cornmeal–sugar–yeast–agar medium. For sampling, 3–7 day old *D. melanogaster* females were immersed in Phosphate-buffered saline solution (PBS) and dissected under a stereomicroscope. Three biological samples were collected in buffer TA (35 mM Tris/HCl, 25 mM KCl, 10 mM $MgCl_2$, 250 mM sucrose, pH 7.5), each composed of 30 pairs of ovaries. All samples were collected on ice in 1.5 ml RNAse free tubes and stored at −80°C until use. Total RNA was extracted using QIAGEN AllPrep DNA/RNA Micro extraction kit (Qiagen 80284) following the manufacturer's guideline. cDNA synthesis was performed on 1 μg of total RNA or water (RT negative control) with BIORAD iScript cDNA systhesis (Biorad 1708891). Primers were designed considering the TE and the exon location with aligned chimeric reads (Supplementary Table 1) and PCRs were performed with Taq'Ozyme (Ozya001-1000).

Sequence and protein analysis of TE-exonized elements

The sequences of TE-exonized elements were extracted from wild-type genomes with *bedtools getfasta* (44), parameter *-s,* using BED files created by ChimeraTE Mode 1. Since the TE reading frame incorporated into the gene transcript is unknown, we recovered all open reading frames (ORFs) with *EMBOSS v.6.6.0 getorf* (64) in the three coding frames, from both strands. Next, the protein domains in these sequences were assessed with *pfam-scan v.1.6* (65), using *Pfam-A.hmm* database from InterPro (66), with default parameters. We applied a filtering step for protein domains with multiple overlapped matches in the same TE insertion, keeping in the downstream analysis only the longest match. We also removed protein domains that had no association with any TE function, following the description from InterPro. Finally, to analyze whether *roo* elements generating TE-exonized embedded transcripts provided a specific motif to the chimeric exons, we extracted their sequences with *bedtools getfasta*, and aligned with *roo* consensus sequence with *MUSCLE v.5.1* (67). Alignment plots were performed with *MIToS v.2.11.1* (68), using the package *Plots*.

Differential expression analysis

The read count for differential gene expression analysis for the four wild-type strains was obtained from Fablet *et al.* (52). The data were normalized with *DEseq2 v.3.10* (69), with a designed model (∼genotype), assuming that the only variable to identify differentially expressed genes between strains must be the genotype. We performed pairwise comparisons with the four wild-type strains, considering differentially expressed genes the ones with adjusted p-value < 0.05.

RESULTS AND DISCUSSION

ChimeraTE predicts chimeric transcripts derived from genes and TEs using two different strategies. Mode 1 is a genome-guided approach that will predict chimeric transcripts from paired-end RNA-seq through chimeric read pair detection. The main advantages of Mode 1 in comparison to Mode 2 are that the first one can detect split reads between exons and TEs, capture chimeric transcripts with low

coverage/expression and classify chimeric transcripts according to the TE position: TE-initiated, TE-exonized, TE-terminated transcripts. However, Mode 1 misses chimeric transcripts derived from TE insertions that are absent from a reference genome. Mode 2, which is a genome-blind approach, performs two alignments against reference transcripts and TE copies and then, similar to Mode 1, predicts chimeric reads between these two alignments. In addition, Mode 2 can optionally perform a *de novo* transcriptome assembly able to detect chimeric transcripts and chimeric reads through split read alignment, improving the sensitivity. Despite the similarities between both Modes, the results from Mode 1 and 2 had an overlap of 50.46% (Supplementary Data, note 1). We have shown that such differences are likely to be associated with the alignment used by the Modes, and read length (Supplementary Data, note 1, Fig S1.1) (70,71). We also observed that increasing the minimum number of chimeric reads (default = 2) causes a decrease of 48.45% for Mode 1 and 33.48% for Mode 2 in the total number of chimeras found in two RNA-seq replicates (Supplementary Data, note 2). Therefore, a substantial part of the results is detected with low coverage of chimeric reads. Finally, we demonstrated that the use of RNA-seq replicates is crucial to reduce the frequency of potential chimeras supported by artifactual chimeric reads (Supplementary Data, note 3) (37,72–77).

Setting up the datasets for ChimeraTE

Each ChimeraTE Mode requires different input datasets. To run Mode 1 (genome-guided), gene and TE annotations, along with a genome fasta file are necessary. We took advantage of available paired-end RNA-seq datasets from the ovaries of four wild-type strains of *D. melanogaster* (52), for which high-quality genome assemblies were also available (54). We performed gene annotation in the new assemblies using *D. melanogaster*'s genome (*dm6*) from Flybase r6.49 as a reference and obtained ∼17,278 genes per genome (Supplementary Table 2). Regarding TE annotation, we found ∼10.48% of TE content in the four wild-type strains, similar to our previous estimates for these strains (54). Filtering out TE insertions smaller than 80 bp, removed ∼5,549 TEs (20.59%) across the four wild-type strains. In addition, to remove potential mapping artifacts, we discarded ∼752 TE copies (2.79%) per strain that harbored SSRs longer than half the TE length. On average across strains, 85% of the TEs filtered out due to the presence of SSRs were *roo* elements (Supplementary Figure 2). These *roo* copies were around 112 bp (Supplementary Figure 3), which is the same length as the known SSR present in the roo consensus sequence (57). In total, 123 TE families in the four genomes were uncovered, comprising a mean of ∼21,402 TE insertions (*standard deviation*= 883). In all genomes the TE content in bp is higher for LTR, then LINE elements, followed by DNA and Helitron families (Supplementary Table 2) (78). In order to run Mode 2 (genome-blind), we used reference transcripts from Flybase r6.49 and performed TE annotation on the reference *dm6* genome. We have obtained 29,086 TE insertions, representing 16.37% of the genome content, following similar TE family proportions as seen for the four wild-type genomes (Supplementary Table 2). Nevertheless, the *dm6* genome TE annotation shows an extra ∼6% of TE content compared to the wild-type genomes, mainly due to LTRs and LINEs (Supplementary Table 2). The *D. melanogaster's* heterochromatin corresponds to ∼32.5% of its genome (79), and includes repeat sequences, as TE copies (80). The wild-type assemblies have a high coverage for euchromatin regions, since 96% of gene content was recovered, but a poor heterochromatin coverage, corresponding to the lack of TEs compared to the *dm6* genome. However, the search for chimeric transcripts with ChimeraTE Mode 1 is restricted to the gene region and its 3 kb boundaries, hence, the lack of heterochromatic TE insertions does not impact the chimeric transcript search.

ChimeraTE mode 1 reveals that 1.12% of genes produce chimeric transcripts in *D.melanogaster* wild-type strains

ChimeraTE mode 1 was run on the four wild-type strain genomes and their respective ovarian RNA-seq data (52). Across all strains, we found 327 genes producing 408 chimeric transcripts, accounting for 1.83% of the total genes in *D. melanogaster*, 4.04% of the expressed genes (FPKM > 1), and representing 1.12% of all genes when normalized by the four strains. Previous studies have found similar proportion of genes generating chimeras in *D. melanogaster.* Among them, Coronado-Zamora et al. is the only study that performed chimeric transcripts search in multiple strains and different tissues, including ovaries, for which the authors found 1.41% of genes with chimeric transcripts when normalized by all strains (81). To verify whether ChimeraTE identified the correct TE family for each chimera, we have compared the genomic coordinates of TEs and genes (3 kb upstream/downstream and inside genes) with *bedtools intersect* (44) and predicted genes with TE copies in these regions. All chimeric transcripts in the four wild-type strains had at least one TE insertion in the expected position from the predicted TE family (Supplementary Tables 3–6). In addition, we randomly selected 100 chimeric transcripts in each wild-type strain to visualize in IGV (63) and confirm the presence of chimeric reads as expected (Figure 3A). All the 400 manually inspected chimeras were correctly found in the genome browser. Among all genes generating chimeric transcripts, around 89% are TE-exonized (61.84% correspond to TE-exonized embedded, 17.18% to TE-exonized intronic, 9.95% to TE-exonized overlapped), 8.64% TE-terminated and 2.36% TE-initiated transcripts (Figure 3B). It is important to note that ChimeraTE does not classify an internal TE copy as either a TE-initiated or a TE-terminated transcript, even if the transcript does indeed start, or end, at the TE sequence. TE-initiated and TE-terminated repeats are necessarily outside of gene regions as *per* ChimeraTE categories (see Materials and Methods). Therefore, one should assume that the high prevalence of TE-exonized transcripts (∼89%) might be associated with potential cases of misannotated TE-initiated and TE-terminated transcripts. Indeed, a previous study in *D. melanogaster* has shown that ∼2.8% of genes have TE-derived promoters in embryos, pupae, and larvae tissues (33).

A recent analysis in *D. melanogaster* ovaries has found 884 chimeric transcripts derived from 549 genes, among

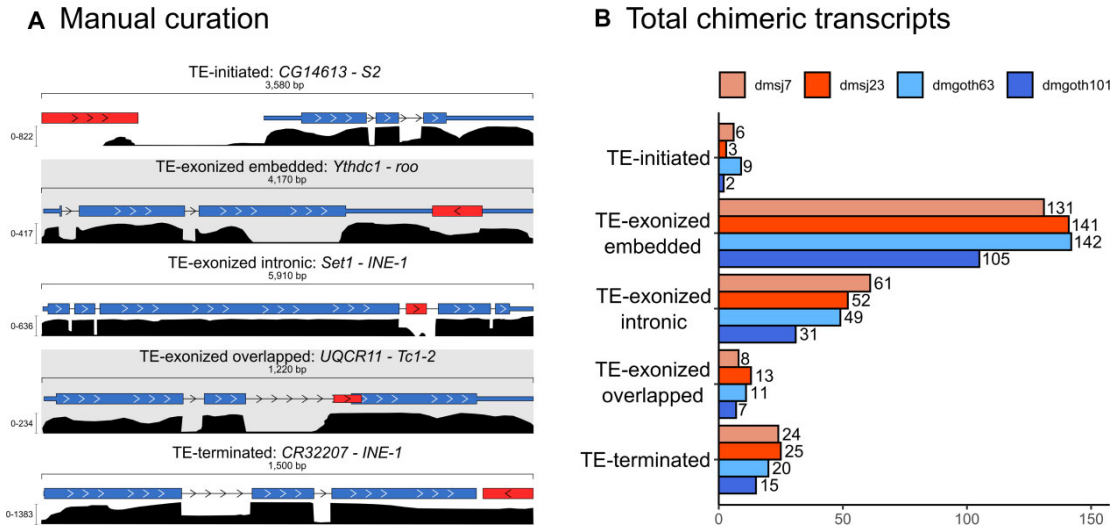


Figure 3. General results from Chimera Mode 1. (**A**) Five examples of chimeric transcripts manually curated with the IGV genome browser. Red boxes: TE insertion; blue boxes: exons and UTRs; black density graphs: coverage of RNA-seq reads; head arrows: transcription sense, blue and red boxes linked by a line: chimeric reads. (**B**) Total genes generating chimeric transcripts, following the TE position classification in the four wild-type strains.

five wild-type strains (81), indicating that a gene is capable of producing multiple chimeric transcripts involving different TE copies. In agreement, TE-exonized chimeric genes detected by ChimeraTE, show evidence of multiple TE sequences acting as chimeric transcripts (22.6% of TE-embedded genes, 30.95% for overlapped, and 30.34% for intronic). Nevertheless, TE-initiated transcripts do not involve more than one TE *per* gene. Among 46 TE-terminated transcripts found in all strains, only five genes were reported with more than one TE copy: (1): *NADH dehydrogenase* (*CG40002*) in dmgoth63; (2,3) *Max* and *rudimentary-like* in dmsj23; and (4,5) *six-banded*, *CG14464* in dmsj23. All of them, except *CG14464*, were associated with two TE insertions from the same TE family, located close to each other (Supplementary Figure 4), suggesting they might be multiple hits from *RepeatMasker* corresponding to only one TE copy (82). Thus, multiple TE copies producing chimeric transcripts exist only when TE copies are within genes. We then tested whether the number of TEs within genes has a correlation with the presence of multiple TE-exonized chimeras. By using all the genes expressed in ovaries (FPKM $>$ 1) with at least one TE insertion within, we found weak positive correlation between them (Supplementary Figure 5), among which the highest was found in dmsj23 (Pearson; $r = 0.22$; $P = 2.6\text{E-}08$), and non-significant for dmgoth101 (Pearson; $r = 0.06$; $P = 0.08$). However, when we take into account the number of TE copies per TE family present within genes, and the respective number of chimeras generated by TE family, we observed a positive correlation (Pearson; $r = 0.87$; $P < 2.2\text{E-}16$) (Supplementary Figure 6). Therefore, TE-rich genes are not the most frequent among the ones generating chimeras, but TE families with high copy number within genes are the ones likely to produce TE-exonized transcripts.

Among all chimeric transcripts predicted by ChimeraTE Mode 1, 11 chimeras have been previously described in *D. melanogaster* using EST data (31). It is important to note that the authors did not consider heterochromatic regions, as well as chimeric transcripts derived from INE-1 elements. From these 11 chimeras, five were found in all strains: *Ssdp* (gene) - *HMS-Beagle* (TE family); *Agpat1-1360*; *anne-1360*; *Atf6-Xelement* and *ctp-HMS-Beagle*. The other six chimeric transcripts are absent from some strains: *CG6191-jockey*; *CG15347-HB*; *CG3164-McClintock; Svil-roo; CHKov1-Doc* and *Kmn1-pogo* (Supplementary Table 7). We manually checked these chimeric transcripts in the IGV genome browser and we confirmed the presence of the TE families predicted by ChimeraTE (Supplementary Table 7). In addition, we also found three chimeras reported previously with RAMPAGE: *CG31999-Tc1-2*; *CkIIα-1366*; *Svil-roo* (33). Finally, in another study on *D. melanogaster* midbrain, 264 genes were shown to produce chimeric transcripts using single-cell RNA-seq data (34). The authors demonstrate that retrotransposons have splice donor and acceptor sites that generate new chimeric isoforms through alternative splicing (34). Despite the differences between tissues and methods, we found 22 chimeric transcripts identified previously using scRNA-seq (34) (Supplementary Table 7), of which two were previously found in ESTs (31). Taken together, the genome-wide analysis performed by ChimeraTE Mode 1 has uncovered 296 genes with chimeric transcripts in the *D. melanogaster* ovarian tissue for the first time.

ChimeraTE mode 2: a method to uncover chimeric transcripts without genome alignment

D. melanogaster has a high rate of TE insertion polymorphism across worldwide population (83–85). To test the ability of ChimeraTE Mode 2 in detecting chimeric transcripts derived from TE insertions absent from a reference genome, we used RNA-seq from the four wild-type strains with transcript and TE annotations from the *D. melanogaster* reference genome (*dm6*). ChimeraTE

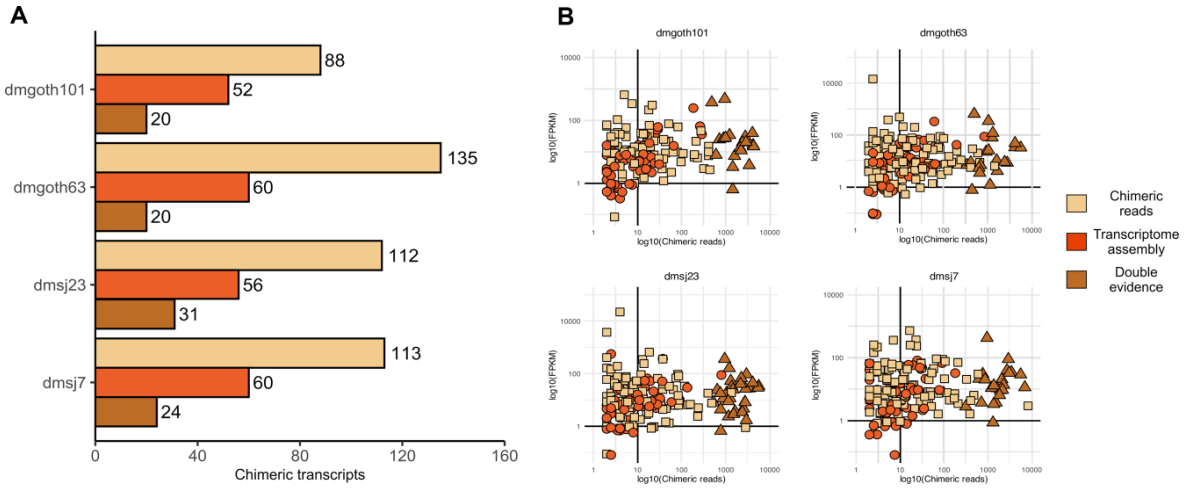


Figure 4. (**A**) Total number of chimeric transcripts found by ChimeraTE mode 2. "Assembled transcripts": chimeric transcripts detected only by the method of transcriptome assembly (Figure 2C). "Chimeric reads": chimeric transcripts detected only by the method of chimeric reads (Figure 2B). "Double evidence": chimeric transcripts detected by both methods. (**B**) Correlation between chimeric read coverage and gene expression of chimeric transcripts found by Mode 2. In all strains, we did not observe correlation between gene expression (FPKM) and the coverage of chimeric reads between the three categories of evidence. All double evidence chimeras were found with high coverage of chimeric reads, suggesting high reliability of chimeras when both methods (chimeric reads and transcriptome) are considered together.

Mode 2 may predict chimeric transcripts based on two types of evidence: either using chimeric reads only or by taking advantage of *de novo* transcriptome assembly. Chimeric transcripts with both evidences are named 'double-evidence' (chimeric reads and transcript assembly - Figure 2B and C). Among the four wild-type strains, ChimeraTE Mode 2 has identified 324 genes (Supplementary Table 8–11) producing 339 chimeric transcripts (Figure 4A), representing 1.81% of the total *D. melanogaster* genes, and 1.07% when normalized by the four wild-type strains. Comparing the 'chimeric read' approach with the 'transcript assembly' one, the former method found ∼51% as many chimeric transcripts than the detection by transcriptome assembly (Figure 4A). This is probably due to the technical limits of *de novo* transcriptome assembly of repetitive-rich transcripts (86), which may lead to the lack of chimeric transcript isoforms. Between both methods, we did not find any correlation with chimeric read coverage and gene expression (FPKM), indicating that even low expressed genes (FPKM < 10) can be detected by both methods independently (Figure 4B). Double evidence chimeras represents ∼14% of all chimeras detected by Mode 2, and they have been found only with high chimeric read coverage (Figure 4B).

As each chimeric transcript was detected based on reference transcripts and TE insertions, we used the Nanopore assemblies to manually inspect the presence of the predicted TE family inside and near (±3 kb) genes, with the intersection of genomic coordinates from genes and TEs with *bedtools intersect* (44). We considered true chimeric transcripts the cases in which we found in the assembly the presence of the predicted TE insertion inside/near the gene. The alignment of RNA-seq libraries against the genome sequence was also used to confirm the presence of the TE insertion, as well as the presence of chimeric reads, with IGV genome browser (63). The manual curation was performed with the three groups of results from Mode 2 ('chimeric reads', 'transcriptome assembly' and 'double-evidence'). We found that 96.30% of all chimeric transcripts predicted by double evidence were true, whereas we observed 90.59% from transcriptome assembly and 73.13% from chimeric reads. Therefore, taken together, ChimeraTE Mode 2 has provided a reliable inference with 81.2% sensitivity (weighted average), based on genomic manual curation.

The main difference of ChimeraTE Mode 2 from the previous pipelines is its ability to detect chimeric transcripts derived from TEs that are absent in a reference genome, using RNA-seq from non-reference individuals/cells/strains. We first identified in the *dm6* reference genome the genes with TEs located upstream, inside, and downstream. We found that 2,239 genes in the *dm6* genome have TEs located 3 kb upstream, 1,863 inside (introns and exons), and 2,320 downstream. These genes were selected as candidates to produce chimeras in the *dm6* genome and then compared with the list of chimeric transcripts generated by ChimeraTE Mode 2 in the four wild-type strains. In addition to the comparison with *dm6* genes harboring TEs inside/near, we have manually curated the TE insertions and the presence of chimeric transcripts with the IGV genome browser (63). Such manual curation was performed with the wild-type genomes, by checking the presence of the TE and chimeric read alignments. For instance, the *Mps1-FB* chimera is an interesting case, since it is the only chimeric transcript specific to French populations, as we found it in both dmgoth101 and dmgoth63. In the *dm6* reference genome, *Mps1* has an overlap between its 3′ UTR and the 3′ UTR of the *alt* gene, located on the other strand (Figure 5A). The same distribution of both genes is found in the Brazilian strains, dmsj23, and dmsj7 (Figure 5B). Conversely, in dmgoth101 and dmgoth63, there is a gap of ∼9,500 bp between *Mps1* and *alt*,

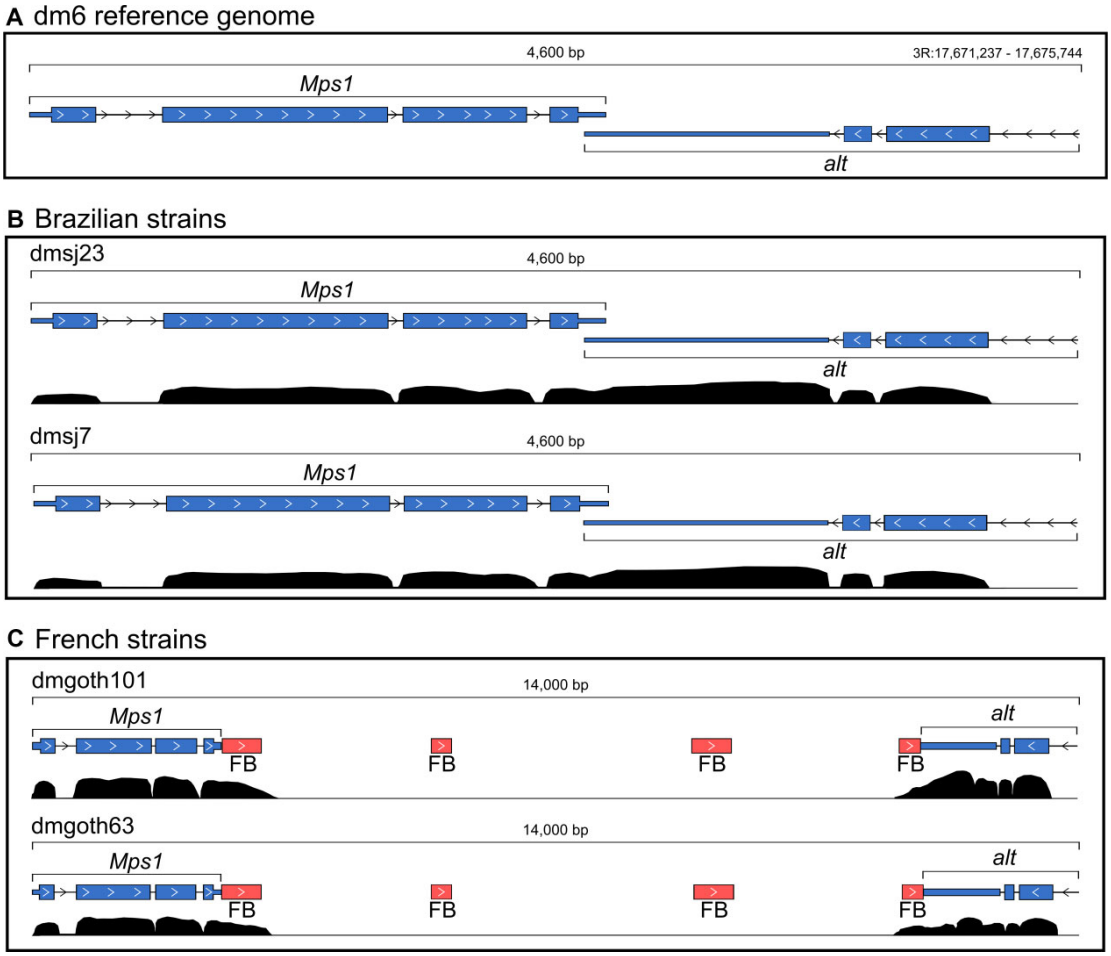


Figure 5. *Mps1* gene and its downstream region in the genomes of the four wild-type strains. (**A**) *Mps1* in the *dm6* reference genome and the *alt* gene located downstream to it, on opposite strands and with overlapped 3' UTRs. (**B**) In the dmsj23 and dmsj7, *Mps1* and *alt* are distributed as found in the reference genome. (**C**) In dmgoth101 and dmgoth63, there is a *FB* insertion located downstream to *Mps1*, which has chimeric reads supporting a TE-terminated downstream in both strains.

with four small *FB* copies of ~412 bp in length (Figure 5C), indicating that it may be an old insertion. Furthermore, one of the *FB* copies is located downstream of the *alt* gene, which has also been identified as TE-terminated downstream in dmgoth63 and dmgoth101. This result shows that ChimeraTE Mode 2 can detect chimeric transcripts derived from TEs that are not present in the reference genome.

Taking into account all chimeric transcripts detected by Mode 2, we found 11 genes (3.40%) generating chimeric transcripts derived from TEs that are absent from the *dm6* genome (Table 1). There are specific chimeras from French strains: *Mps1-FB*, *CG1358-S* and *CG46280-POGON1* genes, but only *Mps1* had chimeric transcripts in both French dmgoth63 and dmgoth101 strains, whereas *CG1358* was found as chimera only in the dmgoth101, and *CG46280* only in dmgoth63. The *Ythdc1-roo* chimera was observed with a strain-specific polymorphic *roo* element inside an exon in the dmgoth63 strain. Regarding the Brazilian strains, we found two TE insertions present in both dmsj23 and dmsj7, involving the genes *cic* and *TrpRS-m*, but they were found as chimeras only in the dmsj23 strain. We also found *rb*, *r-1* and *ArfGAP1* with dmsj23-specific TE insertions, whereas in the dmsj7 we found *caps* and *Ubi-p5E* with specific TE insertions giving rise to chimeric transcripts.

To evaluate Mode 2 accuracy in detecting chimeric transcripts derived from TEs absent in the reference genome, we have performed RT-PCRs for 10 out of 11 cases predicted (it was not possible to design specific primers for the *TrpRSm-protop_a* chimera). The primers were designed considering the TE and the exon location with aligned chimeric reads (Supplementary table 1). For *Mps1-FB*, *caps-pogon1*, *Ythdc1-roo*, *Ubip5E-gypsy*, *CG46280-pogon1*, the amplified fragment sizes were in agreement with ChimeraTE Mode 2 predictions (Supplementary Figure 7A–E). The *r1-stalker2* chimera was amplified in both Brazilian strains, even though the TE insertion generating this chimeric transcript was absent in dmsj7 genome assembly

Table 1. 11 polymorphic chimeric transcripts from TE insertions that are not present in the *dm6* reference genome. The red triangle in a line represents the presence of a chimeric transcript; the white triangle in a line represents the presence of the TE insertion but without a chimeric transcript; the line without triangles represents the absence of the TE insertion

Gene	TE	TE location	Strains			
			dmgoth101	dmgoth63	dmsj23	dmsj7
FBgn0000063 - Mps1	FB4	TE-terminated	▲	▲	—	—
FBgn0003210 - rb	MDG3	TE-terminated	—	—	▲	—
FBgn0003257 - r-l	Stalker2	TE-terminated	—	—	▲	—
FBgn0020655 - ArfGAP1	POGO	TE-terminated	—	—	▲	—
FBgn0023095 - caps	POGON1	TE-exonized	—	—	—	▲
FBgn0027616 - Ythdc1	ROO	TE-exonized	—	▲	—	—
FBgn0033196 - CG1358	S	TE-terminated	▲	△	—	—
FBgn0036763 - TrpRS-m	PROTOP_A	TE-exonized	—	—	▲	△
FBgn0086558 - Ubi-p5E	Gypsy	TE-initiated	—	—	—	▲
FBgn0262582 - cic	POGON1	TE-exonized	—	—	▲	△
FBgn0283438 - CG46280	POGON1	TE-exonized	△	▲	—	—

(Supplementary Figure 8A). We hypothesized that this might be due to a low-frequency TE in the dmsj7, absent from the assembly and without enough read coverage in the RNA-seq to predict it. Indeed, we confirmed the presence of this TE insertion by PCR in both Brazilian strains (Supplementary Figure 8B). The chimeric transcript from *rb-mdg3* was detected only in dmsj23 by ChimeraTE Mode 2, but the RT-PCR assay was successful only in the dmsj7 strain (Supplementary Figure 9). The chimera derived from *CG1358-S* was predicted only in dmgoth101 (Supplementary Figure 10), despite the TE insertion being also present in dmgoth63 genome. By RT-PCR, we detected this chimeric transcript in both strains, suggesting that it was not detected in dmgoth63 by ChimeraTE due to Illumina coverage issues. A similar result was found for the *cic-pogon1* chimera, for which we amplified the chimeric transcript in both Brazilian strains, but it was detected only in dmsj23 by ChimeraTE (Supplementary Figure 11). Altogether, these results indicate that Mode 2 can predict chimeric transcripts from TEs absent from the reference genome, but this prediction depends upon the coverage of chimeric reads in the RNA-seq data.

Evaluation of chimeric transcripts in dmgoth101 through nanopore RNA-seq

Short-read RNA-seq may introduce biases during library preparation (e.g. reverse transcription, second-strand synthesis, PCR amplification) (87), or even during transcriptome assembly, with loss of information regarding isoform diversity, especially for those with low-expression and repetitive regions (88), which might be the case of many chimeric transcripts. In the past years, the third generation of long-read sequencing has been used to mitigate such biases, because there is no need to perform transcriptome assembly. Thus, we validated chimeric transcripts predicted by ChimeraTE Mode 1 with short reads using long reads produced with Oxford Nanopore Technology (ONT) from dmgoth101 ovaries (89). We did not perform the same validation with RNA long reads for chimeras from Mode 2 predicted with short reads, due to the lack of information regarding the genomic position of TEs generating them. The long reads were aligned to the dmgoth101 genome and then checked for the presence of all chimeric transcripts detected by ChimeraTE Mode 1 in dmgoth101. The presence of the chimeric transcripts was confirmed when the long read was aligned into the gene exons, and it had the TE-derived sequence as part of the reads. From two TE-initiated transcripts, we found only *CG14613-S2* chimera by ONT reads. In TE-terminated transcripts, six (40%) out of 16 were also observed by ONT data. Finally, in TE-exonized transcripts, we identified 89 (62.24%) out of 143 embedded TEs, 12 (52.17%) out of 23 overlapped TEs, and 13 (46.43%) out of 28 intronic. The undetected chimeras through ONT reads might be associated with the long-read library construction protocol, TeloPrime, which was used to generate full-length cDNA libraries. Such a protocol might miss weakly expressed isoforms, as well as truncated 3′ ends due to RNA degradation *in vivo* or during RNA manipulation (90). We investigated whether short-read coverage could explain the lack of detection of certain chimeric transcripts with ONT data. Taking all dmgoth101 chimeras from Mode 1, chimeric transcripts found by both ChimeraTE and ONT data had an average chimeric read coverage of 69.44, whereas the undetected ones had 10.83. Therefore, these results indicate that the detection of chimeras with ONT reads is likely to be more efficient with highly expressed chimeric transcripts. Overall, our results demonstrate that 50.16% of all chimeric transcripts detected in the ovary transcriptome of *D. melanogaster* with ChimeraTE Mode 1 can also be found by ONT RNA-seq.

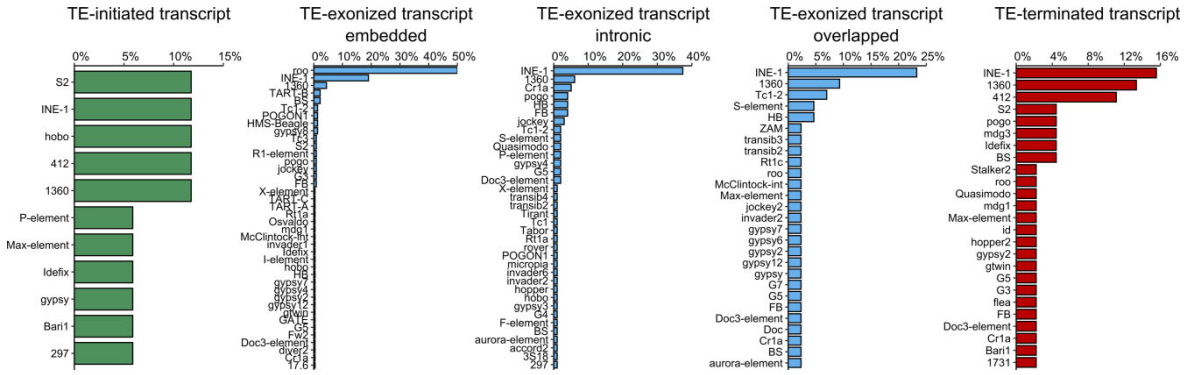


Figure 6. The frequency of the 76 TE families generating chimeric transcripts in the four wild-type strains. In chimeric transcripts derived from TEs near genes, *INE-1* was the most frequent (15%) in TE-terminated downstream, whereas for TE-initiated upstream, *S2*, *INE-1*, *hobo*, *1360* and *412* had the same frequency (11%). Regarding TEs inside genes, the *roo* element has the highest frequency of TE-exonized embedded, representing 50% of all chimeras. In TE-exonized intronic and overlapped, *INE-1* was the most frequent TE family.

The high prevalence of *roo* elements in chimeric transcripts is due to ∼132 nt TE insertions embedded into exons

Several TE families have been associated with chimeric transcripts in *D. melanogaster* (31,34). Here, we found 76 TE families producing at least one chimeric transcript in the four wild-type strains, representing 61.79% of all TE families annotated. We found a positive correlation between the frequency of TE families in chimeric transcripts and their respective abundance in the *D. melanogaster* genomes (Pearson; $r = 0.53$; $P < 2.2E\text{-}16$). In the TE-initiated transcripts, *S2*, *INE-1*, *hobo*, *1360* and *412* had the highest frequencies (Figure 6). In TE-exonized transcripts, the prevalence of *roo* elements is more pronounced in the embedded TEs group, comprising 50% of all chimeras, followed by *INE-1* elements, with 19%.

Interestingly, for both TE-exonized intronic and overlapped, *INE-1* had the highest frequencies, with 37.37% and 23.25%, respectively (Figure 6); whereas *roo* element was observed in only one chimera (*Lk6 - roo*, in dmgoth63 and dmsj23) as a TE-exonized overlapped. Notably, the *1360 element* is among the most frequent TE families in all TE-exonized categories. The other 38 families have an average frequency of 2%. Finally, in TE-terminated transcripts, *INE-1* is the most frequent TE, with 15.5% of all cases, followed by *1360* and *412*, with 13% and 11%, respectively. Taken together, these results suggest that *INE-1, S2, 412* and *1360* are candidate TE families to investigate the exaptation of regulatory motifs since they were the most abundant in TE-initiated and TE-terminated transcripts. On the other hand, in TE-exonized transcripts, *roo* might have a potential for exaptation/domestication of both regulatory and protein domain motifs, since it is observed with high prevalence within exons (UTRs and CDSs). Regarding TE-exonized transcripts derived from TEs at intronic regions and overlapping with exons, *INE-1* and *1360* are the families with the highest frequency of intron retention, suggesting they might harbour splice sites allowing incorporation into the gene transcripts.

We then investigated whether exonized TE insertions generating TE-exonized transcripts could contribute to protein domains, especially *roo* due to its high prevalence in TE-exonized embedded chimeras. Through our analysis with *pfam-scan*, we found in total 28 genes with TE-exonized transcripts from TE insertions harbouring a TE protein domain, corresponding to an average of ∼16 per strain (Supplementary Table 12). We assessed the completeness of these protein domains conserved on these insertions, assuming domains with >75% of completeness as complete, and <75% as partial. Our analysis revealed that ∼45% of all domains are likely to be conserved (Figure 7A), for which 92% correspond to catalytic domains, and 8% to zinc finger binding domains (PRE_C2HC – PF07530.14; zf-CCHC – PF00098.26). In addition, 17.69% of all TE-derived protein domains stem from embedded TE insertions, 27.43% from overlapped with exon, and 54.86% from intron regions. Furthermore, only one *roo* insertion had protein domains detected (Peptidase A17; DUF5641), which is embedded within CR46472 long noncoding RNA gene. Taking into account that 50% of the TE-exonized transcripts are derived from embedded *roo* insertions, this result suggests that despite their high frequency, they are old insertions without catalytic potential.

Since we predict these chimeric transcripts only based on chimeric reads, we are not able to confirm whether these domains have been fully incorporated into gene transcripts, limiting our biological conclusions regarding catalytic or binding effects. There are only three chimeric transcripts with TE-derived protein domains for which we found the whole TE sequence incorporated into the transcript with ONT RNA-seq data (Supplementary Figure 12). Among them, *CG4004 – hobo*, and *nxf2 – TART-A* are present into CDS regions, suggesting a putative functional role of the TE sequences. The third chimera, *CR42653 – Doc*, is a non-coding pseudo-gene, which suggests that the TE-derived protein domain does not have a functional role.

TE insertions disrupting coding regions and promoters are more likely to be deleterious, being consequently eliminated by purifying selection (91), although several exaptation events have been documented (23,92). The presence of *roo* elements in 50% of all TE-exonized embedded transcripts suggests a neutral or adaptive role of this family when incorporated into gene transcripts. *Roo* is an LTR

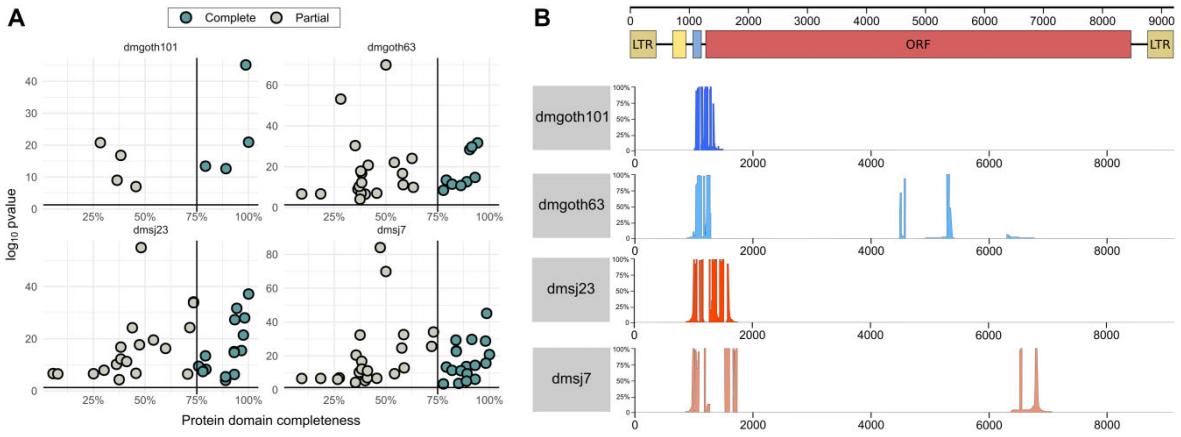


Figure 7. (**A**) Completeness of protein domains identified in TE insertions generating TE-exonized transcripts, with p-value < 0.05. (**B**) Alignment depth with *roo* consensus and embedded *roo* insertions generating TE-exonized transcripts. At the top, the scheme of the full-length roo element: brown boxes: LTRs; yellow box: first tandem repeats at 5' UTR; blue box: second tandem repeat at 5' UTR; red box: Open reading frame (ORF). The coverage depth of the multiple alignments between embedded roo insertions and the consensus is separated by strain.

retrotransposon, encoding three proteins: *gag*, *pol*, and *env*, which have been through domestication events from retroelements in many species, including *Drosophila* (8,93–94). It is the most abundant euchromatic TE family of *D. melanogaster* (1,95,96), and its insertions have been associated with modifications in the expression level of stress response genes due to the presence of TFBSs (97). They have also low enrichment of repressive histone marks (98,99), potentially explaining their high transposition rates (100,101). We analyzed the *roo* sequences that are embedded into exons of TE-exonized transcripts. The length of these chimeric insertions is ∼132 bp, confirming that they are old insertions since the full-length *roo* consensus sequence is 9,250 bp (Figure 7B). Subsequently, we analyzed whether these exonized *roo* insertions are donors of preferential motifs to the chimeric transcripts. All *roo* insertions stem from a specific region, between the 5′ UTR and the *roo* open-reading frame (Figure 7B). Compared to a *roo* consensus sequence, in dmgoth63 and dmsj7, most insertions show a deletion from the 5′ UTR up to the middle of the ORF. Despite the low nucleotide diversity of *roo* insertions in *the D. melanogaster* genome, the 5′ UTR region has a hypervariable region, including deletions and repeats, with several copies missing a tandem repeat of 99 bp (57,102). It has been proposed that this region may have a role in *roo* transposition, by heterochromatinization, recruitment of RNA pol II, and interaction with other enzymes (102). Curiously, a study assessed the nucleotide diversity between *roo* insertions, and characterized the same region as a deletion hotspot (78). A recent work also observed this region as part of TE-exonized transcripts in a transcriptome-wide manner in *D. melanogaster* (81). Why this region has been maintained through evolution, and whether it could be adaptive or neutral, remains unclear.

Polymorphic TE insertions in the wild-type strains generate 76 chimeric transcripts

Polymorphic TE insertions are common across *D. melanogaster* strains (103,104). To quantify how many chimeric transcripts in the wild-type strains are derived from polymorphic TE insertions compared to the reference genome, we used the list of genes with TEs located 3 kb upstream, inside (introns and exons), and 3 kb downstream generated previously for the dm6 genome. These genes were selected as potential sources of chimeras and then compared with the list of chimeric transcripts generated by ChimeraTE Mode 1 in the four strains. We found that 76 genes (23.24%) with chimeric transcripts in the wild-derived strains were generated by TE insertions that are absent in the reference genome (Supplementary Table 13). The number of genes generating chimeras from polymorphic TEs found by Mode 1 is higher than Mode 2 (Table 1; 11 genes). Such result is likely due to the divergence of TE insertions from the reference genome that is used with Mode 2, decreasing the alignment rate, whereas with Mode 1 we use strain-specific insertions. From the 76 genes, we found nine out of the 11 chimeras derived from polymorphic insertions identified by Mode 2 (Table 1). The chimeric transcripts from *CG1358* and *caps* genes were not found by Mode 1, despite the TEs presence on the genomes. Considering the 76 chimeras, we observed twenty-seven TE families generating polymorphic chimeric transcripts: *roo* (44.15%), *412* (6.49%), *pogo* (6.49%), *POGON1* (6.49%), among others. Except for *INE-1*, for which we found two polymorphic chimeras, most of the other TE families are known to be active in *D. melanogaster* (57). *INE-1* chimeras are unexpected since it is an old and inactive TE family in *D. melanogaster* (105), however, *INE-1* polymorphism among *D. melanogaster* populations has previously been shown (106).

The study of TEs in wild-type strains offers new insights into their ability to provide genetic variability. Depending upon the position where TEs are inserted regarding genes, they can contribute to gene expression or protein sequence variation. We then compared whether polymorphic TE insertions generating chimeric transcripts are more likely inserted near or inside genes. We found that 5.19% correspond to TEs located upstream, 74.03% inside (intron/exons) and 20.78% downstream (Figure 8). In the

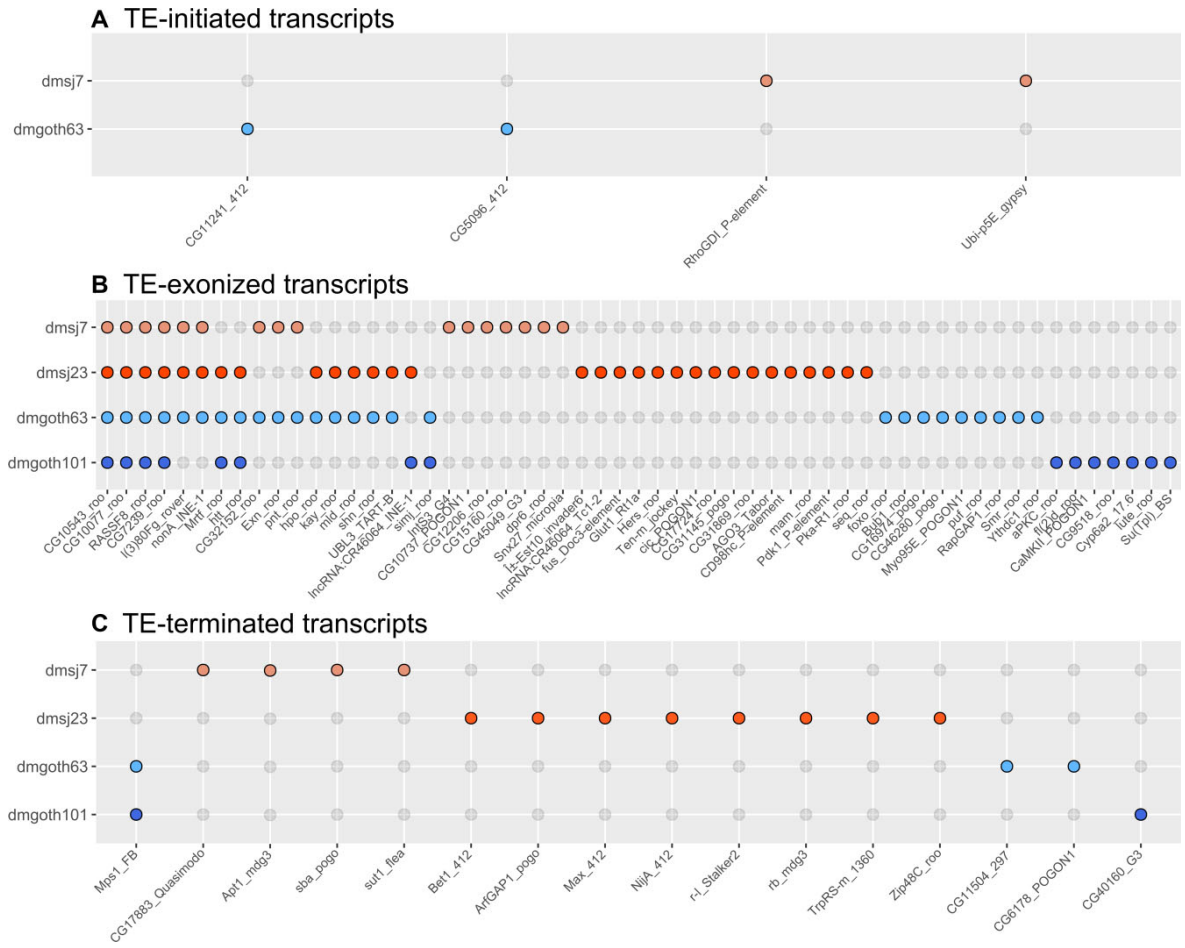


Figure 8. The 76 chimeric transcripts derived from TE insertionsthat are absent in the dm6 reference genome, 5.19% of them correspond to TEs located upstream, 74.03% to TEs located inside genes (introns and exons), and 20.78% to TEs located downstream. (**A**) TE upstream: Chimeric transcripts in which the TE is located up to 3kb upstream of the gene. (**B**) TE inside: Chimeric transcripts with TE insertions located inside the gene region (exons and introns). There are four chimeric transcripts found in all strains, and one specific to French strains. (**C**) TE downstream: Chimeric transcripts in which the TE is located up to 3kb downstream of the gene. Only Mps1-FB was specific to French strains, whereas all the other 15 chimeras are strain-specific.

genomic context, we observed ∼1,767 polymorphic TE insertions in these three regions regardless of the presence of chimeric transcripts, for which we found an equal distribution between TEs upstream (∼32%), TEs inside genes (∼35%) and TEs downstream (∼32%). Such result suggests that polymorphic insertions are not preferentially maintained by selection in specific gene locations, but chimeric transcripts from polymorphic TEs are more likely to be generated when they are within gene regions. In addition, we investigated whether the TE insertions absent from the reference genome could either be population-specific (Brazil or France) or strain-specific. The four TE-initiated transcripts derived from polymorphic TE insertions were strain-specific, two from dmgoth63 and two from dmsj7 (Figure 8A). For TEs inserted inside genes, 66.67% were found only in one strain. Only four chimeric transcripts are common to all strains *CG10543 - roo*, *CG10077 - roo*, *RASSF8 - roo* and *CG7239 - roo* (Figure 8B). We did not find a TE-exonized derived from a TE insertion present only in the Brazilian strains, but we did find one for the French strains, the *simj – roo* chimera. Interestingly, nine chimeras were found between pairs of Brazilian and French strains (Figure 8B), indicating retention of ancestral polymorphisms in these populations due do incomplete lineage sorting (107). Finally, for TEs located downstream genes (Figure 8C), the chimera *Mps1–FB* is the only one found in both French strains, and absent from Brazilian strains, as we observed by Mode 2 (Figure 5). Taken together, these results reinforce the potential of TEs in generating genetic novelty between *D. melanogaster* strains.

Modification in gene expression is not a general rule for chimeric transcripts

TEs can modify gene expression when inserted inside and in their close vicinity. We hypothesized that genes with

strain-specific chimeric transcripts may have differences in their expression levels in comparison to strains without the same chimeras. We first identified 67 strain-specific chimeras per strain, through pairwise comparisons. Then, through differential expression analysis between the four strains, we found that most genes have a similar expression level regardless of the presence of chimeric transcripts (Figure 9). In average, only 8 genes per strain were differentially expressed (Figure 9). Taking all these genes together, these chimeras are composed of 11.34% TE-initiated, 76.28% TE-exonized and 12.37% TE-terminated.

Among differentially expressed genes generating strain-specific chimeras, we observed that ∼47% had $\log_2$ fold-change <|2|, representing a mild transcription modulation. The other ∼53% (36 genes) had on average a log2 fold-change of ∼4.5 indicating higher expression when the chimera is present. Nevertheless, the gene expression quantification does not take into account differences between chimeric and non-chimeric isoforms. Therefore, we cannot differentiate gene expression differences due to chimeric isoform expression. We then focused on differentially expressed genes that are expressed in a given strain, but silent in the other (log2 normalized counts < 1). Only two chimeric cases fell into such category: TE-exonized CR45600 – *Tc1-2; gypsy7* in dmgoth101 compared to dmsj7 (Figure 9A); and the TE-terminated CG30428 – *INE-1* in dmgoth101, dmsj23, and dmsj7 compared to dmgoth63 (Figure 9A, C, D). To verify whether the chimeric transcript has a relatively high contribution to the gene expression, we compared the depth of uniquely mapped reads between exons and the TE insertion. For the CR45600 – *Tc1-2; gypsy7* chimera, we found a similar depth (Supplementary Figure 13A), indicating that the chimeric transcript isoform represents a major contribution to the gene expression; whereas for CG30428 – *INE-1* chimera, the depth in *INE-1* is half of CG30428 exons (Supplementary Figure 13B), suggesting that the TE sequence contributes with only ∼50% of gene expression.

In dmgoth101, we found the gene *Cyp6a2* differentially expressed and involved in a TE-exonized transcript with a *17.6* insertion. Such chimera is absent from dmsj23, dmsj7, and dmgoth63 transcriptome. We checked whether *17.6* could be present on the genome of the other strains and we observed it only in dmgoth63, but undetected by ChimeraTE because *Cyp6a2* is not expressed. The presence of the *17.6* insertion within the *Cyp6a2* 3′UTR generating a chimeric transcript was initially proposed to cause the overexpression of this gene in flies resistant to xenobiotics (108), but later the opposite has been shown (109). Furthermore, regardless of *17.6* presence, the overexpression of this gene has been associated with xenobiotic resistance in *D. melanogaster* strains (110). Here, the differential expression analysis reveals that *Cyp6a2* is up-regulated in dmgoth101 in comparison to dmgoth63 (adj. *P*-value: 0.005; $\log_2$ FC: 4.38), but no significant differences were found when comparing with strains without the *17.6* insertion, dmsj23 (adj. *P*-value: 0.29), and dmsj7 (adj. *P* value: 0.15), reinforcing the lack of contribution of *17.6* insertion to *Cyp6a2* expression (109). Finally, the chimera derived from *Cyp6a14*, with a *1360* insertion embedded into the 3′ UTR has been found only in dmsj7, in comparison to the other three strains. This gene is associated with xenobiotic resistance (111), as *Cyp6a2*. Interestingly, our results reinforce previous findings showing that *Cyp* genes have accumulated more TE insertions when compared to a random sample of genes (112), potentially as a source of alternative regulatory motifs.

Collectively, the presence of chimeric transcripts might not drive gene differential expression. Despite eight differentially expressed genes per strain, the transcriptional activation of specific genes in one genotype may be determined by its genetic background, which can be associated with structural variants as TEs, but also includes pleiotropic effects from regulatory networks, cell physiology and epigenetics (113). Nevertheless, given that TEs might provide TFBSs and epigenetic marks to chromatin accessibility, we propose these differentially expressed genes as candidates to investigate the domestication of regulatory networks in further studies, especially the ones with TEs inserted at regulatory regions.

ChimeraTE uncovers chimeric transcripts in model and non-model species

We sought to demonstrate that both Modes of ChimeraTE are able to identify chimeric transcripts regardless of the species. Therefore, we used ChimeraTE with human, *A. thaliana* and *P. reticulata* (guppy fish) data. In Mode 1, the RNA-seq of the human K562 cell line revealed 67 TE-initiated, 14,999 TE-exonized, and 166 TE-terminated transcripts, derived from 5,170 genes (10.45% of total). Furthermore, Mode 2 revealed 4,372 chimeric transcripts. From these, 3,405 chimeras were detected from chimeric reads, 342 from transcriptome assembly, and 625 from double evidence. The overlap from chimeras found by both Modes was 2,997 (68.55%), suggesting a high amount of chimeric transcripts derived from TE insertions that are not present in the reference genome, or from TEs located farther away than the chosen 3 kb gene flanking regions. In *A. thaliana*, we used ChimeraTE with RNA-seq from leaf tissue. Our results from Mode 1 demonstrated that 266 genes (0.7% out of 38,319) generate 397 chimeras, corresponding to 24 TE-initiated, 338 TE-exonized, and 35 TE-terminated transcripts. Mode 2 has detected 141 genes generating chimeras, for which 63.12% were from chimeric reads, 24.11% from transcriptome assembly, and 12.77% from double evidence. Similar to human, Mode 2 found 39.72% of chimeras detected by Mode 1. Finally, the guppy transcriptome was analyzed with RNA-seq produced from ovary follicular tissue. For the first time, we identified 1,151 genes (5.85% of total genes) generating chimeric transcripts in this species. From these, 39 were TE-initiated, 1,507 TE-exonized, and 136 TE-terminated transcripts. Mode 2 revealed 640 genes with chimeras, having an overlap with 318 genes (49.69%) detected by Mode 1. Taking together our results from human, *A. thaliana*, guppy and *D. melanogaster* data, we demonstrated that ChimeraTE can identify chimeras in different species. We then analyzed how the genome size could affect ChimeraTE Mode 1 and Mode 2 in the processing time, and we have observed a positive correlation between them in

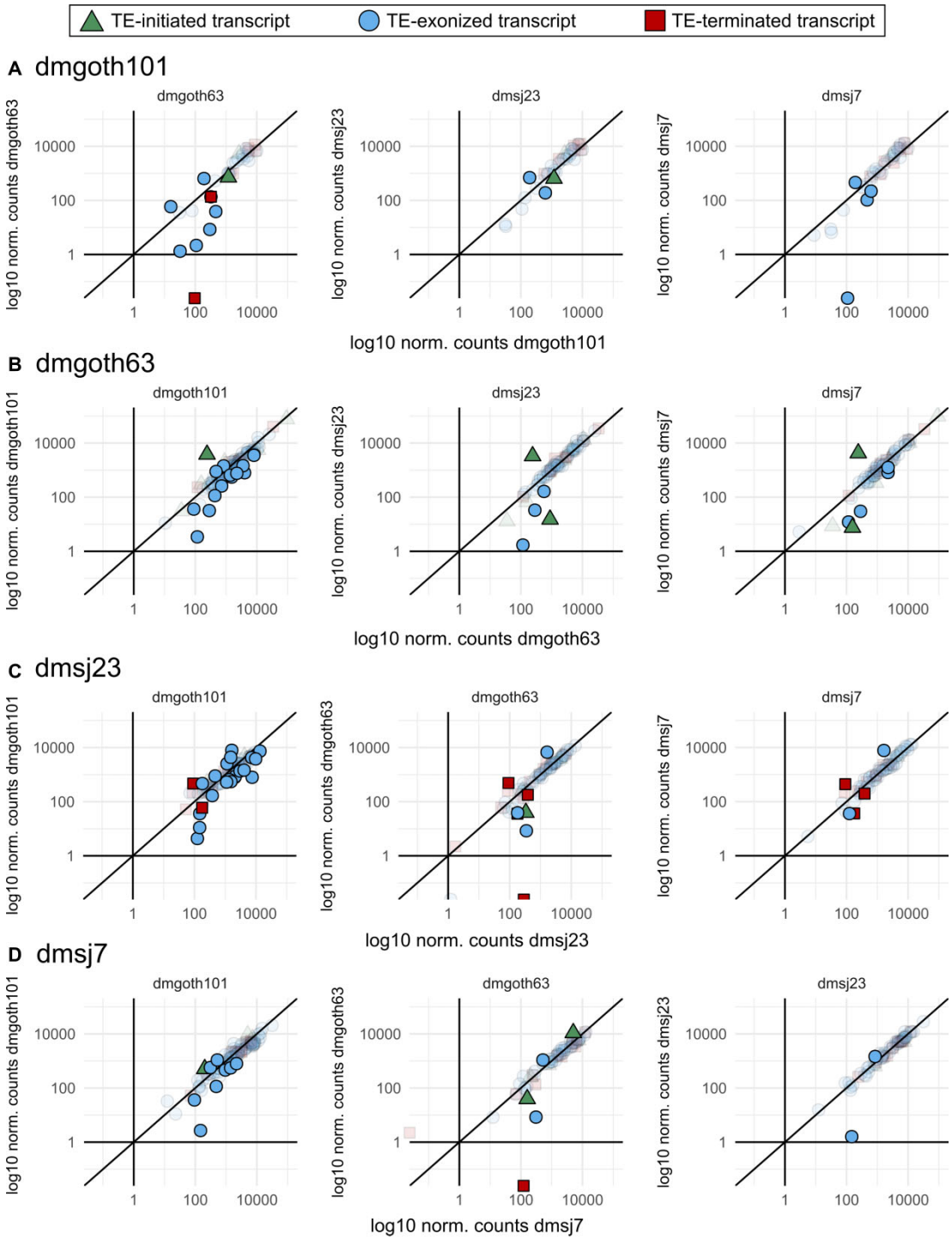


Figure 9. Normalized counts from DEseq2 indicating gene expression of genes generating polymorphic chimeric transcripts among the four wild-type strains. Colorful forms represent differentially expressed genes (adj. p-value < 0.05); transparent forms represent non-differentially expressed genes (adj. p-value). (**A**) Expression level of genes producing chimeric transcripts only in dmgoth101, in comparison to dmgoth63, dmsj23, and dmsj7. (**B**) Expression level of genes producing chimeric transcripts only in dmgoth63, in comparison to dmgoth101, dmsj23, and dmsj7. (**C**) Expression level of genes producing chimeric transcripts only in dmsj23, in comparison to dmgoth101, dmgoth63, and dmsj7. (**D**) Expression level of genes producing chimeric transcripts only in dmsj7, in comparison to dmgoth101, dmgoth63, and dmsj23. Overall, gene expression does not change when comparing strains with and without chimeric transcripts.

Mode 1 (Pearson; $r = 0.99$; $P = 0.01$), as well as for in Mode 2 (Pearson; $r = 0.99$; $P = 0.009$) (Supplementary Figure 14).

CONCLUSION

In the last decades, RNA-seq has provided the opportunity to understand transcriptome plasticity, which can lead to phenotypic divergence, from related species or individuals from a population (114). Among several sources of modification in gene expression and novel isoform transcripts, TEs have been considered fundamental suppliers of transcriptome plasticity, participating in gene networks and incorporating either regulatory sequences or protein domains into gene transcripts. The identification of chimeric transcripts is an important step to the understanding of transcriptome plasticity, since they may be triggered by ectopic conditions, such as cancer, oxidative stress, and heat shock (26,115,116), which may lead to both detrimental and advantageous outcomes (117). Therefore, uncovering the extent of chimeric transcripts between individual/cell/strain transcriptomes is a crucial first step to investigate potential exaptation/domestication events, or gene disruption and loss of function.

Chimeric transcripts have been identified more recently by different methods exploiting RNA-seq data (34–36,81), but none of them provided the possibility to predict chimeras from TEs that are absent from reference genomes. Here, we developed ChimeraTE, a pipeline able to identify chimeric transcripts from TEs that are absent from the reference genome. Mode 1 is a genome-guided approach and may be used either when the user is not interested in chimeras derived from TEs absent from the reference genome, or when the user has a high-quality genome assembly for each individual/cell/strain; whereas Mode 2 is a genome-blind approach, with the ability to predict chimeric transcripts without the assembled genome, but missing chimeras where the TE is smaller than the length of the reads.

We analyzed ovarian RNA-seq from four *D. melanogaster* wild-type strains, for which we have genome assemblies. Altogether, we found that ∼1.12% of all genes are generating chimeric transcripts in ovaries, following the proportion from previous studies obtained with ESTs, and RNA-seq in midbrain tissue (31,34). Furthermore, our results revealed that 50% of all TE-exonized transcripts with TEs embedded derive from *roo* elements, and most specifically, a small region between tandem repeats in the 5′ UTR and the beginning of the *roo* ORF. These results suggest that these *roo* insertions have neutral or advantageous effects as they are maintained within these gene transcripts. However, we did not provide enough support to claim these *roo*-exonized transcripts as exaptation or domestication events, due to the lack of evidence regarding the functional role of these chimeras. Further studies must be performed to clarify this subject, mainly because chimeric transcripts detected by ChimeraTE can be degraded by surveillance pathways that degrade aberrant mRNA, such as non-sense-mediated mRNA decay (118), no-go decay (119), and non-stop decay (120).

Altogether, this new approach allows studying the impact of new mobilization events between populations or between treatment conditions, providing insights into biological questions from a broad community of researchers, ranging from cancer research, to population transcriptomics, and adaptation studies. ChimeraTE implementation will be useful for the next discoveries regarding the evolutionary role of TEs and their impact on the host transcriptome.

DATA AVAILABILITY

ChimeraTE is an open-source collaborative initiative available in the Zenodo repository (https://doi.org/10.5281/zenodo.8143045). The RNA-seq data used in this study are available in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/), under PRJNA795668 accession, while the long-read ONT data is available under PRJNA956863.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Josefa González, Marta Coronado-Zamora, Dixie Mager and Vincent Lacroix for useful discussions and advice. This work was performed using the computing facilities of the CC LBBE/PRABI.

FUNDING

Agence Nationale de la Recherche [Exhyb ANR-14-CE19-0016-01 and LongevitY ANR-20-CE02-0015 to C.V.]; Fondation pour la Recherche Médicale [DEP20131128536 to C.V.]; Idex Lyon fellowship to D.S.O.; Campus France Eiffel [Eiffel Fellowship P769649C to D.S.O.]; TIGER [H2020-MSCA-IF-2014-658726 to R.R.]; National Council for Scientific and Technological Development [308020/2021-9 to C.M.A.C.]; São Paulo Research Foundation [2020/06238-2 to C.M.A.C.]. Funding for open access charge: Fondation pour la Recherche Médicale [DEP20131128536] and Agence Nationale de la Recherche [ExHyb, ANR-14-CE19-0016-01].

Conflict of interest statement. None declared.

REFERENCES

1. Quesneville,H., Bergman,C.M., Andrieu,O., Autard,D., Nouaud,D., Ashburner,M. and Anxolabehere,D. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.*, **1**, e22.
2. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research, Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
4. Sotero-Caio,C.G., Platt,R.N., Suh,A. and Ray,D.A. (2017) Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evolut.*, **9**, 161–177.
5. Danilevskaya,O.N., Arkhipova,I.R., Pardue,M.L. and Traverse,K.L. (1997) Promoting in tandem: the promoter for telomere transposon HeT-A and implications for the evolution of retroviral ltrs. *Cell*, **88**, 647–655.

6. Kapitonov,V.V. and Jurka,J. (2005) RAG1 Core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.*, **3**, e181.
7. Kapitonov,V.V. and Koonin,E.V. (2015) Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol. Direct*, **10**, 20.
8. Volff,J.-N. (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays*, **28**, 913–922.
9. Babaian,A., Romanish,M.T., Gagnier,L., Kuo,L.Y., Karimi,M.M., Steidl,C. and Mager,D.L. (2016) Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene*, **35**, 2542–2546.
10. Daborn,P.J., Yen,J.L., Bogwitz,M.R., Le Goff,G., Feil,E., Jeffers,S., Tijet,N., Perry,T., Heckel,D., Batterham,P. *et al.* (2002) A single P450 allele associated with insecticide resistance in *Drosophila*. *Science*, **297**, 2253–2256.
11. Jordan,I.K., Rogozin,I.B., Glazko,G.V. and Koonin,E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.*, **19**, 68–72.
12. Mateo,L., Ullastres,A. and González,J. (2014) A transposable element insertion confers xenobiotic resistance in *Drosophila*. *PLos Genet.*, **10**, e1004560.
13. Modzelewski,A.J., Shao,W., Chen,J., Lee,A., Qi,X., Noon,M., Tjokro,K., Sales,G., Biton,A., Anand,A. *et al.* (2021) A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. *Cell*, **184**, 5541–5558.
14. Capy,P. (2021) Taming, domestication and exaptation: trajectories of transposable elements in genomes. *Cells*, **10**, 3590.
15. Fueyo,R., Judd,J., Feschotte,C. and Wysocka,J. (2022) Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.*, **23**, 481–497.
16. Lamprecht,B., Walter,K., Kreher,S., Kumar,R., Hummel,M., Lenze,D., Köchert,K., Bouhlel,M.A., Richter,J., Soler,E. *et al.* (2010) Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.*, **16**, 571–579.
17. McGinnis,W., Shermoen,A.W. and Beckendorf,S.K. (1983) A transposable element inserted just 5′ to a *Drosophila* glue protein gene alters gene expression and chromatin structure. *Cell*, **34**, 75–84.
18. Almeida,L.M., Amaral,M.E.J., Silva,I.T., Silva,W.A. Jr, Riggs,P.K. and Carareto,C.M. (2008) Report of a chimeric origin of transposable elements in a bovine-coding gene. *Genet. Mol. Res.*, **7**, 107–116.
19. Sela,N., Mersch,B., Hotz-Wagenblatt,A. and Ast,G. (2010) Characteristics of transposable element exonization within human and mouse. *PLoS One*, **5**, e10907.
20. Sorek,R. (2007) The birth of new exons: mechanisms and evolutionary consequences. *RNA*, **13**, 1603–1608.
21. Bogwitz,M.R., Chung,H., Magoc,L., Rigby,S., Wong,W., O'Keefe,M., McKenzie,J.A., Batterham,P. and Daborn,P.J. (2005) Cyp12a4 confers lufenuron resistance in a natural population of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 12807–12812.
22. Farré,D., Engel,P. and Angulo,A. (2016) Novel role of 3'UTR-embedded Alu elements as facilitators of processed pseudogene genesis and host gene capture by viral genomes. *PLoS One*, **11**, e0169196.
23. Magwire,M.M., Bayer,F., Webster,C.L., Cao,C. and Jiggins,F.M. (2011) Successive increases in the resistance of *Drosophila* to viral infection through a transposon insertion followed by a duplication. *PLoS Genet.*, **7**, e1002337.
24. Cordaux,R., Udit,S., Batzer,M.A. and Feschotte,C. (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8101–8106.
25. Ehrlich,M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239–259.
26. Babaian,A. and Mager,D.L. (2016) Endogenous retroviral promoter exaptation in human cancer. *Mobile DNA*, **7**, 24.
27. Lock,F.E., Rebollo,R., Miceli-Royer,K., Gagnier,L., Kuah,S., Babaian,A., Sistiaga-Poveda,M., Lai,C.B., Nemirovsky,O., Serrano,I. *et al.* (2014) Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E3534–E3543.
28. Shah,N.M., Jang,H.J., Liang,Y., Maeng,J.H., Tzeng,S.-C., Wu,A., Basri,N.L., Qu,X., Fan,C., Li,A. *et al.* (2023) Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat. Genet.*, **55**, 631–639.
29. Faulkner,G.J., Kimura,Y., Daub,C.O., Wani,S., Plessy,C., Irvine,K.M., Schroder,K., Cloonan,N., Steptoe,A.L., Lassmann,T. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571.
30. Babarinde,I.A., Ma,G., Li,Y., Deng,B., Luo,Z., Liu,H., Abdul,M.M., Ward,C., Chen,M., Fu,X. *et al.* (2021) Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res.*, **49**, 9132–9153.
31. Lipatov,M., Lenkov,K., Petrov,D.A. and Bergman,C.M. (2005) Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. *BMC Biol.*, **3**, 24.
32. Batut,P. and Gingeras,T.R. (2013) RAMPAGE: promoter activity profiling by paired-end sequencing of 5′-complete cDNAs. *Curr. Protoc. Mol. Biol.*, **104**, 25.
33. Batut,P., Dobin,A., Plessy,C., Carninci,P. and Gingeras,T.R. (2013) High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.*, **23**, 169–180.
34. Treiber,C.D. and Waddell,S. (2020) Transposon expression in the *Drosophila* brain is driven by neighboring genes and diversifies the neural transcriptome. *Genome Res.*, **30**, 1559–1569.
35. Pinson,M.-E., Pogorelcnik,R., Court,F., Arnaud,P. and Vaurs-Barrière,C. (2018) CLIFinder: identification of LINE-1 chimeric transcripts in RNA-seq data. *Bioinformatics*, **34**, 688–690.
36. Babaian,A., Thompson,I.R., Lever,J. and Gagnier,L. (2019) LIONS: analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinformatics*, **35**, 3839–3841.
37. Kircher,M., Sawyer,S. and Meyer,M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, **40**, e3.
38. Evrony,G.D., Lee,E., Park,P.J. and Walsh,C.A. (2016) Resolving rates of mutation in the brain using single-neuron genomics. *eLife*, **5**, e12966.
39. Quail,M.A., Kozarewa,I., Smith,F., Scally,A., Stephens,P.J., Durbin,R., Swerdlow,H. and Turner,D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, **5**, 1005–1010.
40. Treiber,C.D. and Waddell,S. (2017) Resolving the prevalence of somatic transposition in *Drosophila*. *eLife*, **6**, e28297.
41. Martin Cerezo,M.L., Raval,R., de Haro Reyes,B., Kucka,M., Chan,F.Y. and Bryk,J. (2022) Identification and quantification of chimeric sequencing reads in a highly multiplexed RAD-seq protocol. *Mol. Ecol. Resour.*, **22**, 2860–2870.
42. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
43. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and and 1000 Genome Project Data Processing Subgroupand 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.
44. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
45. Dale,R.K., Pedersen,B.S. and Quinlan,A.R. (2011) Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, **27**, 3423–3424.
46. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
47. Roberts,A. and Pachter,L. (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*, **10**, 71–73.
48. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q. *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.
49. SMIT, Arian FA (2004) RepeatMasker Open 3.0.

50. Storer,J., Hubley,R., Rosen,J., Wheeler,T.J. and Smit,A.F. (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, **12**, 2.
51. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
52. Fablet,M., Salcez-Ortiz,J., Jacquet,A., Menezes,B.F., Dechaud,C., Veber,P., Noûs,C., Rebollo,R. and Vieira,C. (2022) A quantitative, genome-wide analysis in Drosophila reveals transposable elements' influence on gene expression is species-specific. bioRxiv doi: https://doi.org/10.1101/2022.01.20.477049, 22 January 2022, preprint: not peer reviewed.
53. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
54. Mohamed,M., Dang,N.T.-M., Ogyama,Y., Burlet,N., Mugat,B., Boulesteix,M., Mérel,V., Veber,P., Salces-Ortiz,J., Severac,D. *et al.* (2020) A transposon story: from TE content to TE dynamic invasion of *drosophila* genomes using the single-molecule sequencing technology from Oxford Nanopore. *Cells*, **9**, 1776.
55. Shumate,A. and Salzberg,S.L. (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.
56. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., Capy,P., Chalhoub,B., Flavell,A., Leroy,P., Morgante,M., Panaud,O. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
57. Lerat,E., Rizzon,C. and Biémont,C. (2003) Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Res.*, **13**, 1889–1896.
58. Flutre,T., Duprat,E., Feuillet,C. and Quesneville,H. (2011) Considering transposable element diversification in de novo annotation approaches. *PLoS One*, **6**, e16526.
59. Flutre,T., Permal,E. and Quesneville,H. (2012) Transposable element annotation in completely sequenced eukaryote genomes. In: Grandbastien,M-.A. and Casacuberta,J.M. (eds.) *Plant Transposable Elements, Topics in Current Genetics*. Springer, Berlin Heidelberg, Vol. **24**, pp. 17–39.
60. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
61. dos Santos,G., Schroeder,A.J., Goodman,J.L., Strelets,V.B., Crosby,M.A., Thurmond,J., Emmert,D.B., Gelbart,W.M. and the FlyBase Consortiumthe FlyBase Consortium (2015) FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.*, **43**, D690–D697.
62. Lee,C.M., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Gonzalez,J.N., Hinrichs,A.S., Lee,B.T., Nassar,L.R., Powell,C.C. *et al.* (2019) UCSC Genome Browser enters 20th year. *Nucleic Acids Res.*, **48**, D756–D761.
63. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf.*, **14**, 178–192.
64. Rice,P., Longden,I. and Bleasby,A.EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
65. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
66. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L. *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
67. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
68. Zea,D.J., Anfossi,D., Nielsen,M. and Marino-Buslje,C. (2017) MIToS.Jl: mutual information tools for protein sequence analysis in the Julia language. *Bioinformatics*, **33**, 564–565.
69. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
70. Lu,Z., Marand,A.P., Ricci,W.A., Ethridge,C.L., Zhang,X. and Schmitz,R.J. (2019) The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants*, **5**, 1250–1259.
71. Bakoulis,S., Krautz,R., Alcaraz,N., Salvatore,M. and Andersson,R. (2022) Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. *Nucleic Acids Res.*, **50**, 2111–2127.
72. Annala,M.J., Parker,B.C., Zhang,W. and Nykter,M. (2013) Fusion genes and their discovery using high throughput sequencing. *Cancer Lett.*, **340**, 192–200.
73. Kanagawa,T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.*, **96**, 317–323.
74. Houseley,J. and Tollervey,D. (2010) Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, **5**, e12271.
75. Van Der Valk,T., Vezzi,F., Ormestad,M., Dalén,L. and Guschanski,K. (2020) Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.*, **20**, 1171–1181.
76. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
77. Zhou,Y., Zhang,C., Zhang,L., Ye,Q., Liu,N., Wang,M., Long,G., Fan,W., Long,M. and Wing,R.A. (2022) Gene fusion as an important mechanism to generate new genes in the genus Oryza. *Genome Biol.*, **23**, 130.
78. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Svirskas,R., Patel,S., Frise,E., Lewis,S.E., Rubin,G.M., Ashburner,M. and Celniker,S.E. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.*, **3**, RESEARCH0084.
79. Gatti,M. and Pimpinelli,S. (1992) Functional elements in *Drosophila melanogaster* heterochromatin. *Annu. Rev. Genet.*, **26**, 239–276.
80. Carmena,M. and Gonzfilez,C. (1995) Transposable elements map in a conserved pattern of distribution extending from beta-heterochromatin to centromeres in *Drosophila melanogaster*. *Chromosoma*, **103**, 676–684.
81. Coronado-Zamora,M. and González,J. (2022) Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across *Drosophila* natural strains. bioRxiv doi: https://doi.org/10.1101/2022.12.02.518890, 04 December 2022, preprint: not peer reviewed.
82. Bailly-Bechet,M., Haudry,A. and Lerat,E. (2014) "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, **5**, 13.
83. Kapun,M., Barrón,M.G., Staubach,F., Obbard,D.J., Wiberg,R.A.W., Vieira,J., Goubert,C., Rota-Stabelli,O., Kankare,M., Bogaerts-Márquez,M. *et al.* (2020) Genomic analysis of european *Drosophila melanogaster* populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Mol. Biol. Evol.*, **37**, 2661–2678.
84. Rech,G.E., Bogaerts-Márquez,M., Barrón,M.G., Merenciano,M., Villanueva-Cañas,J.L., Horváth,V., Fiston-Lavier,A.-S., Luyten,I., Venkataram,S., Quesneville,H. *et al.* (2019) Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *PLoS Genet.*, **15**, e1007900.
85. Rech,G.E., Radío,S., Guirao-Rico,S., Aguilera,L., Horvath,V., Green,L., Lindstadt,H., Jamilloux,V., Quesneville,H. and González,J. (2022) Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nat. Commun.*, **13**, 1948.
86. Lima,L., Sinaimeri,B., Sacomoto,G., Lopez-Maestre,H., Marchet,C., Miele,V., Sagot,M.F. and Lacroix,V. (2017) Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorith. Mol. Biol.*, **12**, 2.
87. Lahens,N.F., Kavakli,I.H., Zhang,R., Hayer,K., Black,M.B., Dueck,H., Pizarro,A., Kim,J., Irizarry,R., Thomas,R.S. *et al.* (2014) IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.*, **15**, 15.
88. Wang,S. and Gribskov,M. (2016) Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics*, **33**, 327–333.
89. Rebollo,R., Cumunel,E., Mary,A., Burlet,N., Gillet,B., Hughes,S., Oliveira,D.S., Goubert,C., Fablet,M., Vieira,C. *et al.* (2023) Detection and identification of transposable element transcripts using Long Read RNA-seq in Drosophila germline tissues. bioRxiv

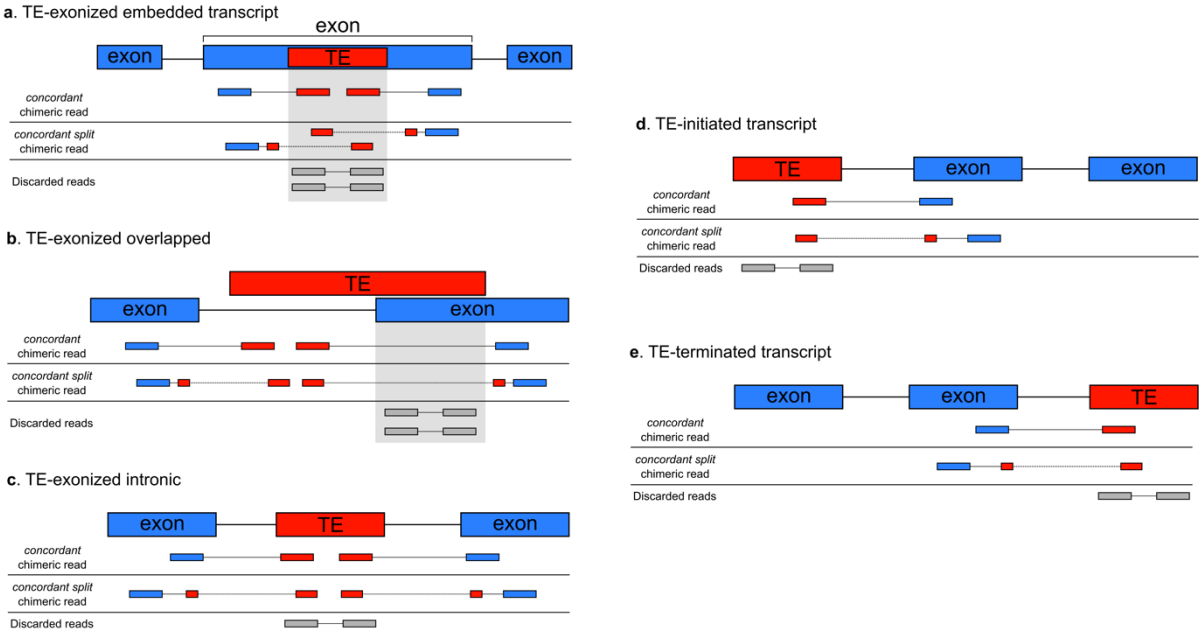
doi: https://doi.org/10.1101/2023.05.27.542554, 30 May 2023, preprint: not peer reviewed.
90. Ura,H., Togi,S. and Niida,Y. (2022) A comparison of mRNA sequencing (RNA-Seq) library preparation methods for transcriptome analysis. *Bmc Genomics [Electronic Resource]*, **23**, 303.
91. Cridland,J.M., Macdonald,S.J., Long,A.D. and Thornton,K.R. (2013) Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Mol. Biol. Evol.*, **30**, 2311–2327.
92. Lerman,D.N. and Feder,M.E. (2005) Naturally occurring transposable elements disrupt hsp70 promoter function in *Drosophila melanogaster*. *Mol. Biol. Evol.*, **22**, 776–783.
93. Nefedova,L.N., Kuzmin,I.V., Makhnovskii,P.A. and Kim,A.I. (2014) Domesticated retroviral GAG gene in *Drosophila*: new functions for an old gene. *Virology*, **450–451**, 196–204.
94. Malik,H.S. and Henikoff,S. (2005) Positive selection of Iris, a retroviral envelope–Derived host gene in *Drosophila melanogaster*. *PLoS Genet.*, **1**, e44.
95. Rahman,R., Chirn,G., Kanodia,A., Sytnikova,Y.A., Bergman,C.M. and Lau,N.C. (2015) Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res.*, **43**, 10655–10672.
96. Vieira,C. and Biemont,C. (2004) Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *drosophila simulans*. *Genetica*, **120**, 115–123.
97. Merenciano,M., Ullastres,A., de Cara,M.A.R., Barrón,M.G. and González,J. (2016) Multiple independent retroelement insertions in the promoter of a stress response gene have variable molecular and functional effects in *Drosophila*. *PLoS Genet.*, **12**, e1006249.
98. Rebollo,R., Horard,B., Begeot,F., Delattre,M., Gilson,E. and Vieira,C. (2012) A snapshot of histone modifications within transposable elements in *Drosophila* wild type strains. *PLoS One*, **7**, e44253.
99. Yasuhara,J.C. and Wakimoto,B.T. (2008) Molecular landscape of modified histones in *Drosophila* heterochromatic genes and euchromatin-heterochromatin transition zones. *PLoS Genet.*, **4**, e16.
100. Díaz-González,J., Vázquez,J.F., Albornoz,J. and Domínguez,A. (2011) Long-term evolution of the *roo* transposable element copy number in mutation accumulation lines of *Drosophila melanogaster*. *Genetics Res.*, **93**, 181–187.
101. Díaz-González,J., Domínguez,A. and Albornoz,J. (2010) Genomic distribution of retrotransposons 297, 1731, copia, mdg1 and roo in the *Drosophila melanogaster* species subgroup. *Genetica*, **138**, 579–586.
102. Díaz-González,J. and Domínguez,A. (2020) Different structural variants of roo retrotransposon are active in *Drosophila melanogaster*. *Gene*, **741**, 144546.
103. Vieira,C., Lepetit,D., Dumont,S. and Biemont,C. (1999) Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol. Biol. Evol.*, **16**, 1251–1255.
104. Barrón,M.G., Fiston-Lavier,A.-S., Petrov,D.A. and González,J. (2014) Population genomics of transposable elements in *Drosophila*. *Annu. Rev. Genet.*, **48**, 561–581.
105. Kapitonov,V.V. and Jurka,J. (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 6569–6574.
106. Kofler,R., Betancourt,A.J. and Schlötterer,C. (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet.*, **8**, e1002487.
107. Maddison,W.P. and Knowles,L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.*, **55**, 21–30.
108. Waters,L.C. (1992) Possible involvement of the long terminal repeat of transposable element 17.6 in regulating expression of an insecticide resistance-associated P450 gene in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 4855–4859.
109. Delpuech,J.M., Aquadro,C.F. and Roush,R.T. (1993) Noninvolvement of the long terminal repeat of transposable element 17.6 in insecticide resistance in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 5643–5647.
110. Brun,A., Cuany,A., Le Mouel,T., Berge,J. and Amichot,M. (1996) Inducibility of the *Drosophila melanogaster* cytochrome P450 gene, CYP6A2, by phenobarbital in insecticide susceptible or resistant strains. *Insect Biochem. Mol. Biol.*, **26**, 697–703.
111. Seong,K.M., Coates,B.S., Berenbaum,M.R., Clark,J.M. and Pittendrigh,B.R. (2018) Comparative CYP-omic analysis between the DDT-susceptible and -resistant *Drosophila melanogaster* strains *91-C* and *91-R*. *Pest Manage. Sci.*, **74**, 2530–2543.
112. Carareto,C.M.A., Hernandez,E.H. and Vieira,C. (2014) Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling *Drosophila* species. *Gene*, **537**, 93–99.
113. Hill,M.S., Vande Zande,P. and Wittkopp,P.J. (2021) Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.*, **22**, 203–215.
114. Marguerat,S. and Bähler,J. (2010) RNA-seq: from technology to biology. *Cell. Mol. Life Sci.*, **67**, 569–579.
115. Oliveira,D.S., Rosa,M.T., Vieira,C. and Loreto,E.L.S. (2021) Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. *J. Evol. Biol.*, **34**, 628–638.
116. Horváth,V., Merenciano,M. and González,J. (2017) Revisiting the relationship between transposable elements and the eukaryotic stress response. *Trends Genet.*, **33**, 832–841.
117. Nicolau,M., Picault,N. and Moissiard,G. (2021) The evolutionary volte-face of transposable elements: from harmful jumping genes to major drivers of genetic innovation. *Cells*, **10**, 2952.
118. Hug,N., Longman,D. and Cáceres,J.F. (2016) Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.*, **44**, 1483–1495.
119. Harigaya,Y. and Parker,R. (2010) No-go decay: a quality control mechanism for RNA in translation. *WIREs RNA*, **1**, 132–141.
120. Vasudevan,S., Peltz,S.W. and Wilusz,C.J. (2002) Non-stop decay—a new mRNA surveillance pathway. *Bioessays*, **24**, 785–788.

Supplementary Material

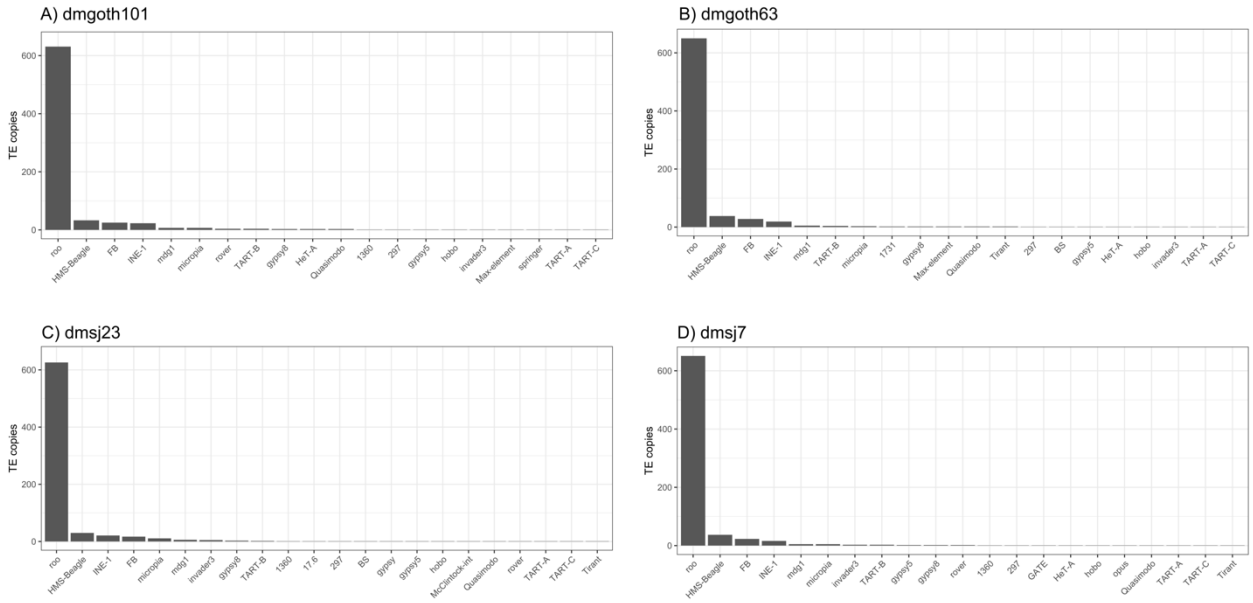
Supplementary Tables

The supplementary tables S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12 and S13 can be accessed in the publicly available link from Nucleic Acids Research journal: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkad671#supplementary-data

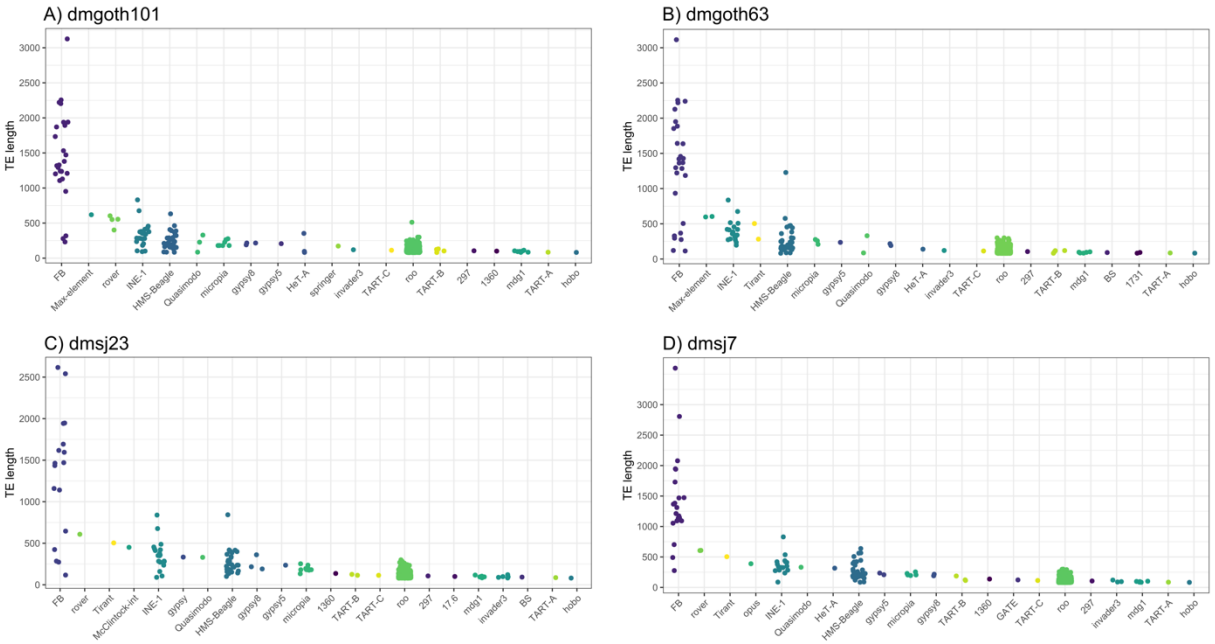
Supplementary Figures



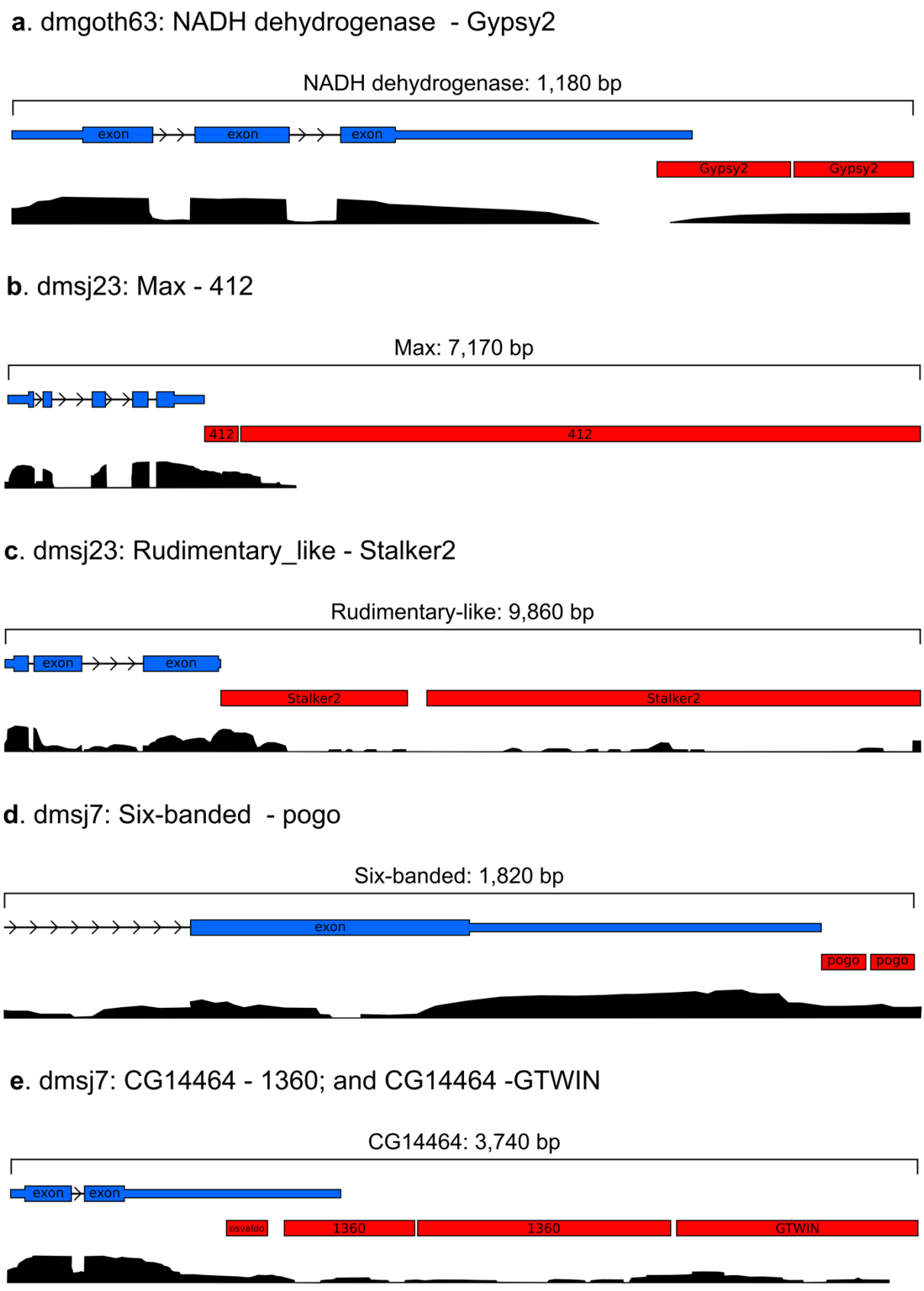
Supplementary Figure 1: ChimeraTE Mode 1 detection of chimeric reads (CR). **a**) TE-exonized embedded transcript: The exon is artificially divided into two portions, based on the TE's start and end position. Reads with both mates aligned solely into the TE region (gray block) are discarded. Concordant *end-to-end* reads are considered as CRs if one mate is aligned into the TE, and the other is aligned into the exon; while *split* reads are counted when a partial region of one mate is aligned into the TE (or exon) and the other is aligned into the exon (or TE). Such approach avoids counting autonomous TE expression as evidence for chimeric transcripts. **b**) Similarly, TE-exonized overlapped does not consider reads aligned solely to the TE as CR. Instead of it, both concordant and split reads are considered as CRs when both mates have concomitantly aligned into exons and TEs. **c**) CRs supporting TE-exonized intronic are computed when both mates are aligned between the TE and an exon. **d**) CRs for TE-initiated transcripts must have one mate, concordant or split, aligned into the TE located 3kb upstream to the gene, whereas the other is aligned into the exon. **e**) CRs for TE-terminated transcripts have the same rationale as TE-initiated, but with the TE insertion located 3kb downstream to the gene. Any exon can have CRs, even though they are not represented in the figure.



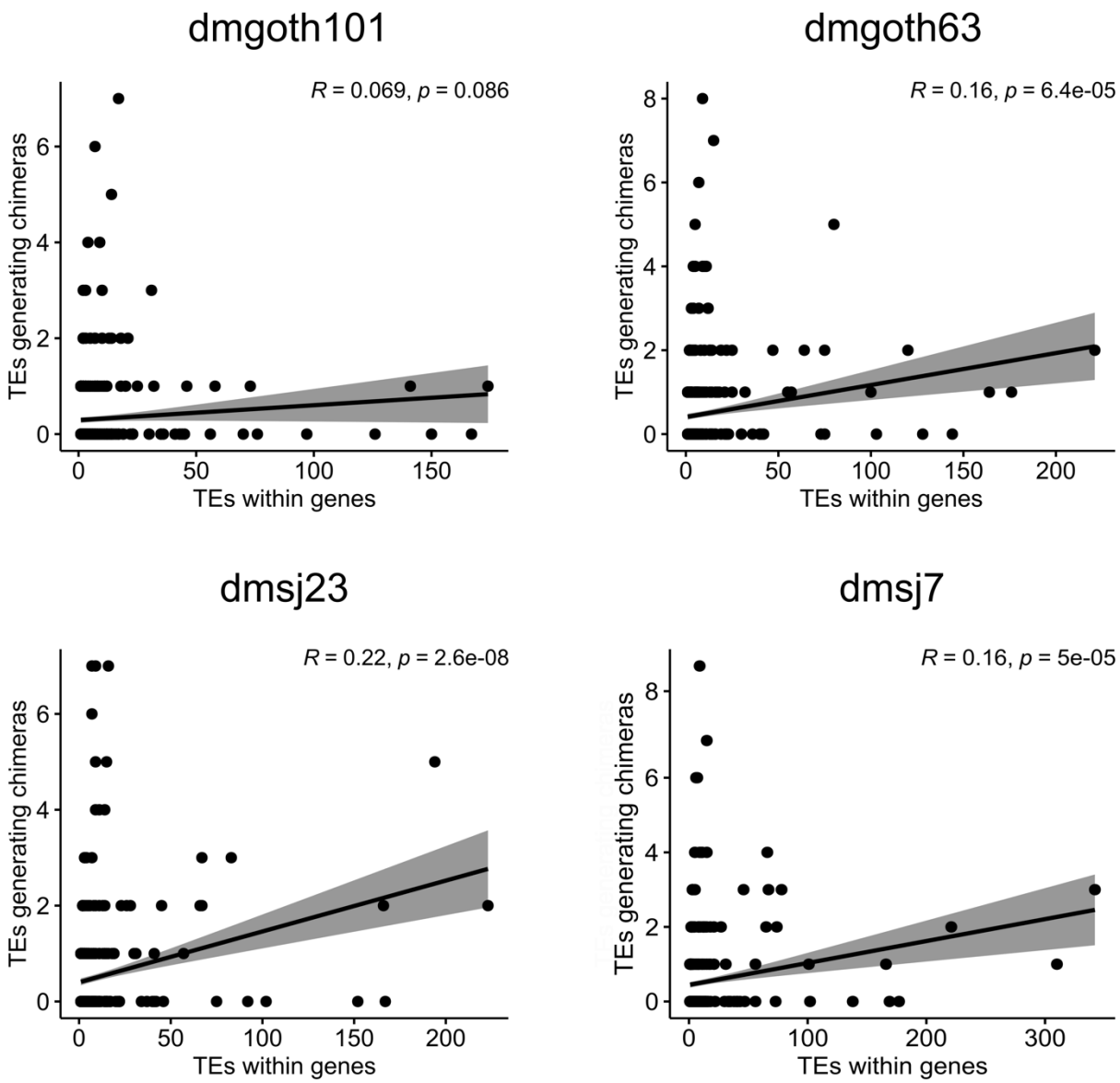
Supplementary Figure 2: Frequency of TE copies by TE family that were filtered out after removal of insertions with SSRs covering more than 50% of their length. *Roo* family comprise more than 85% in the four wild-type strains.



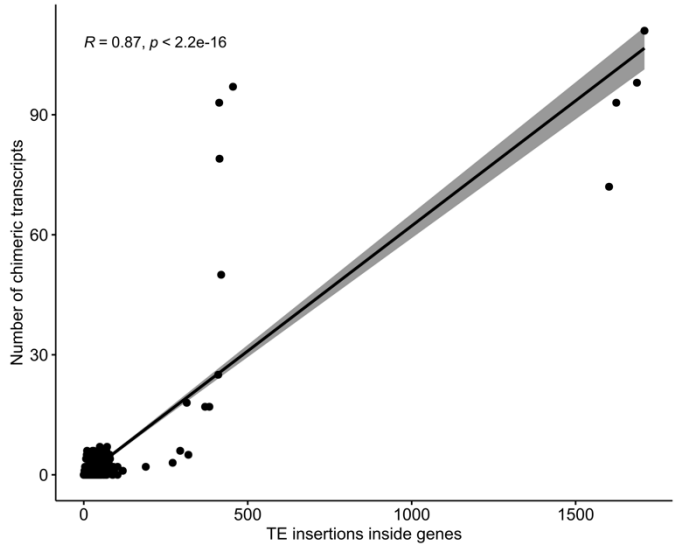
Supplementary Figure 3: Average length of TE insertions removed after applying filtering of SSRs > 50% of TE insertion length.



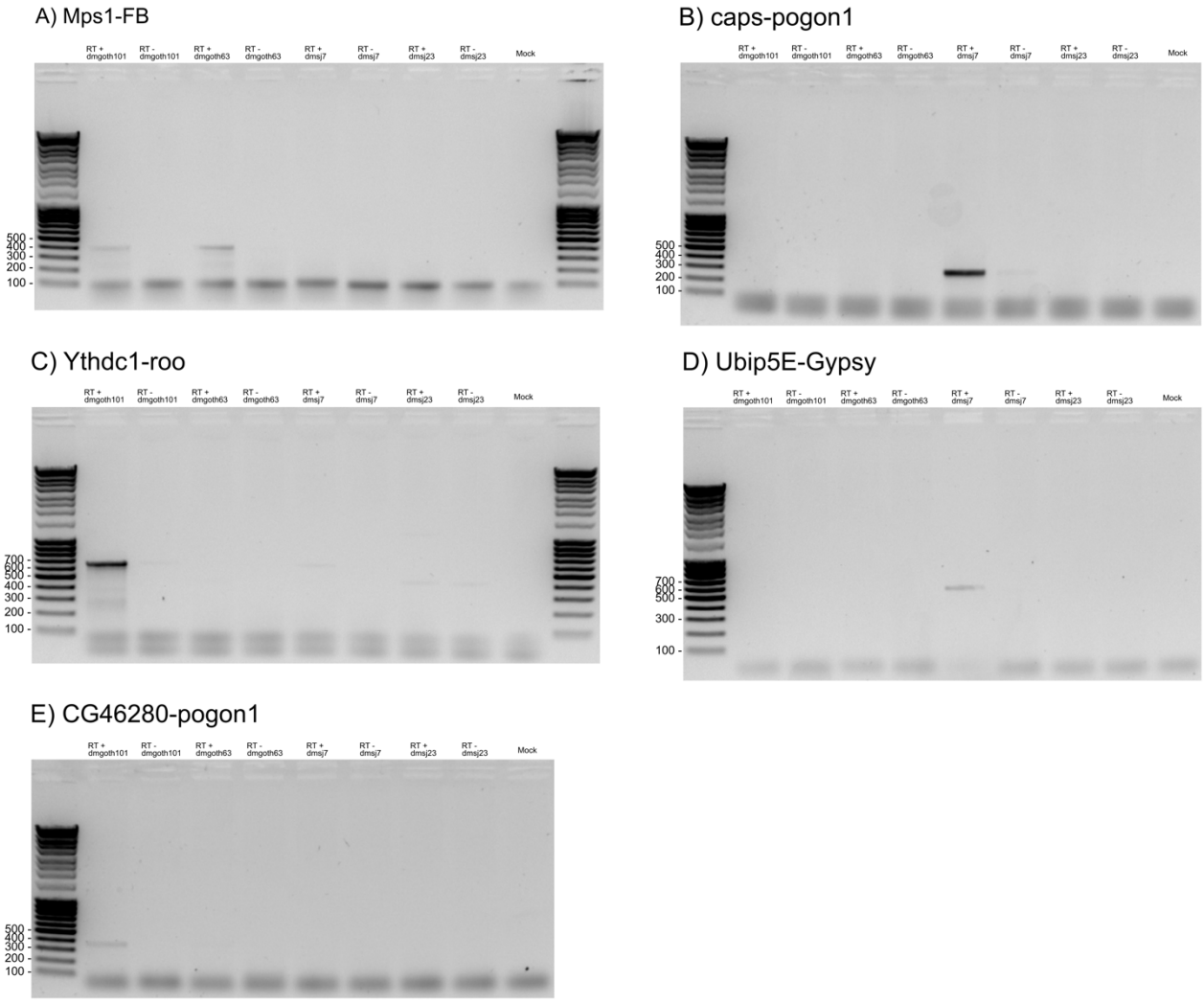
Supplementary Figure 4: TE-terminated transcripts with multiple TE insertions.



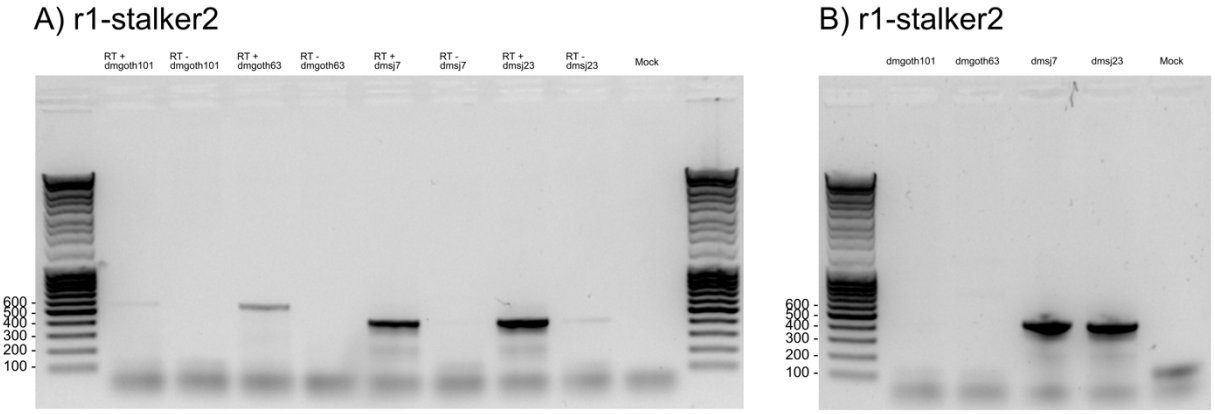
Supplementary Figure 5: Spearman correlation between the number of TEs within genes with FPKM > 1. Y axis represents the number of TE insertions identified as chimeric transcripts, whereas X axis represents the number of TE insertion within gene region.



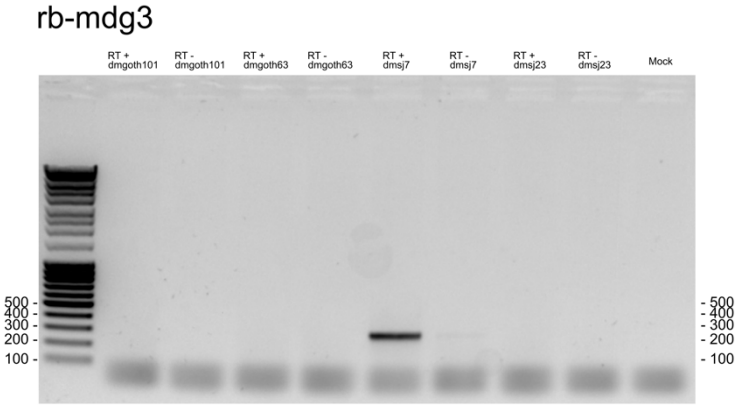
Supplementary Figure 6: Spearman correlation between the number of TE insertion inside genes, per TE family, and its respective number of chimeric transcripts.



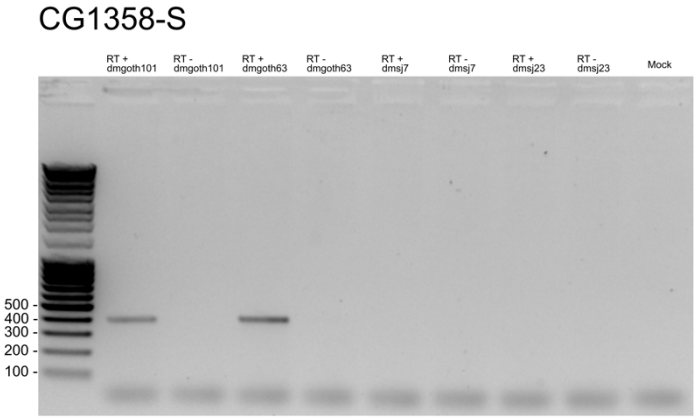
Supplementary Figure 7: Five chimeric transcripts found by ChimeraTE Mode 2 in our four wild-type strains derived from TE insertions absent in the reference genome. All RT-PCR amplifications represent the same result obtained by ChimeraTE Mode 2.



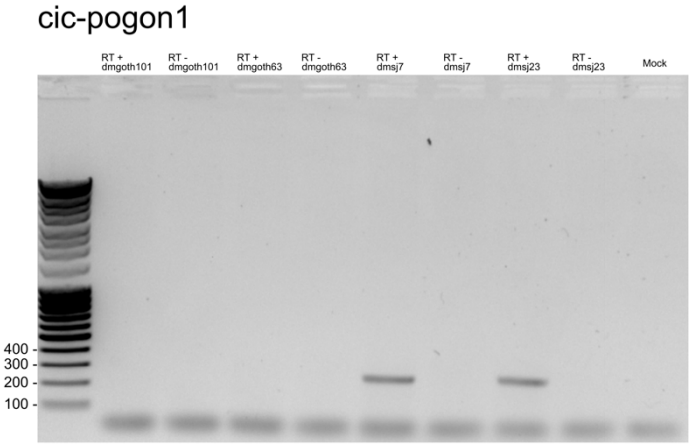
Supplementary figure 8: **A**) Presence of chimeric transcript derived from *r1* gene and *Stalker2* TE in the four wild-type strains. The RT-PCR reveals amplification in the expected length in both dmsj7 and dmsj23, whereas for dmgoth101 and dmgoth63 there are off target amplification. **B**) PCR result for the Stalker2 insertion, demonstrating the presence of such insertion in both Brazilian strains.



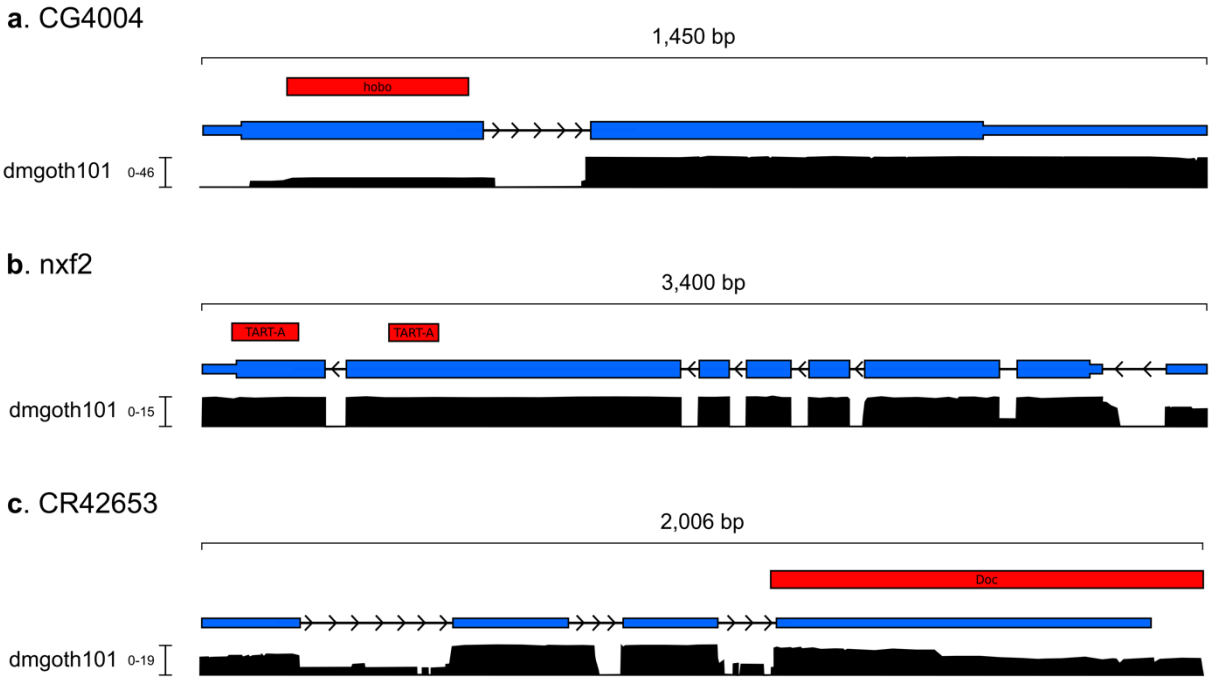
Supplementary figure 9: Presence of chimeric transcript derived from *rb* gene and *mdg3* TE in the four wild-type strains.



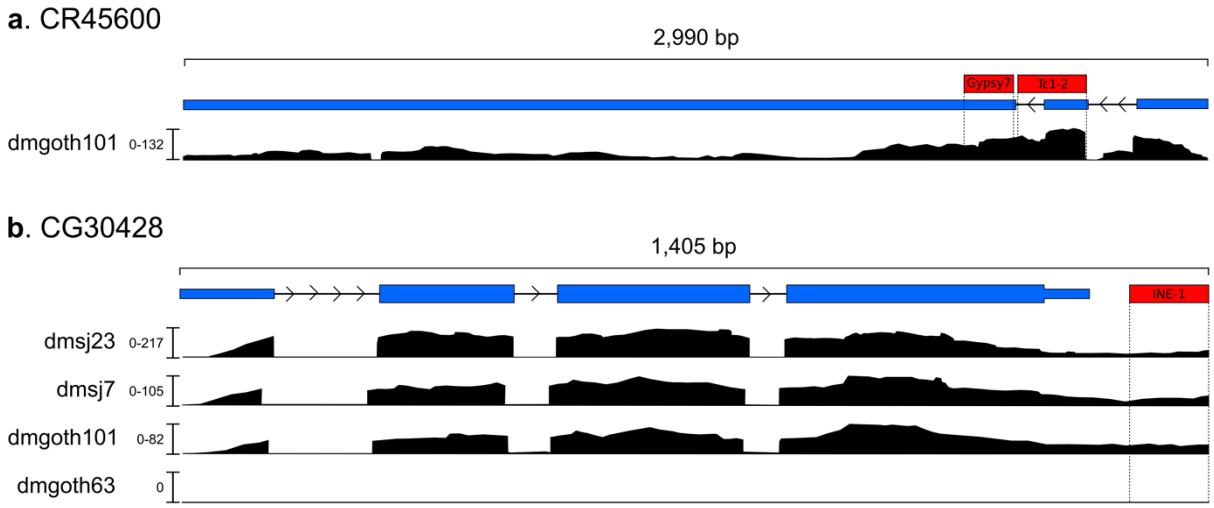
Supplementary figure 10: Presence of chimeric transcript derived from *CG1358* gene and *S* TE in the four wild-type strains.



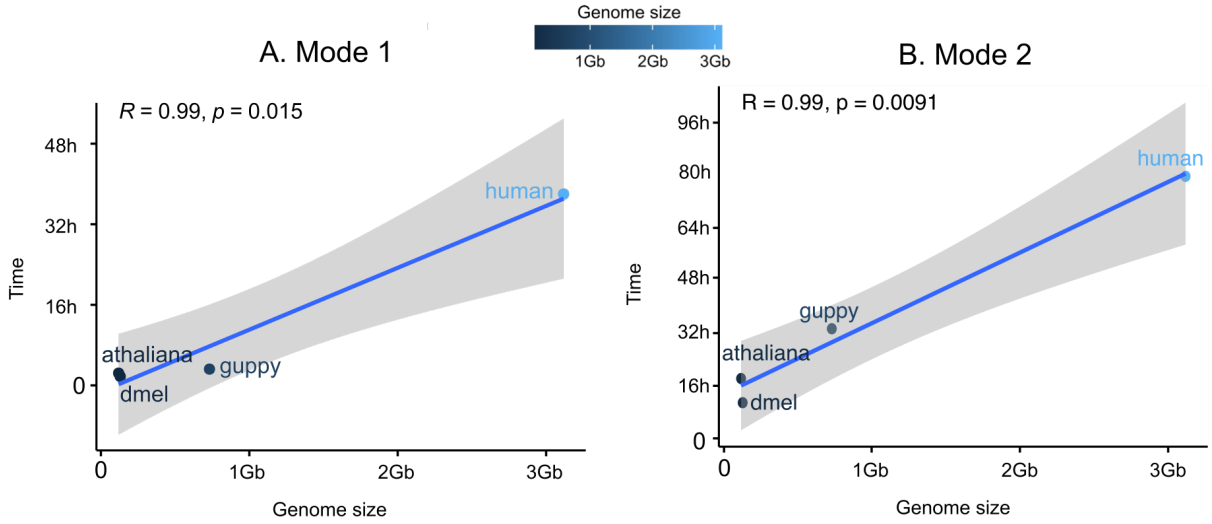
Supplementary figure 11: Presence of chimeric transcript derived from *cic* gene and *pogon1* TE in the four wild-type strains.



Supplementary Figure 12: Chimeric transcripts with TE-derived protein domain found with RNA-seq ONT data.



Supplementary Figure 13: Relative contribution of TE expression to the gene expression, for genes that are inactive in one strain, but generating chimeric transcript in the other. The coverage on the TE sequence region is delimited by dashed lines.



Supplementary Figure 14: Correlation between processing-time of ChimeraTE and genome size of species. Both Mode 1 (**a**) and Mode 2 (**b**) had a strong positive correlation with the genome sizes.

Transposable elements as evolutionary driving force to host shift in cactophilic *Drosophila* species

Oliveira D. S.[1,2], Larue A.[2], Nunes W. V. B.[2], Sabot F.[3], Bodelón A.4, García Guerreiro M. P.4, Vieira C.[2,*], Carareto C. M. A.[1,*]

1 São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, São Paulo, 15054-000, Brazil.

2 Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS, UMR5558, Villeurbanne 69100, France.

3 DIADE, University of Montpellier, Cirad, IRD, 911 avenue Agropolis BP64501, 34394 Montpellier, France.

4 Grup de Genòmica, Bioinformática i Biologia Evolutiva, Departament de Genètica i Microbiologia (Edifici C), Universitat Autònoma de Barcelona, Barcelona, Spain

* Corresponding authors

Daniel Siqueira de Oliveira: daniel.siqueira-de-oliveira@etu.univ-lyon1.fr

Anaïs Larue: anais.larue@univ-lyon1.fr

William Vilas Boas Nunes: william.vilas-boas-nunes@univ-lyon1.fr

François Sabot: francois.sabot@ird.fr

Alejandra Bodelón: Alejandra.Bodelon@uab.cat

María Pilar García Guerreiro: MariaPilar.Garcia.Guerreiro@uab.cat

Cristina Vieira: cristina.vieira@univ.lyon1.fr

Claudia Marcia Aparecida Carareto: claudia.carareto@unesp.br

ABSTRACT

Background

The host shift in insects has been considered a key process with potential to collaborate with reproductive isolation and speciation. Both genomics and transcriptomics variation has been attributed to such a process, in which gene families with functions for host location, acceptance and usage have been proposed to evolve. In this context, cactophilic *Drosophila* species are an excellent model to study host shift evolution, since they use a wide-range of cacti as hosts, and many species have different preferences. Transposable elements are engines of genetic novelty between populations and species, driving rapid adaptive evolution. However, the extent of TEs' contribution to host shift remains unexplored.

Results

We performed genomic and transcriptomic analysis in seven genomes of cactophilic species/subspecies to investigate how TEs interact with genes associated with host shift. Our results revealed enrichment of TEs at promoter regions of host shift-related genes, with 35% of the odorant receptors containing their transcription factor binding sites within TEs. We observed that 60% of these TEs are Helitrons, demonstrating an unprecedented putative cis-regulatory role of *Helitrons* in *Drosophila*. Differential expression analysis between species with different preferred hosts demonstrated divergence on gene expression in head and larvae tissues. Although TEs' presence does not affect overall gene expression, we observed 1.31% of genes generating gene-TE chimeric transcripts, including those with function affecting host preference.

Conclusions

Our combined genomic and transcriptomic approaches provide evidence of TE-driven divergence between species, highlighting the evolutionary role of TEs in the context of host shift, a key adaptive process that can cause reproductive isolation.

Keywords: comparative transcriptomics, adaptation, chimeric transcripts.

Introduction

The molecular consequences of host shift evolution can be closely related to ecological divergence, specialization and ultimately ecological speciation (1). Due to the wide variety of hosts, successful plant-insect interactions require adaptations from the insect. In many species, reproductive isolation arises as a result of successive adaptations driven by host shift (HS), which results from divergent selection between different ecological environments (1–4). Thus, natural selection increases the frequency of alleles that confer fitness benefits when HS occurs. *Drosophila* species use a wide range of necrotic host tissues as feeding and breeding sites, providing a suitable model to study the HS process and its role in insect speciation. Among them, the *repleta* group comprises a clade with ~100 cactophilic *Drosophila* species (5). The specialization in cacti represents an important ecological challenge to the flies, due to the lack of nitrogen and phosphorus compared to other hosts, as well as the presence of toxic compounds (6). In the *repleta* species, the HS was initially to the *Opuntia* sp. cacti, and secondarily several independent shifts happened to columnar cacti (6), considered a host with higher toxicity than *Opuntia* sp. (7). Such HS is observed between sibling species with recent divergence in two clusters of the subgroup mulleri: cluster *buzzatii* and cluster *mojavensis*. In the cluster *buzzatii*, *Drosophila buzzatii* and *D. koepferae* are sibling species (divergence 4-5 Mya (8)) with different primary hosts. The former species use preferentially *Opuntia* sp. cacti, whereas the latter is a columnar cacti dweller (9). In the cluster *mojavensis*, *D. arizonae* and the four allopatric subspecies of *D. mojavensis* have a high diversity regarding cacti hosts' preference. *D. arizonae* use both *Opuntia* sp. and columnar cacti, whereas *D. m. mojavensis* uses *Ferocactus cylindraceus* as preferential host, *D. m. wrigleyi* uses only *Opuntia* sp., *D. m. baja* uses *Stenocereus gummosus*, and *D. m. sonorensis* uses *S. thurberi* (10). Therefore, this set of species from *buzzatii* and *mojavensis* clusters provide an excellent model to investigate the molecular mechanisms underlying HS evolution.

The radiation of cactophilic *Drosophila* species across different hosts is an evolutionary outcome that gave rise to specific adaptations (11), which may be summarized into three steps: localization, acceptance and host usage (12). In the location step, odorant-binding proteins (OBPs) and odorant receptors (ORs) are key proteins to distinguish specific host in the environment, through chemoreceptors that integrate the visual and olfactory systems. Subsequently, in the acceptance step occurs the evaluation of the nutritional compounds present on the host, as well as the presence of competitors, predators, pathogens and parasites (12). Several receptors are associated with this process, such as gustatory receptors (GRs) (13) and

ionotropic receptors (IRs). (14–16). Finally, the flies must be able to complete the feeding and breeding processes, but it depends upon the overcoming of toxic substances that might be present on the host. Several gene families are associated with detoxification, such as P450s (CYPs), Glutathione S-Transferases (GSTs), UDP-Glicosiltransferases (UGTs), esterases (ESTs) and ATP binding-cassette transporters (ABCs). Altogether, these nine gene families associated with the Host Location, Acceptance and Usage will henceforth be referred to as HLAU.

A previous study with 14 *Drosophila* genomes, representing both *Sophophora* and *Drosophila* subgenera, has proposed that 82 genes, including GSTs, CYPs, and UDPs, are under positive selection (11). Signatures of positive selection were also found between *D. mojavensis* subspecies and *D. arizonae* in subsets of GRs and ORs (16). In addition to the selective pressure acting into HLAU gene families, modifications in the gene expression also play an important role in the HS adaptation in *Drosophila* species. It has been shown that CYPs, GSTs and OBPs are differentially expressed between *D. m. wrigleyi* and *D. m. sonorensis* (17), indicating that gene expression in the nervous system might play a role into detoxification and olfactory systems, hence contributing to HS adaptation in *D. mojavensis*.

Transposable elements (TEs) are also likely to play a role in environmental adaptation because of their ability to generate mutations. In most cases, mutations caused by TEs are likely to be deleterious or neutral. Throughout evolutionary time, TEs that remain in the genome tend to be silenced by epigenetic control, accumulating mutations, and losing their transposition ability (18). Despite the fact that they become degraded, the remaining TE sequences may still contain either regulatory motifs or protein domains (19). These sequences might be co-opted by the cell machinery modifying gene expression or protein sequences of the nearby genes (reviewed in (20)). Depending upon the positive outcome of the cooption into the individual fitness, these TEs can increase their frequency in the population, contributing to populational evolution and hence determined as adaptive insertions. For instance, in D. melanogaster, the gene CHKov1 has a truncated mRNA derived from a DOC TE located into the intron, resulting in resistance to viral infection and insecticides (21). Such exaptation and domestication events can be often identified by the occurrence of chimeric transcripts, which are mRNAs with both gene and TE-derived sequences (22). A recent transcriptome-wide study has identified 327 genes of *D. melanogaster* genes produce chimeric transcripts in different populations (23). Among all genes, 76 generate chimeric transcripts from TE insertions that were present in one

strain, but absent in another, highlighting the potential of TEs as a source of genetic novelty between different *Drosophila* ecotypes.

Many aspects involving HS adaptation in *repleta* species have been shown previously, such as detoxification pathways (24), morphology (25), life history traits (26), behavior (27), and genomic and transcriptomic differences (17,28–31). However, the contribution of TEs to the evolution of HLAU genes has not yet been assessed. Here we aimed to uncover the extension of variability derived from TEs in cactophilic *Drosophila* species, using both genome- and transcriptome-wide analysis. We tested the hypothesis regarding the contribution of TEs to the HS in cactophilic species. The genome-wide analysis revealed that HLAU gene families are enriched with TEs in regulatory regions, particularly in ORs. Furthermore, the results show that approximately 35% of ORs with TEs in their promoter regions contain transcription factor binding sites (TFBSs) within the TE sequence, of which around 60% are Helitrons, highlighting their putative important regulatory role in *Drosophila*. The transcriptomic analysis revealed that HLAU genes are differentially expressed between cactophilic species. Additionally, we also show that 1.31% of genes produce chimeric transcripts in head and larvae tissues, with HLAU genes are also being involved.

Results

Genome assemblies and gene annotation

To investigate the potential role of TEs into HS of cactophilic *Drosophila* species, we performed Nanopore long-reads sequencing on species that have different preferential cacti as hosts: *D. buzzatii* (*Opuntia* sp.), *D. koepferae* (columnar cacti), *D. arizonae* (*Opuntia* sp. or columnar cacti), *D. m. mojavensis* (*F. cylindraceus*), *D. m. wrigleyi* (*Opuntia* sp.), *D. m. sonorensis* (*S. thurberi*), and *D. m. baja* (*S. gummosus*). We obtained high resolution assemblies for all genomes, with an average of 567 scaffolds, 13.8 Mb of N50 (Sup. Table 1), and ~98% of BUSCO genes (Sup. Fig. 1). Furthermore, the gene annotation for these genomes covered 98.63% of all reference genes in the *D. mojavensis* subspecies, and 95.98% in *D. arizonae*. The de novo gene annotation in *D. buzzatii* and *D. koepferae* revealed a total of 18,050 and 17,848 coding genes, for which 63.77%, and 64.13% were successfully assigned as one-to-one orthologs with *D. mojavensis*, respectively. These results represent the higher gene repertoire identified for *D. buzzatii* and *D. koepferae* species, altogether with its ortholog relationship with *D. mojavensis* and *D. arizonae*. Since analysis on HLAU genes is the focus of this study, their efficient annotation is relevant. They are represented by nine gene families associated with the

three steps on the use of host resources: (1) Location: OBPs, ORs; (2) Acceptance: GRs, IRs; and (3) Host usage: ABCs, ESTs, GSTs, UDPs, CYPs (Sup. Table 2). In *D. arizonae* and in the *D. mojavensis* subspecies, we recovered ~98.21% of the total HLAU genes from the reference genome. In *D. buzzatii* and *D. koepferae* the number of annotated HLAU genes covered ~95% in both species compared to *D. mojavensis* (Table 1).

Species/subspecies	Total genes	Host location, acceptance and usage genes								
		Location		*Acceptance*		*Host use*				
		OBP	OR	GR	IR	ABC	EST	GST	UDP	CYP
Ref. *D. m. wrigleyi*	**15,045**	**25**	**59**	**45**	**22**	**27**	**17**	**28**	**21**	**59**
D. m. mojavensis	14,810	25	59	44	22	27	17	27	20	59
D. m. baja	14,797	25	57	44	22	27	17	27	19	59
D. m. wrigleyi	14,948	25	59	45	22	27	17	27	20	59
D. m. sonorensis	14,808	25	57	44	22	27	17	27	20	58
D. arizonae	14,441	25	56	41	21	27	17	27	20	59
D. buzzatii	18,050	25	52	34	21	27	16	27	20	68
D. koepferae	17,848	25	45	34	24	27	13	25	23	62

Table 1. Total number of genes annotated in seven *Drosophila* cactophilic species. The "Ref. *D. m. wrigleyi*" represents the annotation in the reference genome used to perform the gene annotation for our long-read Nanopore genomes. HLAU genes are separated according to their functions: Location: odorant-binding proteins (OBP), odorant receptor (OR); Acceptance: gustatory receptor (GR), ionotropic receptor (IR); Host usage: ATP-binding cassette (ABC), carboxylesterase (EST), glutathione S-Transferase (GST), glycosiltransferase (UDP), and cytochrome P450 (CYP).

Molecular evolution of HLAU genes between species with preference for Opuntia sp. and columnar cacti

The use of *Opuntia* sp. has been proposed to be an ancestral state in cactophilic *Drosophila* species, whereas columnar cacti is the derived state (6). To investigate whether species with preferential columnar cacti usage have HLAU genes that undergone positive selection, we used a branch-site approach to assess their evolutionary rates compared to species that use *Opuntia* sp. For this analysis, we removed *D. arizonae* due to its non-preferential usage between columnar and *Opuntia* sp. cacti (6). Moreover, we added D. navojoa, which is a basal species of the *mojavensis* cluster that use solely *Opuntia* sp. (32). Therefore, we carried out the branch-site selection test with *D. buzzatii*, *D. m. wrigleyi*, and *D. navojoa* as *Opuntia*-related branches (background), and *D. m. baja*, *D. m. mojavensis*, *D. sonorensis*, and *D. koepferae* as columnar-related branches (foreground). It is important to highlight that due to the need of one-to-one

ortholog genes between the seven genomes, our analysis has decreased the total number of HLAU genes to ~42% of the initial number depicted on Table 1. From 128 genes, we did not observe any gene under positive selection with FDR < 0.05 (Sup. Table 3). To further explore the evolutionary forces driving gene HLAU genes evolution, we also performed a test for relaxed selection. We observed the genes Obp56d (LOC6580688) and Cyp304a1 (LOC6572829) with intensified selection strength (k > 1 and FDR < 0.05). Other six HLAU genes had significant relaxed selection, but with k < 1 (Sup. Table 4).

Evolutionary dynamics of TEs in cactophilic Drosophila species

We used a combination of automatic and manual methods to build a curated TE-library for each species. The method was designed to perform polishing steps in both TE consensus and copies to remove artefacts (see Methods). The total number of TE consensuses (LTRs separated of internal sequences) ranged from 243 in *D. m. baja* to 313 in *D. m. mojavensis*. The TE content in the genomes ranged from 14.07% in *D. buzzatii* to 20.4% in *D. m. mojavensis* (Fig. 1A). In addition, we observed a positive correlation between TE content and genome size estimated from the assemblies (Pearson correlation: p = 0.013; r = 0.86) (Sup. Fig. 2). Following the correlation, both species from the *buzzatii* cluster had the lower genome sizes and TE contents compared to species of the *mojavensis* cluster (Fig. 1A). Regarding the proportional load of TE orders in the genomes, LTRs contribute with 31.21% and 31.68% in the sister species *D. buzzatii* and *D. koepferae*, respectively. The main differences in the mobilomes are due to Helitrons and LINEs (Fig. 1B).

In the *mojavensis* cluster (*D. arizonae* and the four *D. mojavensis* subspecies), the abundance of TEs in the genome varied substantially (Fig. 2B), despite their recent divergence time of 1.5 Mya (33). *D. m. mojavensis* has the TE-richest genome (20.45%), with the abundance following from the highest to the lowest: LTR > LINE > Helitron > DNA > Unknown (Fig. 1B). The distribution of TEs in *D. arizonae* and the other three *D. mojavensis* does not follow such order. All of them have LTRs with the highest abundance, but followed by Helitrons, LINEs, DNA, and Unknown (Fig. 1B). The observed differences are often associated with the divergence on the evolutionary patterns of the TE classes between the genomes. Our analysis revealed that Helitrons had a recent burst on *D. m. mojavensis* and *D. m. wrigleyi*, followed by a decline (Fig. 1C). In *D. arizonae* the patterns reveal high genome content derived from TEs with low Kimura 2-parameters divergence, which indicates conserved and potentially active copies. The antagonistic patterns of TE dynamics observed in the seven

analyzed genomes suggest that TEs contribute differently to genome evolution across these species.

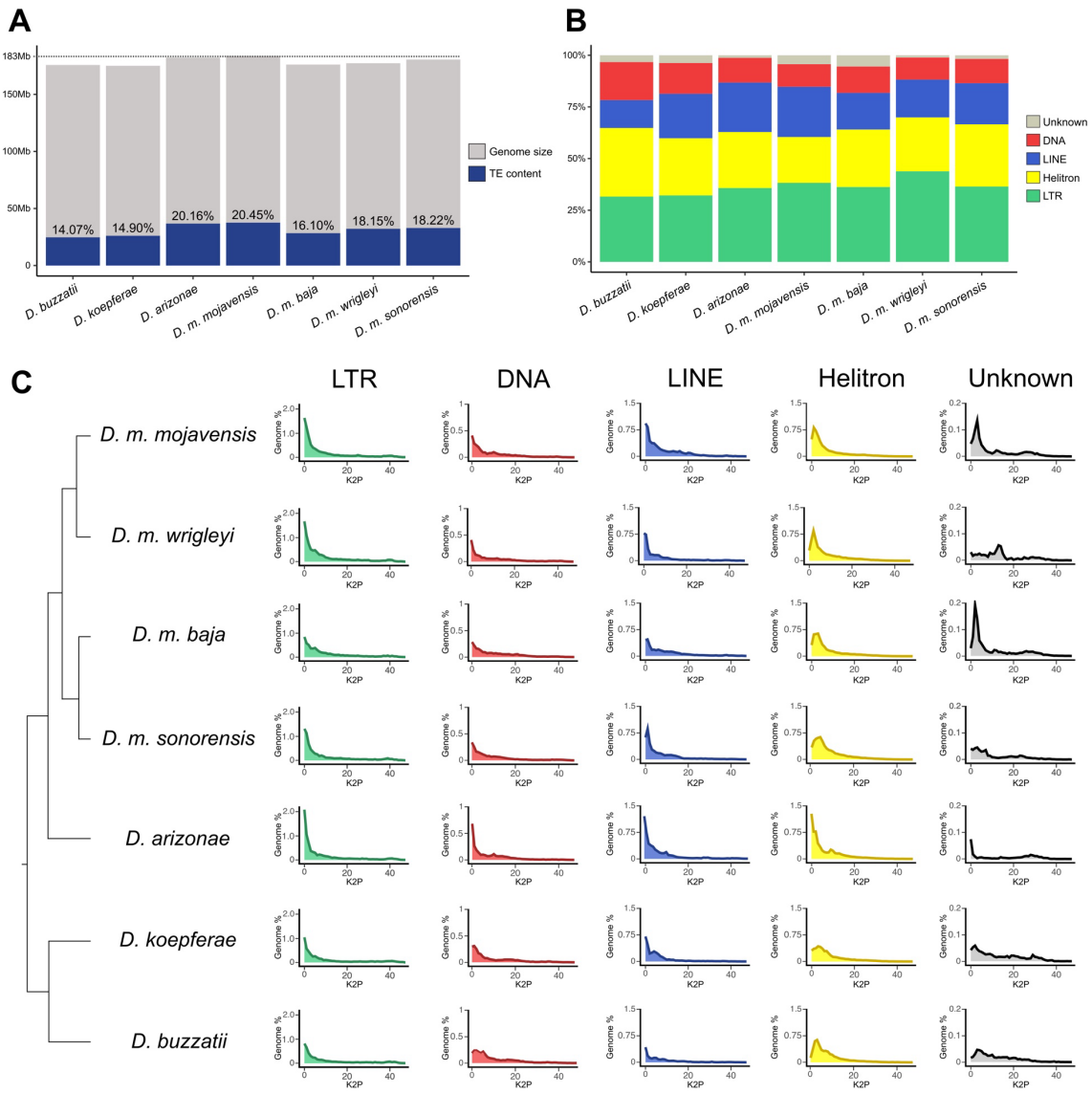


Figure 1. A) Genome size of the seven cactophilic *Drosophila* species sequenced in this work, and their respective TE content. B) Proportional distribution of the total TE content annotated in the genomes, divided by DNA transposon, LINE, *Helitron*, LTR, and Unknown. C) Phylogenetic relationship among species (adapted from (24,34)), alongside with TE families' landscapes representing TE content in genome % (y-axis), and the divergence of TE insertions compared to the consensuses (x-axis), based on Kimura 2-parameters (K2P) corrected with CpG. The TE content located on the left of the graph is represented by TE insertions with low divergence with their respective consensus, hence inferring that they might be conserved/recent/active, whereas TE content on the right part represents copies with high divergence due to the accumulation of mutations, corresponding to old/inactive copies.

HLAU genes are enriched in TEs at regulatory regions

To investigate whether TEs may have been maintained over time in a non-random distribution in the genomes relative to genes, we compared their positions and analyzed their frequencies in each genome. We analyzed whether specific TE orders are preferentially located within genes, which would suggest possible co-evolution. To do so, we compared the frequency of TE families that are preferentially inserted within gene regions (>50% of all insertions overlapping with genes), with a random distribution of TE insertions from the same order. We did not observe any TE order enriched within genes, which is expected with the hypothesis of purifying selection against TEs disrupting gene sequences (Sup. Fig. 3).

Despite TEs were not enriched on gene regions overall, a few insertions may be selected in the promoter of genes if they provide beneficial outcome to the individual fitness, through donation of histone marks enrichment, or cis-regulatory elements (35,36). We performed permutation tests to investigate whether TEs are enriched in the promoters (2 Kb upstream of the TSS) from the nine HLAU gene families in the seven genomes. To make a comparison of our findings with non-host shift related genes, we also included zinc-finger and serine/threonine kinases as two non-HLAU gene families. Regarding genes associated with location (OBPs and ORs), we observed that ORs have enrichment of TEs at the promoters in all genomes from species of the *mojavensis* cluster (p-value < 0.05), (Fig. 2A). For acceptance genes, *D. koepferae* had enrichment of TEs in the promoters of GRs (Fig. 2A). Finally, enrichment of TEs upstream to detoxification genes was observed in *D. arizonae* and *D. m. baja*. The latter had the enrichment for the CYP and Esterase families, whereas the latter had ABCs enriched by TEs (Fig. 2A). Overall, detoxification genes had z-score towards depletion of TEs on promoters, for which we observed significant depletion on *D. buzzatii* (Fig. 2A). In the non-HLAU gene families, we observed the expected result by chance: either z-score indicated the expected number of genes with TEs on promoters in a genome-wide comparison (z-score near 0), or it indicated depletion of TEs, as observed in *D. buzzatii* and *D. koepferae*. This result reinforces our findings indicating the significant and evolutionary unexpected enrichment of TEs, especially on the promoters of ORs.

The results depicting the histograms with frequency of genes with TEs for each permutation test, and p-values are shown in the Sup. Fig. 4 and Sup. Table 5. To obtain further insights in the potential regulatory role of TEs, we also analyzed the enrichment at the 2 kb downstream region. In *D. buzzatii* and *D. koepferae*, we did not observe significant enrichment, indicating

that the frequency of TEs downstream to HLAU genes follows the genomic distribution in these species. However, in the species of the *mojavensis* cluster, ORs had enrichment for TEs in the downstream region, as we have also observed for the 2 kb upstream. Finally, in the four gene families associated to detoxification, we observed that esterases are enriched by TEs in *D. arizonae*, while GSTs are depleted by TEs in *D. m. wrigleyi*. Our results suggest a possible ancient evolutionary path of TE cooption in species of the *mojavensis* cluster, since enrichment of TEs in ORs was observed in the five genomes.

The enrichment of TEs upstream and downstream of ORs indicates a possible coevolution of this gene family with TEs. Alternatively, it may also indicate that ORs are located in TE-rich regions of the genomes. To investigate this hypothesis, we identified TE-rich regions (Sup. Fig. 5) and analyzed the frequency of ORs. We observed on average 6 and 3 ORs in such regions in species of the *mojavensis* and *buzzatii* clusters, respectively (Sup. Table 6). Therefore, we rejected the hypothesis that OR genes are enriched by TEs in their vicinity due to their TE-rich genomic location. Other factors might be involved in the evolutionary maintenance of TEs nearby ORs, such as the cooption of regulatory motifs.

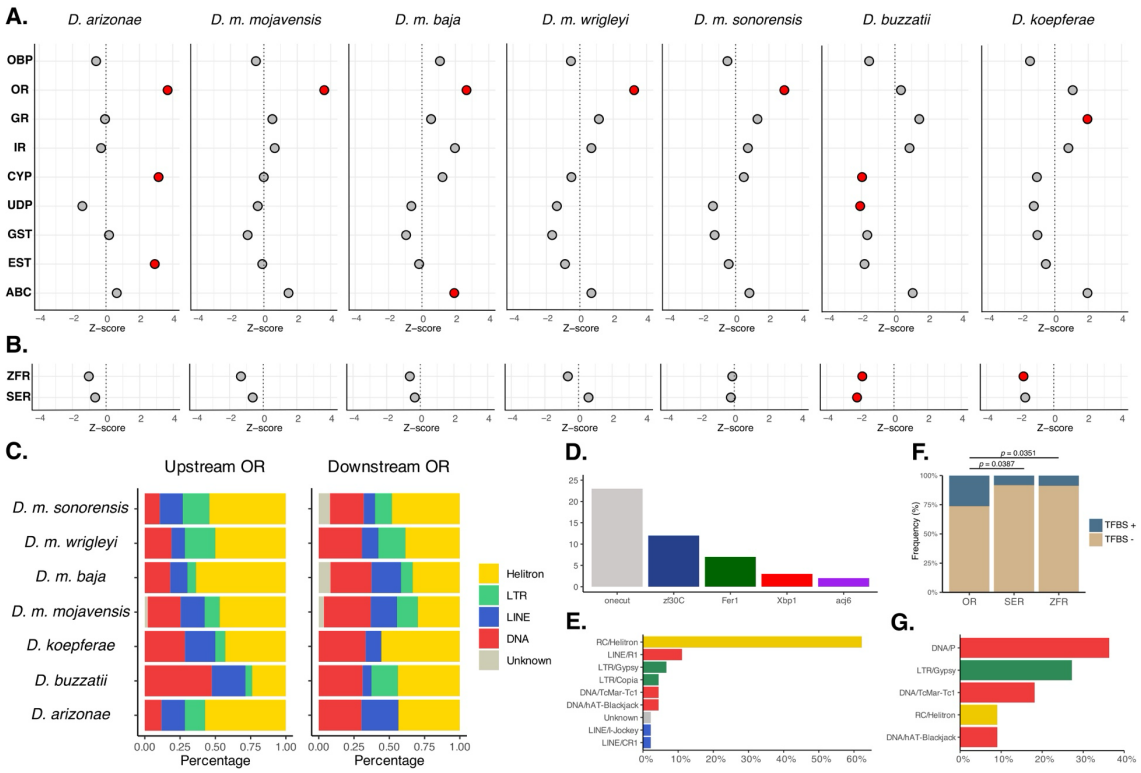


Figure 2. A) Enrichment of TEs in the promoter region (2kb upstream) of HLAU gene families. Z-score > 0 indicates enrichment, and Z-score < 0 depletion. Red dots: p-value < 0.05, grey dots: p-value > 0.05. B) The same enrichment analysis as depicted on A, but for non-HLAU gene families, represented by zinc-finger proteins (ZFR)

and Serine/Threonine Kinases (SER). C) Proportion of TE orders located upstream and downstream ORs. D) Transcription factor binding sites (TFBSs) frequency on TEs located upstream OR genes in all species under study. E) Proportional frequency of TE families with TFBSs embedded on their sequences from TE copies located upstream OR genes. F) Significant Fisher's exact test between OR genes with OR-related TFBSs from TE insertions compared to the frequency observed in the non-HLAU gene families Serine/Threonine Kinases (SER) and zinc-finger proteins (ZFR). G) Proportion of TE orders carrying OR-related TFBSs in non-HLAU gene families. Helitrons represent only 10% of the copies.

A few TE families have been shown to be more prone to contribute with cis-elements than others, such as retroelements in *D. melanogaster* (37). Therefore, we investigated whether the observed enrichment of TEs in the HLAU gene families is TE-specific. Considering all species, 50.42% and 40% of the insertions located upstream and downstream of ORs correspond to Helitrons, respectively (Fig. 3C). Helitrons have been documented to act on gene evolution of bats by carrying gene sequences (38), mainly 5' UTRs and the first exons, within their own internal regions (39). Thus, we analyzed whether these *Helitron* insertions upstream to HLAU genes, mainly ORs, could be part of the first exons from genes. By intersecting exons from genes enriched with their respective upstream Helitrons, we observed that only Or13a and Or22c have their first exons overlapping with the TEs. The former was observed only in *D. m. wrigleyi*, suggesting a subspecies-specific insertion, whereas the latter is observed in the four *D. mojavensis* subspecies, likely to be inherited from their common ancestor (Sup. table 7). At the ORs' downstream region, we observed five genes with *Helitrons* overlapping exons: Or71a, Or10a, Or13a, Or82a and Or30a. Interestingly, the same *Helitron* in the Or82a was observed in the four *D. mojavensis* subspecies, highlighting the inheritance from the common ancestor and maintenance of the insertion overtime (Sup. table 7). Taken together, our results rejected the hypothesis regarding the possible role of Helitrons carrying HLAU genes' fragments. However, these elements could still act on gene regulation if they carry cis-elements as TFBSs.

TEs nearby gene regions have been shown to modulate gene expression, through donation of TFBSs or polyA sites to the ancestral regulatory motifs. To investigate the potential functional role of the observed enrichment of TEs in the upstream region of ORs, we carried out TFBSs identification for all TE insertions located in such position. We used the transcription factors (TFs) previously described to participate on OR's transcription in *D. melanogaster*: acj6, onecut, xbp1, fer1 and zf30C (40). Our analysis revealed that 35.71% of the ORs with TEs at the promoter region have the TFBS located within the insertion (Sup. table 8). Taking all species together, the most frequent TFBSs found was for the TF onecut (48.93%), followed by zf30C (25.53%), and Fer1 (14.89%) (Fig. 3D). *Helitrons* were the TEs with the highest frequency,

representing 59.57% of all insertions carrying ORs' TFBSs (Fig. 3E). To test whether these results are specific to ORs, we also performed the same TFBSs' detection and frequency of TEs ZFR SER genes as non-HLAU families. Our results show that ZFR and SER have on average 9.75% and 9.09% of the genes with OR-related TFBSs, respectively, which can be considered as noise from the analysis. Comparing these proportions with the ones found in ORs, we had significantly more TFBSs compared to ZFR (Fisher's exact test: p-value = 0.0387) and SER (Fisher's exact test: p-value = 0.0351) (Fig. 2F and Sup. Table 9). In addition, *Helitrons* represented only 10% of the insertions carrying TFBSs in ZFR and SER genes (Fig. 2G). These comparative results suggest that our findings regarding the presence of OR-related TFBSs from TE insertions, and the enrichment of *Helitrons*, are unlikely to be due to chance. A few cases of ancestral inheritance were observed, for instance, the Or22c and Or82a have a conserved *Helitron* insertion carrying the onecut TFBS in the five genomes from the *mojavensis* cluster (Sup. table 7). We also observed a case of loss, where the Or85d has the TFBS Fer1 in the three *D. mojavensis* subspecies that uses primarily columnar cacti as host, but absent in *D. m. wrigleyi*. These results highlight a putative functional role of TEs as donators of cis-elements to ORs in the cactophilic species *D. mojavensis* and *D. arizonae*.

HLAU genes are differentially expressed between Opuntia sp. and columnar cacti dwellers

Differential expression of genes has been associated with adaptation to new hosts in insects (30). Here, aiming to understand how gene expression divergence is associated with adaptation to different cactus hosts, we performed differential expression analysis between species with preference for *Opuntia* sp. and columnar cacti, in head and larvae tissues. Overall, the principal component analysis (PCA) from HLAU gene expression revealed remarkable differences between species in both tissues (Fig. 3C). The variance of gene expression resembles the phylogenetic divergence of the species, with the highest variance between the *buzzatii* and *mojavensis* clusters (PC1), followed by divergence between *D. buzzatii* and *D. koepferae* (PC2). The species of the *mojavensis* cluster were clustered together with a slight separation between *D. arizonae* and the subspecies of *D. mojavensis*.

We performed pairwise analysis of HLAU genes expression between species with preference for *Opuntia* sp. (*D. m. wrigleyi* and *D. buzzatii*) versus species with preference for columnar cacti, and *D. arizonae* as non-preferential cacti. In head, *D. m. wrigleyi* had 27, 24, and 19 HLAU genes with higher expression compared to *D. m. mojavensis*, *D. m. baja*, and *D. m. sonorensis*, respectively (Fig. 3A). Four genes were found with higher expression in the head

of *D. m. wrigleyi* in the three comparisons: UDP-Ugt4, Gst-theta3, Abc-a3, Cyp6a21. In the subspecies with preference for columnar cacti, we observed on average 26 HLAU genes with higher expression compared to *D. m. wrigleyi*. Among them, seven were recurrent in the three subspecies: Ir21a, EstB1, UDP2, UDP-Ugt5, Cyp9b2, UDP-Ugt5-1, Cyp12a2. Since they were observed in all subspecies with preference to columnar cacti, the higher expression of these genes can be related to the adaptation to columnar cacti. Finally, the same analysis on larvae revealed that *D. m. wrigleyi* has three genes with recurrent higher expression compared to the other subspecies: Or82a, Obp59a, and Ir21a; whereas in the other three *D. mojavensis* subspecies we observed 17 common up-regulated genes compared to *D. m. wrigleyi*, which comprise seven CYP genes, three GSTs, four UDPs, one Esterase and one OBP (Fig. 3B). Regarding *D. buzzatii* and *D. koepferae*, we observed 27 HLAU genes in head and 54 in larvae with higher expression in *D. buzzatii*, and 63 in head and 29 with higher expression in *D. koepferae*. We also compared the head expression of HLAU genes between *D. m. wrigleyi* and its sister species without preferential cacti host, *D. arizonae*. In the *Opuntia* sp. dweller *D. m. wrigleyi* had 31 HLAU genes with higher expression compared to its counterpart *D. arizonae*. In larvae, *D. m. wrigleyi* had 43 HLAU genes with higher expression compared to *D. arizonae* (Fig. 3D). The result demonstrates that HLAU genes have divergent differential expression in both head and larvae tissues.

To further investigate whether differential expression of HLAU genes have higher prevalence compared to non-HLAU genes, we compared the proportion of differentially expressed genes associated to location, acceptance, and usage with the proportion of differentially expressed genes from the non-HLAU gene families Ser/Thr Kinases and Zinc-finger. In head, our analysis revealed that location and usage have higher differential expression than non-HLAU genes between *D. buzzatii* and *D. koepferae* (Sup. Fig. 6). In *D. mojavensis* subspecies, we observed higher differential expression for location genes between *D. m. wrigleyi* and *D. m. mojavensis*, and for usage between *D. m. wrigleyi* and *D. m. baja*. In larvae, we did not observe significant difference between non-HLAU and HLAU genes, but the pairwise differential expression between *D. m. wrigleyi* and three subspecies revealed higher prevalence of gene families associated with usage (detoxification) compared to non-HLAU (Sup. Fig. 6). In *D. m. wrigleyi* and *D. m. baja*, we also observed higher differential expression of location genes than non-HLAU. Finally, the differential expression between *D. m. wrigleyi* and *D. arizonae* demonstrated significantly higher proportion of genes associated with location and usage compared to non-HLAU in larvae, and higher prevalence of differentially expressed

usage genes in head (Sup. Fig. 6). Taken together, the quantitative results suggest that adaptations that have allowed the shift to columnar cacti may arise from the evolution of gene expression in both head and larvae.

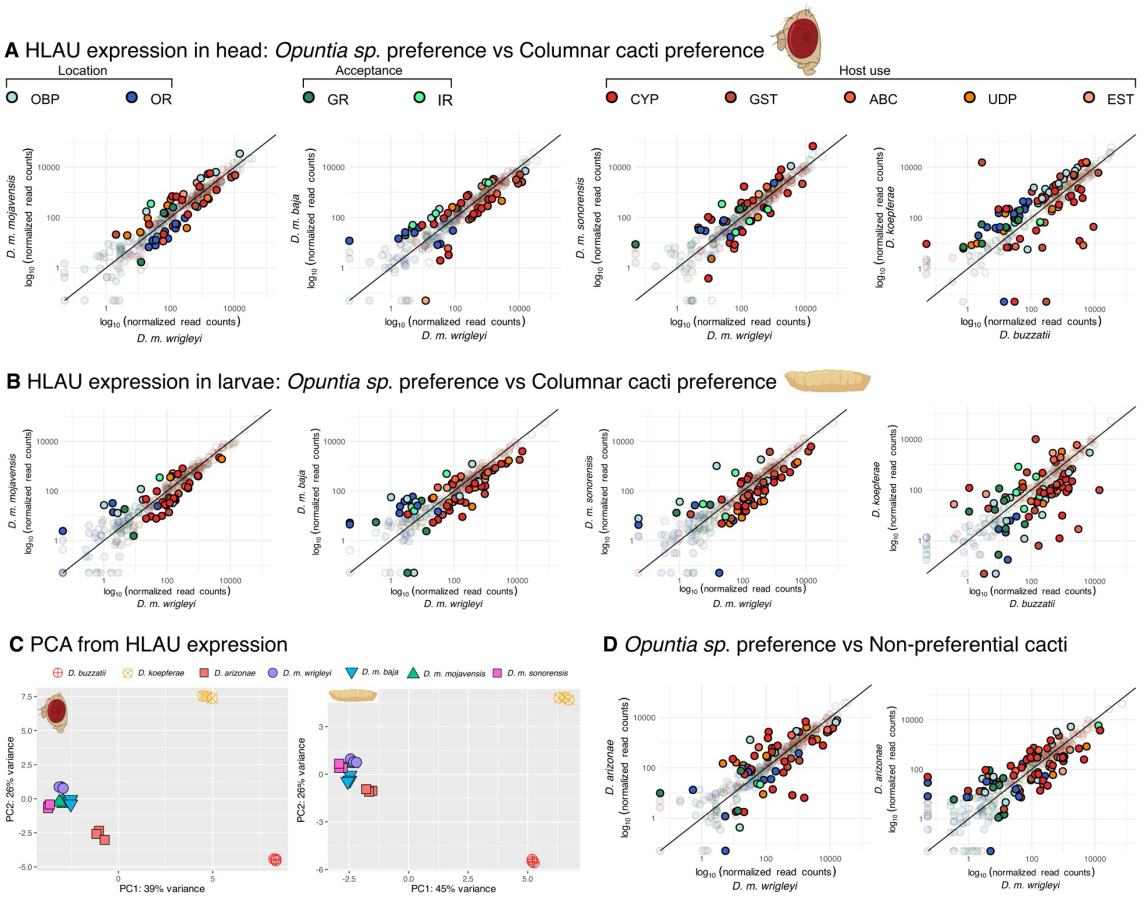


Figure 3: A) HLAU expression in head: differential expression analysis of HLAU genes from head in the *Opuntia* sp. dweller *D. m. wrigleyi* versus the three columnar dwellers *D. mojavensis*, and *D. buzzatii* versus *D. koepferae*. In A and B: Filled dots represent HLAU genes with log2 fold-change > |1| and FDR < 0.05; transparent dots are non-significant differential expression. B) HLAU expression in larvae: The same pairwise species as in A with *Opuntia* sp. dweller vs columnar dwellers, but from larvae tissue. C) Principal component analysis of HLAU genes' expression in head and larvae tissues. Both tissues had HLAU genes variance representing the phylogenetic distance between species: *D. buzzatii* and *D. koepferae* compared to *mojavensis* cluster corresponds to the PC1 (45%), and 26% between them in PC2. Species from the *mojavensis* cluster had strong similarity, with *D. arizonae* slightly separated from *D. mojavensis*. D) HLAU genes' expression in the *Opuntia* sp. dweller *D. m. wrigleyi* versus its sister species *D. arizonae*, which does not have preference between *Opuntia* sp. and columnar cacti.

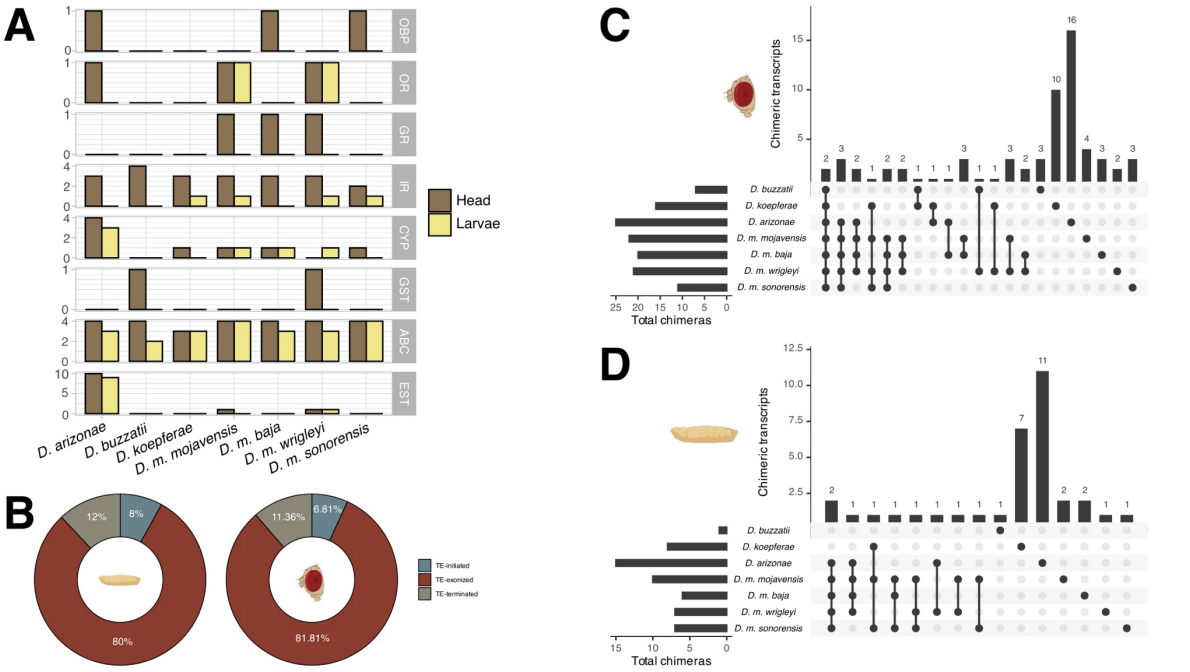
The enrichment of TEs on HLAU genes is not associated with changes on gene expression

Since we observed enrichment of TEs on the regulatory regions of a few HS families, mainly ORs, and we also found differentially expressed HLAU genes between species, we tested whether the presence of TE insertions is associated with differential expression. To do

so, we selected expressed HLAU genes (see Methods) and performed X2 independency test, splitting them into two variables with two categories each: 1) expression as columns: differentially expressed, and non-differentially; 2) TE presence as rows: with TEs at 2kb upstream, and without TE at 2kb upstream. Considering all pairwise analysis (Fig. 3A,B-D), the X2 results demonstrated that the differential expression of HLAU genes is not associated with the presence of TEs on the promoter region (Sup. Table 10), except in *D. m. sonorensis* compared to *D. m. wrigleyi* (X2 = 8.18, p-value = 0.004). Since we did not observe overall effect of TEs on the expression of all HLAU genes together, we investigated the association of TEs with differential expression of HLAU genes from families with TE enrichment (Fig. 2A). None of the enriched gene families had a significant association between the presence of TEs and their expression level (Sup. Table 10). Thus, enrichment of TEs at the regulatory region is not likely to be associated with the differential expression of HLAU genes.

HLAU genes generate chimeric transcripts

To investigate further cases where TEs produce chimeric transcripts with HLAU genes, and their possible role in the host preference between species, we identified all chimeras in head and larvae (Additional file 1) and focused our analysis on the nine HLAU gene families. In the head and larvae tissues, we observed an average of 5 and 3 HLAU genes producing chimeric transcripts per species/subspecies, respectively. We found more chimeric transcripts in head than in larvae for all HLAU gene families, except for UDPs (Fig. 4A). For instance, only head tissue had chimeras in the OBP, GR and GST gene families (Fig. 4A). TE-initiated transcripts constitute 11.36% of all chimeras found in the head, and 12% in larvae (Fig. 4B). TE-exonized transcripts had a slightly higher proportion in larvae (80%) compared to head (81.81%). Finally, TE-terminated transcripts were the ones with the biggest difference between tissues, with 6.81% in head and 8% in larvae. Taken together, these results demonstrate unprecedented transcriptional interaction between TEs and HLAU genes in a tissue-specific manner.


A
OBP
OR
GR
IR
CYP
GST
ABC
EST
Head
Larvae
D. arizonae
D. buzzatii
D. koepferae
D. m. mojavensis
D. m. baja
D. m. wrigleyi
D. m. sonorensis
B
12%
8%
80%
11.36%
6.81%
81.81%
TE-initiated
TE-exonized
TE-terminated
C
Chimeric transcripts
Total chimeras
D
Chimeric transcripts
Total chimeras

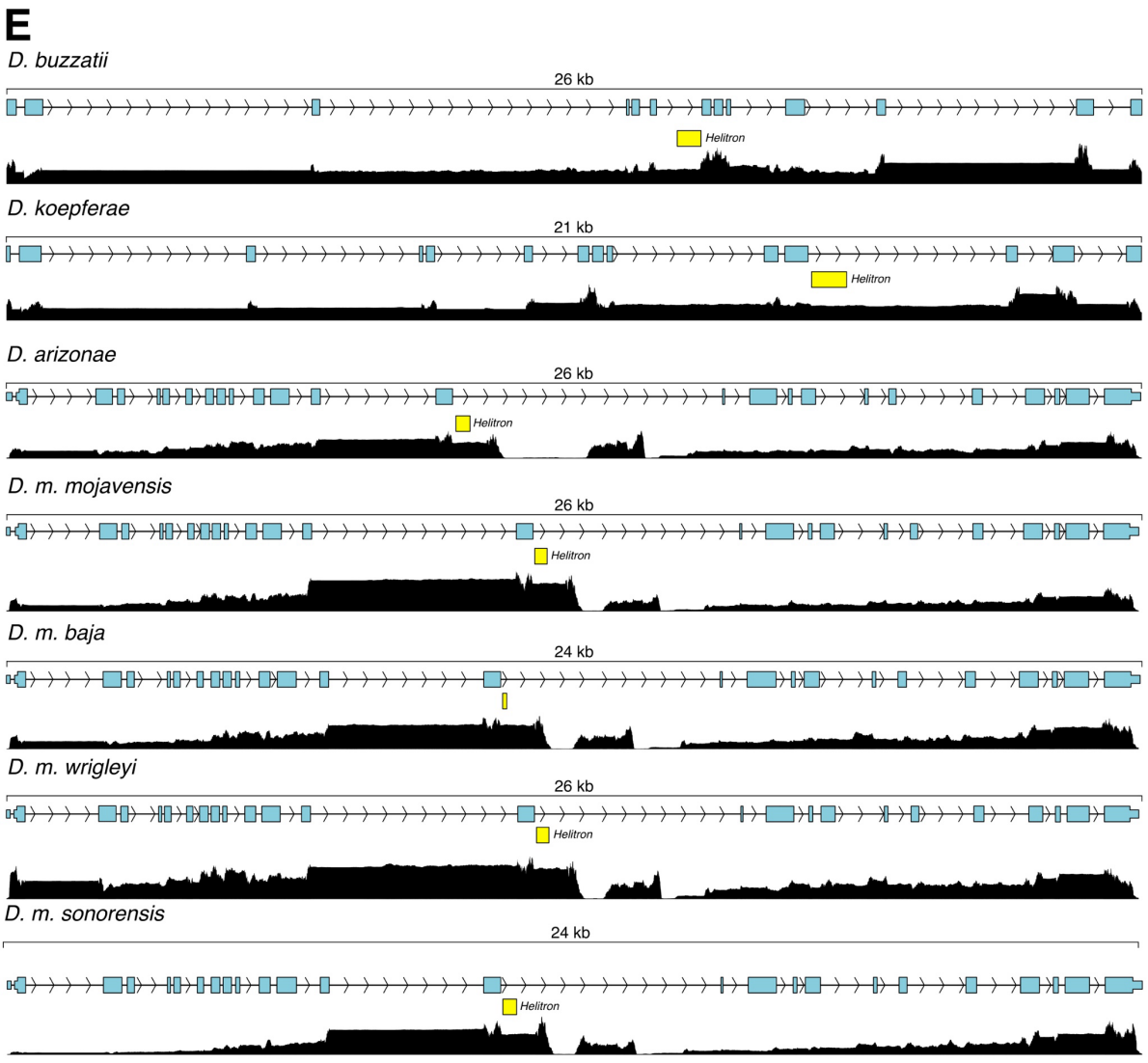

E
D. buzzatii
26 kb
Helitron
D. koepferae
21 kb
Helitron
D. arizonae
26 kb
Helitron
D. m. mojavensis
26 kb
Helitron
D. m. baja
24 kb
D. m. wrigleyi
26 kb
Helitron
D. m. sonorensis
24 kb
Helitron

Figure 4: A) Number of HLAU genes generating chimeric transcripts in each genome. B) Categories of chimeric transcripts in head and larvae tissues. C-D) Common chimeric transcripts between the seven genomes in head (D) and larvae (E) tissues. E) TE-exonized transcript from Glutamate IR, kainate 2 (LOC6585822) gene, with an *Helitron* insertion located in the intron. The same chimera was found in the seven genomes analyzed in this study.

The quantitative analysis on chimeric transcripts revealed a clear transcriptional polymorphism on HLAU genes due to TEs. We investigated common chimeras present in more than one species, highlighting their retention over evolutionary time, as well as species-specific chimeras, possibly associated with recent insertions. Our results revealed that 62.12%and 73.52% of the chimeric transcripts derived from HLAU genes are species-specific in head and larvae tissues, respectively (Fig. 4C-D). *D. arizonae* and *D. koepferae* were the ones with the highest number of HLAU genes with species-specific chimeric transcripts. In head, *D. arizonae* had 16 and *D. koepferae* 10; whereas in larvae we found 11 and 7, respectively. Regarding common chimeras, there are two expressed in head observed in all analyzed transcriptomes (Fig. 4E). The AbcD (LOC6585769) and Glutamate IR, kainate 2 (LOC6585822) have TE-exonized transcripts expressed in the head with *Helitron* insertions. In AbcD, the copy is located in the intron, ~100nt away from the nearby exon, representing a case where the chimeras generated through an exon retention process. In the IR, the same process is observed since the TE is also located in the intron. Regarding the buzzatii cluster, we observed the gene AbcA3 expressed only in *D. buzzatii* and *D. koepferae* with multiple insertions generating chimeras (Sup. Table 12). In the *mojavensis* cluster, two genes from the ABC gene family produced chimeras in the five transcriptomes. Abc10 had a chimeric transcript with an DNA/P insertion, which is embedded in the 5' UTR. Although the insertion generates the same chimera in the seven genomes, the element is not conserved between species. In *D. m. sonorensis*, it is a 3.9kb insertion overlapping the 5' UTR of the Abc10, whereas in the other subspecies the TE is fragmented in 3 small insertions. The second gene is AbcF2, which has a TE-exonized transcript derived from LINE/R1 insertion in all species, except *D. m. sonorensis*, which has an *Helitron*. Although not the same TE in all species, the copies are located partially in the intron and partially in the AbcF2's exon. Both recurrent chimeras in the species of the *mojavensis* cluster were observed in larvae and head. When narrowing down our analysis for chimeras observed only among *D. mojavensis* subspecies, we did not find any.

Discussion

Insect host shift and its role in speciation have been the subject of research to understand the evolution of reproductive isolating barriers and adaptation to local environments. Indeed,

many insects have finely tuned adaptations to locate, feed, reproduce, and develop on the host tissue. Adaptations to new hosts can lead to reproductive isolation, wherein ecological speciation arises as a consequence within divergent natural populations inhabiting distinct environments. In this context, we investigated whether TEs could play a role contributing to genetic novelty of HLAU genes facilitating rapid adaptive radiation to new hosts. To test this hypothesis, we selected seven *Drosophila* species/subspecies that belong to the *repleta* group and have different primary cacti hosts.

Relaxed selection on HLAU genes

The necrotic tissue of each cactus-host used by cactophilic *Drosophila* species differs in its composition in terms of yeasts and alkaloids (41,42). Therefore, species have adapted to feed and breed on different hosts, as demonstrated by behavioral and physiological response when species feed and breed in non-preferential hosts. Positive selection in the HS process acts as an evolutionary force, favoring mutations that lead to amino acid substitutions. These substitutions enhance the ability to recognize a broader spectrum of odors, metabolize nutrients, and activate detoxification pathways (43). Indeed, specialist species are likely to lose some of their genes related to sensory pathways, while the genes that are retained are subject to positive selection (44,45). A preliminary genome-wide analysis of signatures of positive selection in the four *D. mojavensis* subspecies revealed high ω rates on genes associated with chemosensory, perception, immunity, behavior, detoxification, and reproduction (46). Although a previous genome-wide analysis has identified 1,294 genes under positive selection (47), here we focused the analysis to investigate specifically HLAU gene families. Our results with branch-site model did not have any HLAU gene under positive selection when considering the species' host preference. We argue that perhaps the branch-site approach was too stringent. The unprecedent method used in this work focused on the common differences of species adapted to columnar cacti, rather than independent evolutionary events. The lack of genes with statistical support for positive selection may be also associated with technical limitations, since we based our analysis on 1:1 orthologs, and a certain proportion of the genes evolve too fast to be correctly assigned during orthologs identification (48). Additionally, the ω values are higher than 1.5, for instance as Gr59d with $\omega = 3.49$, which despite not significant still indicates ongoing divergence between species/subspecies.

Due to the lack of positive selection signatures from our branch-site test, we decided to choose to apply the RELAX model to identify evidence of relaxed selection across HLAU

genes. This model is particularly suited for detecting shifts from strong purifying selection to relaxed selection (49), which may not be identified by conventional methods focusing solely on positive selection, such as the one employed by CODEML. This is critical because evolutionary processes often involve not only episodes of adaptive evolution but also phases of selection relaxation, where selective pressures diminish but still influence gene evolution (50). The two HLAU genes (Obp56d and Cyp304a1) with significant relaxed selection suggest undergoing evolutionary transitions in response to changes in host preference. This result demonstrates that selective pressures on certain genes may diminish either during HS or after, leading to a period of relaxed selection as the species adapt to novel hosts. This relaxation of selection could indicate a functional change or reduced evolutionary constraint as these species exploit new ecological niches (51). Given that the species analyzed in this study have different primary hosts, the observed relaxed selection across these genes could reflect their involvement in the physiological or behavioral adaptations required for successfully exploiting new hosts.

OR promoters are enriched with TE insertions

The literature has recurrently demonstrated that TEs can create new regulatory elements, including gene promoters (52–54), enhancers (55,56), and insulators (57,58). In mammals, TE-derived sequences contribute up to 40% of genome-wide binding sites for transcription factors located in TE-rich regions (59). It has been proposed that such co-option of TE-regulatory elements for gene regulatory networks can provide substantial modification of gene expression over short time scales, contributing to generate rapid genetic variability on the transcriptome level (60–62). Because the invasion of new habitats is accompanied by modifications in behavior, which has been associated with alterations in sensory cues and environment perception (63,64), our results demonstrate the potential of TEs to change gene regulation of specific HLAU gene families on several species/subspecies, and thus contributing to the HS from *Opuntia* sp. to columnar cacti.

We observed species-specific enrichment of TEs in the promoter region for CYP, GR, EST, and ABC genes, but OR gene family had the highest enrichment in five out of seven genomes from our cactophilic *Drosophila* species. In ants, ORs have also been found enriched in TEs, but in this case they were associated with tandem duplication by crossing over, allowing massive duplication events (65). In our case, since there is no expansion of ORs compared to other *Drosophila* species, the enrichment in TEs is likely to have another function, probably related with gene regulation due to its specific location in promoters. *Helitrons* have been

extensively reported on promoter regions of other eukaryotic species (66–69), in which a few cases of embedded TFBSs were reported (70,71). Here, we demonstrated that ~60% of the enrichment in TEs observed at the promoter region of ORs might be due to the exaptation of TFBSs mostly derived from Helitrons. In addition, the recurrence of the enrichment in species of the *mojavensis* cluster might demonstrate potential cases of TE co-option inherited by common ancestor. Our genomic analysis provides a robust list of potential candidate genes in which Helitrons might play a regulatory role, an interesting hypothesis that can be determined through cutting-edge technologies such CRISPR methods to delete insertions in site-specific manner (72).

Head transcriptome reveals subspecies specific patterns

The differential expression of HLAU genes in the head tissue revealed that location and host use (detoxification) gene families have the highest divergence across the subspecies, since we observed more differentially expressed genes for these families. Furthermore, in *D. m. wrigleyi*, which is the only subspecies with preference for *Opuntia* sp., we observed that CYP genes had on average 12 up-regulated genes compared to other species. Taken together, these results highlight that the divergence of the host preference between *D. mojavensis* subspecies might be explained by modifications in gene expression of HLAU genes in the head tissue. In addition, despite lower toxicity in *Opuntia* sp. compared to columnar cacti (73), we found a similar number of up-regulated host use genes between *D. m. wrigleyi* and the other subspecies, demonstrating that *D. m. wrigleyi* exhibits a constitutive transcriptional response for detoxification that is not driven by the host. This is a plausible explanation since *Opuntia* sp. produces less benzaldehyde volatiles than the columnar cacti *S. thurberi*, but both cacti are lethal to *D. melanogaster* after 48 hours (73). Therefore, the adaptation to use *Opuntia* sp. must have a constitutive gene expression of detoxification pathways in *D. m. wrigleyi*.

The extension of differentially expressed genes corresponds to 6.17% of all genes analyzed. These results demonstrate that the regulation of host location genes might be an important factor that evolved and allowed host shift. Previous studies with *D. melanogaster* have shown that the divergence in the amino acid sequences due to single polymorphic mutations in ORs and OBPs have been suggested as factor of the modulation in odor sensitivity in *Drosophila* (64,74), which could drive to divergence of host location. In addition, many ORs are involved in a large repertoire of ecological relationships, influencing not only behavior, but also oviposition and feeding (75). Another study has identified 18 ORs differentially expressed

between heads of *D. m. wrigleyi* and *D. m. mojavensis* subspecies, demonstrating that modifications in the sensory transcriptome have evolved between these species, as previously reported for the same tissue (76). Here, we extended the comprehension of such divergence with the pairwise comparison between *Opuntia* sp. and columnar cacti dwellers, and pointed out that each species has a large set of differentially expressed HLAU genes.

Gene expression in larvae is one of the factors shaping the adaptation to new hosts

In the larvae tissue, we found a common gene expression pattern between columnar cacti dwellers, which in turn diverge from species that use *Opuntia* sp. Location and acceptance genes had mild alterations, whereas gene families associated with host use (detoxification) had on average more up-regulated genes than the others, especially due to overexpression of CYP genes. As we have observed for the head tissue, there is also an increase of CYP genes expression in *D. m. wrigleyi* larvae compared to the other subspecies. Indeed, in *D. m. wrigleyi*, we found the highest number of CYP up-regulation, with 23 CYP genes overexpressed in comparison to *D. m. sonorensis*. Among the other subspecies that preferentially use different species of columnar cacti as hosts, we also found detoxification genes with the highest numbers of differentially expressed genes. Similarly with our observations in the head tissue, these results indicate that larvae tissue exhibits subspecies-specific expression, demonstrating part of the genetic background that allows these subspecies to be specialized in specific cacti species.

The variance in gene expression in larvae revealed that species of the *buzzatii* cluster have higher divergence compared to *mojavensis* cluster, and lower between *D. buzzatii* and *D. koepferae*, following the phylogenetic relationship between the species. Although we have species with preference for *Opuntia* sp. and for columnar cacti in both clusters, *D. koepferae* and *D. buzzatii* are adapted to different hosts for a longer time than those of the *mojavensis* cluster. In addition, a preliminary study demonstrated that *D. buzzatii* have a high divergence on larvae gene expression compared to *D. koepferae* (77), reinforcing the variance that we have observed. Indeed, differences in larvae behavior and survival has been observed when *Opuntia*-related species are reared on columnar cacti (78,79). Our results show that divergence in the regulation of gene expression in larvae is one of the factors shaping the adaptation to new hosts.

The up-regulated HLAU genes were mostly represented by host use gene families, which are associated with detoxification. A previous study demonstrated that when *D. m. mojavensis*, which has *S. thurbery* as primary host, is reared on media with *S. gummosus*, the larval viability had a reduction of 60% (80). The authors also proposed that 21% of all genes would be

differentially expressed between larvae reared in the primary and alternative cacti-host. Many genes were reported with detoxification, energy production, among others. Another similar study with larvae transcriptome, but with *D. buzzatii* and *D. koepferae*, proposed that *D. buzzatii* flies reared in columnar cacti media have a stronger detoxification response in comparison with *D. koepferae* flies, suggesting that *D. koepferae* has a canalized transcriptome response to the toxic alkaloids from columnar cacti (77). We expected to find a higher number of detoxification genes in the subspecies with preference for columnar cacti compared to *D. m. wrigleyi* and *D. buzzatii* that use primarily *Opuntia* sp. However, we observed a similar number of differentially expressed detoxification genes regardless of the host preference. Since CYP genes also have other functions than detoxification, and they were overrepresented in our results, we suggest that their differential expression might also be associated with developmental pathways in larvae (81,82). Therefore, further analysis in the biological role of up-regulated CYP genes in *D. m. wrigleyi* and *D. buzzatii* might be assessed. Alternatively, despite *Opuntia* sp. has lower toxic compounds compared to columnar cacti, it also has a basal level of toxicity (73). Thus, any cactophilic species is expected to have evolved a certain level of detoxification ability, which explains our observations. In addition, the high frequency of common HLAU genes associated with detoxification on columnar cacti dweller species is expected and already demonstrated in previous works. Here, for instance, we observed that the three *D. mojavensis* subspecies have the same seven genes with higher expression compared to *D. m. wrigleyi*. Most of them are associated with detoxification (EstB1, UDP2, UDP-Ugt5, Cyp9b2, UDP-Ugt5-1, Cyp12a2), expect Ir21a. While the up-regulation of detoxification genes may be associated with the adaptation to higher concentration of alkaloids, the Ir21a is likely to play an important role in columnar cacti recognition and general chemosensation.

Transposable elements are a source of transcriptome variability between cactophilic species

TEs contribute to genome and transcriptome evolution by rewiring regulatory networks, and by incorporating protein domains into gene transcripts. In both cases, TE copies that are inserted within genes can integrate their sequences into the mRNA, generating chimeric transcripts. A recent study with different ecotypes of *D. melanogaster* has shown that these chimeras have the potential to generate transcriptome novelties (23). These chimeras have also been demonstrated to be active in a tissue-specific manner (83). Although the production of chimeric transcripts derived from polymorphic TE insertions is well-known, there is no study addressing the question on how the transcriptomic variability derived from TEs could impact the HS between species/subspecies. Here, we took advantage of our unprecedented genomic

and transcriptomic data to analyze the extent of genetic variability derived from chimeric transcripts in cactophilic species, as well as HLAU genes producing chimeric transcripts.

A preliminary transcriptome-wide study with *D. melanogaster* revealed that ~1.12% of the genes produce chimeric transcripts (23). Here we observed ~1.31% on average between head and larvae tissues. The slightly higher number of genes in *repleta* species might reflect the difference between their mobilome composition compared to D. melanogaster, which is depleted of active *Helitrons* for example (84). Despite this difference, the proportion of TE-exonized, TE-terminated, and TE-initiated transcripts follows similar values between them, in which TE-exonized have the higher frequency with ~70% of all chimeras (23,83). In our analysis, TE-initiated transcripts represent chimeric transcripts in which the TSS of the mRNA is located within the TE insertion (23). *Helitrons* were the most frequent TEs falling in the category in four of the seven species. They have been extensively reported as a source of TFBSs when located at the promoter region of genes. In the bat *Myotis lucifugus*, 1.4% of the transcripts from the salivary gland have *Helitron*-derived sequences, for which ~6% correspond to promoters (67). In the plant Brassica napus, an *Helitron* insertion disrupts the ancestral promoter of the gene BnSP11-1, altering the mating system of the populations carrying this insertion, which in turn contributes to the adaptation and speciation (85). In *Arabidopsis thaliana*, *Helitrons* have been responsible for the genome-wide distribution of the TFBS for the PHE1 TF, which is associated with crucial endosperm regulation (70). In *Caenorhabditis* species, TFBSs for the TF heat shock factor 1 have been widely spread by *Helitrons*, rewiring and diversifying heat shock response between species and strains (86). Combining our results from TFBSs from OR genes with our transcriptome-wide analysis of TE-initiated transcripts, the results reinforce that Helitrons have a cis-regulatory role in cactophilic *Drosophila* species. Regarding TE-terminated transcripts, the high prevalence of *Helitrons* in six transcriptomes also highlights their role on the termination of gene transcription. This phenomenon has been already described in bats, in which the genes Otud3 and Tor1b have novel PASs from *Helitrons* (39). Another study with A. thaliana revealed by poly(A)-tag sequencing data that ~48% of all TE-derived PASs are from *Helitrons* (87). Although there is vast documentation of TEs acting as alternative PASs (88), here we provide the first transcriptome-wide report of *Helitrons* as a major source of putative PASs in *Drosophila*.

Chimeric transcripts reveal recent and old insertions interacting with HLAU genes

A comparative analysis between generalist and specialist *Drosophila* species revealed that niche amplitude is not likely to play a role in TE dynamics in the genomes (89). Nevertheless, several studies have proposed that the successful adaptation to new environments in invasive species might be associated with TE activity, despite reduced genetic diversity caused by founder effects (90,91). Through TE-initiated and TE-terminated transcripts, we provide evidence for TE insertions playing a regulatory role on HLAU genes, whereas TE-exonized transcripts suggest a putative source for TE-driven coding sequence evolution. Here, we demonstrate that TEs interact with genes associated with HS, providing a transcriptomic variability in seven cactophilic species with different hosts preferences. We observed that species *D. arizonae*, which does not have preferential host usage between *Opuntia* sp. and columnar cacti, are the ones with the highest production of unique HLAU chimeras in both head and larvae. This result demonstrates a possible relationship of broader niche and TE-driven adaptation.

The detection of chimeric transcripts is a reliable way to infer transcriptional interactions between genes and TEs. Nevertheless, even applying strict parameters as we did, the contribution of the chimeric isoform to the total gene transcript abundance is unknown. Thus, perhaps the integration of TE-derived sequences into the gene transcript is not beneficial, but still maintained in the transcriptome due to its low expression. Furthermore, chimeric transcripts might be degraded by surveillance pathways that control for aberrant mRNA production, such as no-go decay (92), non-stop decay (93), and non-sense-mediated RNA (94). Our study provides a set of HLAU genes producing chimeric transcripts, in which part of them might have phenotypic impact and ultimately association with host preference. However, it is fundamental to perform comparative phenotypic studies of host adaptation cues with mutant strains for the TEs interacting with HLAU genes identified in this study. The application of techniques of RNAi aiming chimeric transcripts, or deletion of TE insertions producing chimeras with CRISPR (72) will be able to confirm the predictions of our results.

Transposable elements driving host shift

Here, we demonstrated that HLAU genes are enriched by TEs in their regulatory regions, and produce different chimeric transcripts across species with different primary hosts. Host specificity is a key mechanism of reproductive isolation in plant pathogens, and is regulated by the repertoire of effector genes within each pathogen. In *Phytophthora* sp., the host specificity

is partly regulated by RXLR class effectors that facilitate host exploitation. Notably, synthetic chimeras of a short interspersed element (SINE) linked to an effector gene in *P. infestans* induce silencing of effector genes (95). This silencing likely occurs naturally, as transcriptional inactivation of effectors is known to occur, and over half of RXLR effectors in the *P. infestans* genome are located in TE-rich regions (96). Consequently, TEs inserted near these genes may have influenced HS in *P. infestans*. Our findings demonstrate multiple interactions between genes and TEs, highlighting the contribution of TEs to the genome evolution of cactophilic species. Through the TE-cooption process, we postulate that TEs can play a role in the HS in these species. Although many TEs located on promoters of HLAU genes can have low frequency in natural populations, or chimeric transcripts identified in this work can be a result of ectopic expression instead of adaptive, our results demonstrate the first transcriptome-wide evidence of TE co-option in cactophilic species. The genes producing chimeras with TEs must be further studied, since the lack of gene-phenotype association between HLAU genes is not fully reported, preventing us to pinpoint specific relation between chimeras and possibly reproductive isolation.

The existence of reproductive barriers between populations that undergone a host shift is the key process to ecological speciation. In our model, with species from the *buzzatii* and *mojavensis* clusters, incipient reproductive isolation has been reported between *D. m. sonorensis* and *D. m. baja* (25,97,98). The incipient premating isolation between them occurs due to the difference on cuticular hydrocarbons, which participate in the signaling pathway of mate recognition (99). Importantly, artificially shifting the cactus-host was the main cause of changes on the hydrocarbon profiles. Thus, the source of reproductive barriers is dependent of the cacti species on which the flies feed and breed (100,101). The close relationship between HLAU genes to the HS process, and the observed contribution of TEs to the evolution of these genes, supports our proposal that TEs may play a role in the HS of cactophilic species. Hence, such process rises the hypothesis that TEs can play a role in HS. Hence, such process rises the hypothesis that TEs could drive reproductive isolation via adaptation to different environments, ultimately driving ecological speciation.

Conclusions

TEs have been shown to contribute to genome evolution of cactophilic *Drosophila* species. Their enrichment in promoter regions of HLAU gene families, along with potential presence of TFBSs within their sequences highlights their potential role in the evolution of

adjacent genes. Additionally, their frequent exonization reveals a mechanism by which TEs increase genetic diversity between populations, potentially contributing to host shifts. Taken together, our combined genomic and transcriptomic approaches provide new insights regarding TE-related mechanisms that promote genome evolution between species and subspecies with different primary cactus hosts.

Material and methods

Fly stocks and genome sequencing

Most of the strains were obtained from the UC San Diego *Drosophila* Stock Center: [*D. m. mojavensis*: Anza (01), Anza Borrego Desert, California, USA, Stock Center n°: 15081-1352.01; *D. m. wrigleyi*: CI (22), Catalina Island, California, USA, Stock Center n°: 15081-1352.22; *D. m. baja*: BC (20), Cape Region, Santiago, Baja California Sur, México, Stock Center n°: 15081-1352.20 e *D. m. sonorensis*: AG (26), Agiabampo Bay, Sonora, México, Stock Center n°: 15081-1352.26] and *D. arizonae* [HI (17), collected in Metztitlan, Hidalgo, Mexico, Stock Center n°: 15081-1271.17]. *D. buzzatii* (Bu28) and *D. koepferae* (Ko2) stocks correspond to two laboratory lines used in previous works (102,103). Both stocks were maintained by brother-sister mating for more than a decade and then kept by mass culturing.

We used the same following protocols of DNA extraction and genome sequencing to all samples. DNA was extracted from 10 males and 10 females from each species using the Qiagen DNeasy Blood&Tissue kit. Then, we evaluated the genomic DNA amount and quality with NanoDropTM One UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and Qubit ® 1.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Three micrograms of DNA were repaired using the NEBNext FFPE DNA Repair Mix (NEB M6630). We performed end repair and dA-tailing using the NEBNext End repair/dA-tailing Module (E7546, NEB). Ligation was then assessed with the Ligation Sequencing Kit 1D. MinION sequencing was performed according to the manufacturer's guidelines using R9.4.1 flow cells (FLO-MIN106, ONT) and a Nanopore MinIon Mk1b sequencer (ONT) controlled by ONT MinKNOW v.18.3.1. Base calling was performed after sequencing using Guppy v.4.0.5 in high accuracy mode for isogenic wild-type strains v.3.1.5.

Genome assemblies and gene annotation

The quality control for Nanopore reads was performed with Nanoplot v.1.10.2 (https://github.com/wdecoster/NanoPlot). Reads with quality lower than 7 were removed from

downstream analysis. In order to assemble contigs, Flye v.2.8 (104) was used with default parameters, except –plasmids. All contigs were aligned with minimap2 v2.16 (105), with –x map-ont option. The alignment was used to perform the polishing of contigs with four rounds of RACON v1.3.2 (106), default parameters. The assembly quality metrics was assessed with Assembly-Stats v1.0.1 (https://github.com/sanger-pathogens/assembly-stats). Assembly incongruences were manually visualized with D-genies v1.2.0 (107) and corrected with samtools v1.9.0. (108), function faidx; and Gepard v1.4.0(109) for determination of breaking points. The super scaffolding of all corrected assemblies was performed with RaGOO v.1.1 (110), with –s and –t 4 parameters, using the respective reference genome assembly for each species. Subsequently, the chromosome-scale assemblies were submitted to a Benchmarking Universal Single-Copy Orhologs (BUSCO) v.5.4.3 analysis (111) with lineage diptera_odb10 (-l parameter) with nucleotide sequences. The gene annotations for *D. mojavensis* subspecies and *D. arizonae* were performed with the software Liftoff v.1.6.3 (112), with the reference genome of *D. mojavensis* (GCA_018153725.1) (113). We have used the parameters –exclude_partial; -a 0.5 and –s 0.75 to keep annotated genes with at least 75% of the reference gene length. We removed from our gene annotation HLAU genes flagged as “low quality protein” and “unestablished gene function” in the reference *D. mojavensis* genome.

D. buzzatii and *D. koepferae* genomes were annotated with a de novo strategy with the software BRAKER3 v.3.0.8 (114). To optimize the prediction of gene structures and repertoire of isoforms, we used a mix of RNA-seq data from different tissues publicly available and produced in this work (Sup. Table 11), in addition to protein sequences from *D. mojavensis* as reference. Ortholog genes between *D. buzzatii*, *D. koepferae*, *D. mojavensis* and *D. arizonae* were obtained with Orthofinder v.2.5.4 (115) using DIAMOND v.0.9.14 (116) as search engine and default parameters. One-to-one orthology were firstly obtained directly from orthofinder, and remaining proteins without an ortholog pair were analyzed with blastp v.2.5.0+ (117), for which we considered as one-to-one proteins only those with a single homologous with parameters -qcov_hsp_perc 95 and identity >= 90%. Since Orthofinder performs its search with protein sequences, any non-coding RNA gene has been identified. Then, we used Liftoff v.1.6.3 (118) with all non-coding RNA genes from *D. mojavensis* as reference to maximize our homology-based gene annotation.

Molecular evolution analysis

To identify HLAU genes with signatures of positive selection in species that have preference for columnar cacti compared to species with preference for *Opuntia* sp., we selected the four *D. mojavensis* subspecies, D. navojoa, *D. koepferae* and *D. buzzatii*. *D. arizonae* species was removed from this analysis due to its generalist preference for both columnar and *Opuntia* cacti. The D. navojoa genome was retrieved from NCBI (GCF_001654015.2), which was produced with Illumina reads (119). For the other Nanopore genomes sequenced in this work, we performed a polishing with Illumina paired-end reads. We used the software NextPolish (120) to fix base errors in the assembled genomes. The four *D. mojavensis* subspecies were polished with genomic Illumina reads sequenced previously (121), whereas for *D. koepferae* and *D. buzzatii* we used publicly available genomic Illumina reads (122) and RNA-seq reads (47), respectively. In the latter, to increase the read coverage and optimize the polishing to the higher number of genes and exons as possible, we merged RNA-seq reads from adult male and female (47). All genomes were corrected using two rounds of polishing, as recommended by the authors (120). Then, we extracted the longest CDS sequences for all genes combining AGAT (123) agat_extract_sequences.pl parameters -t mrna, samtools faidx (124), and seqtk (https://github.com/lh3/seqtk). Posteriorly, we intersected the occurrence of orthologous genes between the seven genomes, recovering a list of genes that were present in all of them following Orthofinder analysis. The sequences were aligned using MAFFT v.7.487 [67] and the non-aligned codons were removed using pal2nal (125). To obtain a strongly supported tree for this analysis, we inferred the species phylogeny tree with 2,043 shared BUSCO genes. The phylogeny was estimated using the concatenated alignment (3,730,914 sites) with IQ-TREE v. 2.2.2.6 (126), using the best model for each gene, 10,000 replicates of ultrafast bootstrap and 1000 replicates for Approximate Likelihood-Ratio Test (ALRT) (parameters -m MFP -bb 10000 -alrt 1000). Finally, we carried out analysis to detect signatures of selection analysis for each gene individually with CODEML from PAML v.4.10.7 (127), using branch-site approach. In order to find signatures of positive selection in species that use columnar cacti, we marked the respective branches of *D. koepferae*, *D. m. baja*, *D. m. mojavensis* and *D. m. sonorensis*, with the same label (#1), to compare them with the background species that use *Opuntia* sp. (*D. buzzatii*, *D. m. wrigleyi*). For this, we tested two hypotheses: (i) H0 – Null hypothesis, in which all branches evolved under negative selection (model = 2, Nssites = 2, fix_omega = 1, ômega = 1); (ii) H1 – Alternative hypothesis, in which the labeled branches (foreground) have sites with signatures of positive selection (model = 2,

NSsites = 2, fix_omega = 0, omega = 1.5). The hypotheses were tested through chi-square test, calculated by the log-likelihood (lnL) ratio between H1 and H0 (2ΔlnL) and degrees of freedom obtained from the difference between the number of estimated parameters for H1 and H0 (Δnp). Sites under positive selection were identified by Bayes Empirical Bayes and posterior probability, implemented in CODEML.

We used the software HYPHY (128) to test relaxed selection with the same phylogeny framework used for branch-site analysis, using the RELAX model (49). The phylogenetic tree had its branches corresponding to columnar cacti dwellers as foreground, and *Opuntia* sp. as background. We used the parameters --test Foreground and --model Minimal. Multiple testing correction was performed with False discovery rate (FDR). Genes with FDR < 0.05 and k > 1 were considered under intensified selection.

RNA extraction and sequencing

We performed RNA-seq for larvae and head tissues from *D. buzzatii*, *D. koepferae*, *D. arizonae* and the four *D. mojavensis* subspecies. The RNA extraction was assessed with RNeasy Plus Kit for 10 female heads (10 days old) and 15 larvae with three technical replicates. The library preparation was done with stranded mRNA-seq/standard quantity (ligation), and 100 bp paired-end reads were produced in an Illumina Hiseq 4000. Illumina adaptors and low quality reads were removed for downstream analysis with Trimommatic v.0.39 (129), parameters: LEADING:3; TRAILING:3; SLIDINGWINDOW:4:15; HEADCROP:10; CROP:90.

TE annotation

For each sequenced genome, we developed a novel pipeline combining automatic and manual approaches to construct TE consensuses. First, we used Earl Grey v.2.2 (130) to build de novo TE libraries with RepeatModeler2 v.2.0.4 (131), which uses RECON v.1.08 (132) and RepeatScout v.1.0.6 (133). Then, in order to retrieve full-length TE consensus, the pipeline performs an extension of each predicted consensus sequence using a method named “Blast, Extract, Extend” (BEE) (134). Such method extracts the sequences of the 20 full-length (or longest) TE insertions for each consensus, and then it retrieves the 1 kb flanking regions from them. Posteriorly, insertions altogether with flanking regions are aligned with MAFFT v.7.487 (135), and the alignment is trimmed with trimAl v.1.4 (136), parameters –gt 0.6 and –cons 60. Subsequently, the alignment is analyzed to check whether the flanking regions from the full-length insertions have high quality, meaning that they are part of the TE rather than genomic

regions. Finally, the consensus sequence for each TE family is updated with EMBOSS v.6.6.0.0 (137). The BEE method is repeated up to five times for each consensus. Once the extended TE libraries were obtained for the seven genomes, they were submitted to a CDS filtering step, in which we removed all consensus with high similarity with *D. mojavensis* wrigleyi CDSs using blastn, parameter -qcov_hsp_perc 80. A last filter step was applied to remove TE consensus with less than 3 RepeatMasker hits across the genome, with at least 80% of its length. A similar cutoff has been applied to D. suzukii TE annotation (90).

TE consensus that had passed in our filtering thresholds were then submitted to the classification analysis. We developed a pipeline to perform similarity analysis of non-reference consensus with a database provided by the user (https://github.com/OliveiraDS-hub/Pipelines-Cactophilic-*Drosophila*-Species). The pipeline identifies similarity analysis between the non-ref. consensus and a database provided. All consensuses were masked with RepeatMasker v.4.1.4 (138), parameters -lib, -cutoff 250, -s, ¬and -norna. Then, the pipeline performs different steps depending on the number and variety of hits that the non-ref. consensus has: (1) only one hit is considered unique and the total length of the masked hit is compared with the total length of the non-ref insertion; (2) multiple hits, but all from the same class and family: the length of the hits is summed up, and the total length is compared with the length from the non-ref consensus; (3) multiple hit, but different classes and/or families: in this case each class is treated separately. Due to the possibility of overlapped hits from TE families of the same class in the non-ref consensus, the pipeline removes the extent of overlapped regions from two insertions, hence decreasing the length of the insertions. Once all hits from a given class have no overlap, the pipeline sum up the total extent of the hits and compare it with the non-ref consensus. The process is repeated for all classes masked. Finally, the pipeline classifies only TEs with a corresponding match in >= 50% over their total length. In this work, we have used Dfam v.3.7 (139) as reference database. Due to the "curated" library contains only *D. melanogaster* consensuses, we used the "uncurated" library from Dfam. We extracted *Drosophila* TEs with famdb.py v.0.4.3, parameters -i Dfam.h5 families --include-class-in-name -f fasta_name -ad 'Diptera'. Then, we filtered out non-*Drosophila* sequences, and "Unknown" TEs with the unix commandline: grep -i '*Drosophila*' file.fa | grep -v -i 'unknown', obtaining 49,916 out of 344,562 consensuses. Finally, we used all obtained IDs with seqtk v.1.3 (https://github.com/lh3/seqtk) to subset the complete fasta file generated by famdb.

We used the polished TE libraries constructed to the seven cactophilic *Drosophila* genomes to perform TE annotation with RepeatMasker v.4.1.4 (138), parameters -cutoff 250, -

s, and -norna. Then, we have used RepeatCraft (140) with LTR Finder v.1.2 results to merge LTRs and their respective internal regions. Subsequently, we removed all TE insertions shorter than 80 nucleotides, following the classification rule 80-80-80 (141). Finally, we performed a last filtering step in the TE annotation to remove insertions comprised by simple sequence repeat (SSR). We used TRF v.4.09 (142) in all TE insertions from each genome, with parameters 2, 5, 6, 75, 20, 50, 500, -m, -h, and removed TEs that had SSRs over more than 50% of their sequences. Such method has been previously used to remove TE copies that are indistinguishable from simple repeats (23,143).

TE enrichment at regulatory regions

To identify TE insertions within and nearby genes for each genome, we created bed files with the genomic positions of all insertions. Then, we created bed files for the gene regions corresponding to the 2 Kb upstream to the transcription start-site (TSS). In order to test whether the prevalence of TEs in each gene region was significantly higher than expected (enriched) compared to random genomic distributions of TEs in the same region of other genes, we performed permutation tests with the R package regioneR (144), parameter randomization = resampleRegions, with 2,000 random samples. Therefore, for each gene family with n genes, the number of genes with TEs was compared to other random n genes 1,000 times. Such analysis allowed us to infer whether there are HLAU gene families with the observed frequency of genes with TEs at the promoter region higher than expected compared to the rest of the genome. We also selected zinc-finger proteins and serine/threonine kinases as two non-HLAU genes families to compare our findings. Taking into account that TEs might harbor regulatory and/or coding motifs in both strands (62), our analysis considered TEs within the investigated gene regions regardless of the strand.

Identification of TE-rich genomic regions

We split the genomes in bins of 200 kb, generating a bed file with such intervals across the genomes. Then, we used the function intersect from bedtools v.2.30.0 (145) to identify the number of TE insertions located in each genomic bin. Finally, we calculated the 3rd quartile with the frequency of TEs in each region and defined as TE-rich genomic regions those with more TE insertions than the 3rd quartile. We used the same function from bedtools to identify whether genes were present in the TE-rich regions, or in the TE-poor (less TEs than the 3rd quartile).

Transcription factor binding sites on TE insertions

We assessed the identification of transcriptions factor binding sites (TFBSs) with FIMO (146), from the MEME v. 5.0.5 suite tools (147). We extracted the sequences of all TE insertions located 2kb upstream of ORs with bedtools v. 2.30.0, function getfasta, using parameter -s (145). We obtained the TFBS motifs described for *D. melanogaster* from Jaspar database (148), for the transcription factors (TFs) acj6, zf30c, fer1, onecut, and xbp1, which have been demonstrated to activate ORs (40). To control our findings with OR genes, we used IR as a HLAU gene family member, and Serine/Threonine Kinases as a non-HLAU gene family. We performed The same process of TFBSs identification performed for ORs on these two gene families. We only considered TEs with TFBSs those with p-value < 0.01 and score > 250.

Differential expression of genes

We aligned the RNA-seq reads with STAR v.2.7.3 (149), using the assembled genomes sequences and annotations of their respective species. The alignments were used to create the count table reporting the expression of each gene, with htseq v.0.13.5 (150), parameters –order pos; -s reverse; –i gene; -t exon; the latter indicates that the count for each gene corresponds to the sum of all reads aligned into exons (CDSs and UTRs). We have merged the counts from species of the *buzzatii* and *mojavensis* clusters separately, maintaining only ortholog genes that have at least 90% of the length of the annotated gene in the reference genome. Such threshold allowed us to avoid misinterpretation of expression level by comparing genes with different lengths. Then, we performed independent statistical analysis for species of the *buzzatii* and *mojavensis* clusters with DEseq2 v.1.28.1 (151), using single factor experimental design, which considers only the factor that different species may have differentially expressed genes. Due to the phylogenetic distance between species of the *buzzatii* and *mojavensis* clusters, we performed the differential expression analysis separately for each group. We only considered genes with adjusted p-value < 0.05 and |log2 fold-change| > 1 as differentially expressed. To test for association between TE enrichment and differential expression, we selected HLAU genes with the mean of normalized counts between replicates >= 100. Then, we build the contingency table to apply X2 independence test splitting them into four categories: differentially expressed, non-differentially expressed, with TEs at promoter region, and without TEs at the promoter region. We carried out the test for all pairwise analysis between cactophilic species with preference for *Opuntia* sp and columnar cacti. The proportional differences between location, acceptance and usage genes compared to non-HLAU genes (Ser/Thr Kinases

and zinc-finger gene families) were assessed with Fisher's exact test, with contingency table containing average number of genes per each group: 1) non-HLAU differentially expressed genes; 2) non-HLAU non-differentially expressed genes; 3) HLAU differentially expressed genes; and 4) HLAU non-differentially expressed genes. Fisher's Exact test < 0.05 were considered as significant.

Identification of gene-TE chimeric transcripts

We detected gene-TE chimeric transcripts with ChimeraTE v.1.1.1 Mode 1 (23). Each species was analyzed using its respective genome, gene annotation, and TE annotation. Shortly, ChimeraTE identifies chimeras based on paired-end reads spanning from exons to TE sequences. Based on the TE position compared to the gene (upstream, exon, intron, downstream), chimeric transcripts are classified in categories. We have used default parameters, except --replicate 3, which consider true chimeras only those identified in the three RNA-seq replicates, and –coverage 10 to obtain chimeras with at least 10 chimeric reads on average between replicates.

Declarations

Availability of data and materials

The quality control data from both ONT and RNA-seq data, the genome assemblies and their respective gene and TE annotation (libraries and gtf files), and the data used to produce phylogenetic are available in the Zenodo repository, doi: https://zenodo.org/records/13117512. The code to reproduce the analyses are available on Github repository: https://github.com/OliveiraDS-hub/Pipelines-Cactophilic-*Drosophila*-Species.

Competing interests

The authors declare that they have no competing interests.

Funding

The present study has been funded by Brazilian agencies FAPESP- The São Paulo Research Foundation [grant number 2020/06238-2] and from the National Council for Scientific and Technological Development (308020/2021-9) to C. M. A. C; the Eiffel Fellowship Campus France [P769649C] and CAPES – PrInt [88887.716810/2022-00] to D. S. O.; by Ministerio de Ciencia e Innovación (Spain) [grant number PID2021-127107NB-I00], and Generalitat de

Catalunya (Spain) [grant number 2021 SGR 00526] to M. P. G. G.; Udl-FAPESP ANR-IDEX-0005 to C. V.; and by the Agence Nationale de la Recherche [ExHyb, ANR-14-CE19-0016-01] to C. V. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 to DSO.

Authors' contributions

A. L., A. B, M. P. G. G, performed experiments; F.S., Performed genome assemblies; W. V. B. N. performed phylogenetic analysis; D. S. O performed computational analysis for genomics and transcriptomics analysis; D. S. O. wrote and edited the paper; C. V. and C. M. A. C. designed the study, and reviewed the manuscript. All authors read, corrected, and approved the final manuscript.

Acknowledgements

We thank Séverine Chambeyron for useful discussions and advice. This work was performed using the computing facilities of the CC LBBE/PRABI.

References

1. Whiteman NK, Pierce NE. Delicious poison: genetics of *Drosophila* host plant preference. Trends in Ecology & Evolution. 2008;23(9):473–8.
2. Forbes AA, Devine SN, Hippee AC, Tvedte ES, Ward AKG, Widmayer HA, et al. Revisiting the particular role of host shifts in initiating insect speciation. Evolution. 2017 May;71(5):1126–37.
3. Rundle HD, Nosil P. Ecological speciation. Ecology Letters. 2005;8(3):336–52.
4. Howard DJ, Berlocher SH. Endless Forms: Species and Speciation. Oxford University Press; 1998. 486 p.
5. O'Grady PM, Markow TA. Rapid morphological, behavioral, and ecological evolution in *Drosophila*: comparisons between the endemic Hawaiian *Drosophila* and the cactophilic repleta species group. In: Singh RS, Xu J, Kulathinal RJ, editors. Rapidly Evolving Genes and Genetic Systems [Internet]. Oxford University Press; 2012 [cited 2024 Jul 1]. p. 0. Available from: https://doi.org/10.1093/acprof:oso/9780199642274.003.0018
6. Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila* repleta species group. Molecular Phylogenetics and Evolution. 2012;64(3):533–44.
7. Stintzing FC, Carle R. Cactus stems (Opuntia spp.): A review on their chemistry, technology, and uses. Molecular Nutrition & Food Research. 2005;49(2):175–94.
8. Gomez GA. Transpecific Polymorphisms in an Inversion Linked Esterase Locus in *Drosophila* buzzatii. Molecular Biology and Evolution. 2003;20(3):410–23.
9. Hasson E, Naveira H, Fontdevila A. The breeding sites of the Argentinian species of the *Drosophila* mulleri complex (subgenus *Drosophila*-repleta group). Rev Chil Hist Nat. 1992 ;65:319–26.

10. Newby BD, Etges WJ. Host preference among populations of *Drosophila mojavensis* (Diptera: Drosophilidae) that use different host cacti. Journal of Insect Behavior. 1998;11(5):691–712.
11. Rane RV, Clarke DF, Pearce SL, Zhang G, Hoffmann AA, Oakeshott JG. Detoxification Genes Differ Between Cactus-, Fruit-, and Flower-Feeding *Drosophila*. Journal of Heredity. 2019;110(1):80–91.
12. Markow TA. Host use and host shifts in *Drosophila*. Current Opinion in Insect Science. 2019;31:139–45.
13. Montell C. A taste of the *Drosophila* gustatory receptors. Current Opinion in Neurobiology. 2009;19(4):345–53.
14. Ni L. The Structure and Function of Ionotropic Receptors in *Drosophila*. Front Mol Neurosci. 2021;13:638839.
15. Chen Y, Amrein H. Ionotropic receptors mediate *Drosophila* oviposition preference through sour gustatory receptor neurons. Current Biology. 2017;27(18):2741-2750.e4.
16. Gomez-Diaz C, Martin F, Garcia-Fernandez JM, Alcorta E. The Two Main Olfactory Receptor Families in *Drosophila*, ORs and IRs: A Comparative Approach. Front Cell Neurosci. 2018;12:253.
17. Benowitz KM, Coleman JM, Allan CW, Matzkin LM. Contributions of cis- and trans-regulatory evolution to transcriptomic divergence across populations in the *Drosophila mojavensis* larval brain. Genome Biology and Evolution. 2020;12(8):1407–18.
18. Fedoroff NV. Transposable elements, epigenetics, and genome evolution. Science. 2012;338(6108):758–67.
19. Capy P. Taming, domestication and exaptation: Trajectories of transposable elements in genomes. Cells. 2021;10(12):3590.
20. Drongitis D, Aniello F, Fucci L, Donizetti A. Roles of transposable elements in the different layers of gene expression regulation. IJMS. 2019;20(22):5755.
21. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive Increases in the Resistance of *Drosophila* to Viral Infection through a transposon insertion followed by a duplication. PLoS Genet. 2011;7(10):e1002337.
22. Lipatov M, Lenkov K, Petrov DA, Bergman CM. Paucity of chimeric gene-transposable element transcripts in the *Drosophila* melanogaster genome. BMC Biol. 2005;3(1):24.
23. Oliveira DS, Fablet M, Larue A, Vallier A, Carareto CMA, Rebollo R, et al. ChimeraTE: a pipeline to detect chimeric transcripts derived from genes and transposable elements. Nucleic Acids Research. 2023;51(18):9764–84.
24. Matzkin LM. The molecular basis of host adaptation in cactophilic *Drosophila*: Molecular Evolution of a Glutathione S -Transferase Gene (GstD1) in *Drosophila mojavensis*. Genetics. 2008;178(2):1073–83.
25. Pfeiler E, Castrezana S, Reed LK, Markow TA. Genetic, ecological and morphological differences among populations of the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. Journal of Natural History. 2009;43(15–16):923–38.
26. Etges WJ. Genetics of host-cactus response and life-history evolution among cncestral and derived populations of cactophilic *Drosophila mojavensis*. Evolution. 1993;47(3):750–67.
27. Crowley-Gall A, Date P, Han C, Rhodes N, Andolfatto P, Layne JE, et al. Population differences in olfaction accompany host shift in *Drosophila mojavensis*. Proc R Soc B. 2016;283(1837):20161562.
28. Rajpurohit S, Oliveira CC, Etges WJ, Gibbs AG. Functional genomic and phenotypic responses to desiccation in natural populations of a desert drosophilid. Molecular Ecology. 2013;22(10):2698–715.

29. De Panis D, Dopazo H, Bongcam-Rudloff E, Conesa A, Hasson E. Transcriptional responses are oriented towards different components of the rearing environment in two *Drosophila* sibling species. BMC Genomics. 2022;23(1):515.
30. Matzkin L, Markow TA. Transcriptional differentiation across the four subspecies of *Drosophila mojavensis*. In: Speciation: Natural Processes, Genetics and Biodiversity. New York: Nova Scientific Publishers; 2013. p. 119–35.
31. Benowitz KM, Allan CW, Jaworski CC, Sanderson MJ, Diaz F, Chen X, Matzkin LM. Fundamental patterns of structural evolution revealed by chromosome-length genomes of cactophilic *Drosophila*. Genome Biology and Evolution. 2024; 16(9):evae191.
32. Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila* repleta species group. Molecular Phylogenetics and Evolution. 2012;64(3):533–44.
33. Sanchez-Flores A, Peñaloza F, Carpinteyro-Ponce J, Nazario-Yepiz N, Abreu-Goodger C, Machado CA, et al. Genome Evolution in Three Species of Cactophilic *Drosophila*. G3 Genes|Genomes|Genetics. 2016;6(10):3097–105.
34. Smith G, Lohse K, Etges WJ, Ritchie MG. Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*: Phylogeography of *Drosophila mojavensis*. Molecular Ecology. 2012;21(13):3293–307.
35. Guio L, Vieira C, González J. Stress affects the epigenetic marks added by natural transposable element insertions in *Drosophila* melanogaster. Sci Rep. 2018;8(1):12197.
36. Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E, Jeffers S, et al. A Single P450 allele associated with insecticide resistance in *Drosophila*. Science. 2002;297(5590):2253–6.
37. Chang NC, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. Genome Res. 2022;32(7):1408–23.
38. Pritham EJ, Feschotte C. Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proceedings of the National Academy of Sciences. 2007;104(6):1895–900.
39. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. Genome Biology and Evolution. 2014;6(10):2595–610.
40. Jafari S, Alkhori L, Schleiffer A, Brochtrup A, Hummel T, Alenius M. Combinatorial activation and repression by seven transcription factors specify *Drosophila* odorant receptor expression. PLoS Biol. 2012;10(3):e1001280.
41. Fogleman JC, Danielson PB. Chemical Interactions in the Cactus-Microorganism-*Drosophila* Model System of the Sonoran Desert. American Zoologist. 2001;41(4):877–89.
42. Fogleman JC, Heed WB, Kircher HW. *Drosophila* mettleri and senita cactus alkaloids: Fitness measurements and their ecological significance. Comparative Biochemistry and Physiology Part A: Physiology. 1982;71(3):413–7.
43. Markow TA. Host use and host shifts in *Drosophila*. Current Opinion in Insect Science. 2019;31:139–45.
44. McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila* sechellia. Proceedings of the National Academy of Sciences. 2007;104(12):4996–5001.
45. McBride CS, Arguello JR. Five *Drosophila* Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily. Genetics. 2007;177(3):1395–416.

46. Allan CW, Matzkin LM. Genomic analysis of the four ecologically distinct cactus host populations of *Drosophila mojavensis*. BMC Genomics. 2019;20(1):732.
47. Guillén Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, et al. Genomics of Ecological Adaptation in Cactophilic *Drosophila*. Genome Biology and Evolution. 2015;7(1):349–66.
48. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. Briefings in Bioinformatics. 2011;12(5):379–91.
49. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution. 2015;32(3):820–32.
50. Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, et al. Relaxed selection in the wild. Trends in Ecology & Evolution. 2009;24(9):487–96.
51. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution. 1994;11(5):725–36.
52. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41(5):563–71.
53. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 2013;9(4):e1003470.
54. Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. Convergent Evolution of Endometrial Prolactin Expression in Primates, Mice, and Elephants Through the Independent Recruitment of Transposable Elements. Molecular Biology and Evolution. 2012;29(1):239–47.
55. Ye M, Goudot C, Hoyler T, Lemoine B, Amigorena S, Zueva E. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. Proc Natl Acad Sci USA. 2020;117(14):7905–16.
56. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013;45(7):836–41.
57. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, et al. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell. 2012;148(1–2):335–48.
58. Wang J, Lunyak VV, Jordan IK. Genome-wide prediction and analysis of human chromatin boundary elements. Nucleic Acids Research. 2012;40(2):511–29.
59. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24(12):1963–76.
60. Gill RA, Scossa F, King GJ, Golicz AA, Tong C, Snowdon RJ, et al. On the Role of Transposable Elements in the Regulation of Gene Expression and Subgenomic Interactions in Crop Genomes. Critical Reviews in Plant Sciences. 2021;40(2):157–89.
61. Hof AE van't, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016;534(7605):102–5.
62. Mateo L, Ullastres A, González J. A Transposable Element Insertion Confers Xenobiotic Resistance in *Drosophila*. Feschotte C, editor. PLoS Genet. 2014;10(8):e1004560.
63. Dekker T, Ibba I, Siju KP, Stensmyr MC, Hansson BS. Olfactory Shifts Parallel Superspecialism for Toxic Fruit in *Drosophila* melanogaster Sibling, D. sechellia. Current Biology. 2006;16(1):101–9.

64. Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. Odorant-Binding Proteins OBP57d and OBP57e Affect Taste Perception and Host-Plant Preference in *Drosophila* sechellia. PLoS Biol. 2007;5(5):e118.
65. McKenzie SK, Kronauer DJC. The genomic architecture and molecular evolution of ant odorant receptors. Genome Res. 2018;28(11):1757–65.
66. Cultrone A, Domínguez YR, Drevet C, Scazzocchio C, Fernández-Martín R. The tightly regulated promoter of the xanA gene of Aspergillus nidulans is included in a helitron: A fungal helitron includes a regulated promoter. Molecular Microbiology. 2007;63(6):1577–87.
67. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage. Genome Biology and Evolution. 2014;6(10):2595–610.
68. Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK. A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol. 2005;57(1):115–27.
69. Huang K, Li CF, Wu J, Wei JH, Zou Y, Han MJ, et al. Enhancer activity of Helitron in sericin-1 gene promoter from Bombyx mori: Enhancer activity of Helitrons in sericin-1. Insect Science. 2016;23(3):396–405.
70. Batista RA, Moreno-Romero J, Qiu Y, van Boven J, Santos-González J, Figueiredo DD, et al. The MADS-box transcription factor PHERES1 controls imprinting in the endosperm by binding to domesticated transposons. eLife. 2019;8:e50541.
71. Carareto CMA, Hernandez EH, Vieira C. Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling *Drosophila* species. Gene. 2014;537(1):93–9.
72. Merenciano M, Aguilera L, González J. Two-step CRISPR-Cas9 protocol for transposable element deletion in D. melanogaster natural populations. STAR Protocols. 2023;4(3):102501.
73. Wright CR, William S. Characterization of volatiles of necrotic Stenocereus thurberi and Opuntia littoralis and toxicity and olfactory preference of *Drosophila* melanogaster, D. mojavensis wrigleyi, and D. mojavensis sonorensis to necrotic cactus volatiles. Natural Product Communications. 2014;9(6):1934578X1400900833.
74. Hickner PV, Rivaldi CL, Johnson CM, Siddappaji M, Raster GJ, Syed Z. The making of a pest: Insights from the evolution of chemosensory receptor families in a pestiferous and invasive fly, *Drosophila* suzukii. BMC Genomics. 2016;17(1):648.
75. Mansourian S, Stensmyr MC. The chemical ecology of the fly. Current Opinion in Neurobiology. 2015;34:95–102.
76. Crowley-Gall A, Date P, Han C, Rhodes N, Andolfatto P, Layne JE, et al. Population differences in olfaction accompany host shift in *Drosophila mojavensis*. Proc R Soc B. 2016;283(1837):20161562.
77. De Panis D, Dopazo H, Bongcam-Rudloff E, Conesa A, Hasson E. Transcriptional responses are oriented towards different components of the rearing environment in two *Drosophila* sibling species. BMC Genomics. 2022;23(1):515.
78. Ruiz A, Heed WB. Host-Plant specificity in the cactophilic *Drosophila* mulleri species complex. The Journal of Animal Ecology. 1988;57(1):237.
79. Mateus RP, Nazario-Yepiz NO, Ibarra-Laclette E, Ramirez Loustalot-Laclette M, Markow TA. Developmental and transcriptomal responses to seasonal dietary shifts in the cactophilic *Drosophila mojavensis* of North America. Journal of Heredity. 2019;110(1):58–67.

80. Matzkin LM. Population transcriptomics of cactus host shifts in *Drosophila mojavensis*: Population Transcriptomics of D. mojavensis. Molecular Ecology. 2012;21(10):2428–39.
81. Feyereisen R. Insect CYP Genes and P450 Enzymes. In: Insect molecular biology and biochemistry. Academic Press, 2012. p. 236-316.
82. Tijet N, Helvig C, Feyereisen R. The cytochrome P450 gene superfamily in *Drosophila* melanogaster: Annotation, intron-exon organization and phylogeny. Gene. 2001;262(1–2):189–98.
83. Coronado-Zamora M, González J. Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across *Drosophila* natural strains. Genome Res. 2023;33(9):1541–53.
84. Bartolomé C, Maside X, Charlesworth B. On the Abundance and Distribution of Transposable Elements in the Genome of *Drosophila* melanogaster. Molecular Biology and Evolution. 2002;19(6):926–37.
85. Gao C, Zhou G, Ma C, Zhai W, Zhang T, Liu Z, et al. Helitron-like transposons contributed to the mating system transition from out-crossing to self-fertilizing in polyploid Brassica napus. Sci Rep. 2016;6(1):33785.
86. Garrigues JM, Tsu BV, Daugherty MD, Pasquinelli AE. Diversification of the Caenorhabditis heat shock response by Helitron transposable elements. eLife. 2019;8:e51139.
87. Guo C, Spinelli M, Liu M, Li QQ, Liang C. A Genome-wide Study of "Non-3UTR" Polyadenylation Sites in Arabidopsis thaliana. Sci Rep. 2016;6(1):28060.
88. Lee JY, Ji Z, Tian B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. Nucleic Acids Research. 2008;36(17):5581–90.
89. Fonseca PM, Moura RD, Wallau GL, Loreto ELS. The mobilome of *Drosophila* incompta, a flower-breeding species: comparison of transposable element landscapes among generalist and specialist flies. Chromosome Res. 2019;27(3):203–19.
90. Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, et al. The Worldwide Invasion of *Drosophila* suzukii Is Accompanied by a Large Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions. Molecular Biology and Evolution. 2021;38(10):4252–67.
91. Marin P, Jaquet A, Picarle J, Fablet M, Merel V, Delignette-Muller ML, et al. Phenotypic and Transcriptomic Responses to Stress Differ According to Population Geography in an Invasive Species. Genome Biology and Evolution. 2021;13(9):evab208.
92. Harigaya Y, Parker R. No-go decay: a quality control mechanism for RNA in translation. WIREs RNA. 2010;1(1):132–41.
93. Vasudevan S, Peltz SW, Wilusz CJ. Non-stop decay—a new mRNA surveillance pathway. BioEssays. 2002;24(9):785–8.
94. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res. 2016;44(4):1483–95.
95. Whisson SC, Vetukuri RR, Avrova AO, Dixelius C. Can silencing of transposons contribute to variation in effector gene expression in Phytophthora infestans? Mobile Genetic Elements. 2012;2(2):110–4.
96. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, et al. Genome Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. Science. 2010;330(6010):1540–3.
97. Zouros E, d'Entremont CJ. Sexual Isolation Among Populations of *Drosophila mojavensis*: Response to Pressure from a Related Species. Evolution. 1980;34(3):421–30.
98. Markow TA. Sexual Isolation among Populations of *Drosophila mojavensis*. Evolution. 1991;45(6):1525–9.

99. Howard RW, Blomquist GJ. Ecological, Behavioral, and biochemical aspects of insect hydrocarbons. Annual Review of Entomology. 2005;50(Volume 50, 2005):371–93.
100. Stennett MD, Etges WJ. Premating Isolation Is Determined by Larval Rearing Substrates in Cactophilic *Drosophila mojavensis* III. Epicuticular Hydrocarbon Variation Is Determined by Use of Different Host Plants in *Drosophila mojavensis* and *Drosophila arizonae*. J Chem Ecol. 1997;23(12):2803–24.
101. Havens JA, Etges WJ. Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IX. Host plant and population specific epicuticular hydrocarbon expression influences mate choice and sexual selection. Journal of Evolutionary Biology. 2013;26(3):562–76.
102. Romero-Soriano V, Burlet N, Vela D, Fontdevila A, Vieira C, García Guerreiro MP. *Drosophila* Females undergo genome expansion after interspecific hybridization. Genome Biol Evol. 2016;8(3):556–61.
103. Bodelón A, Fablet M, Veber P, Vieira C, García Guerreiro MP. High Stability of the Epigenome in *Drosophila* Interspecific Hybrids. Genome Biology and Evolution. 2022;14(2):evac024.
104. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
105. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
106. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio] [Internet]. 2013 May 26; Available from: http://arxiv.org/abs/1303.3997
107. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958.
108. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
109. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007;23(8):1026–8.
110. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20(1):224.
111. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
112. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37(12):1639–43.
113. Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, et al. Highly contiguous assemblies of 101 drosophilid genomes. eLife. 2021;10(e66405):33.
114. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34(8):1-9.
115. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
116. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
117. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25(17):3389–402.
118. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37(12):1639–43.

119. Vanderlinde T, Dupim EG, Nazario-Yepiz NO, Carvalho AB. An Improved Genome Assembly for *Drosophila* navojoa, the basal species in the mojavensis cluster. Journal of Heredity. 2019;110(1):118–23.
120. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36(7):2253–5.
121. Banho CA, Mérel V, Oliveira TYK, Cararetto CMA, Vieira C. Comparative transcriptomics between *Drosophila mojavensis* and D. arizonae reveals transgressive gene expression and underexpression of spermatogenesis-related genes in hybrid testes. Sci Rep. 2021;11(1):9844.
122. Moreyra NN, Almeida FC, Allan C, Frankel N, Matzkin LM, Hasson E. Phylogenomics provides insights into the evolution of cactophily and host plant shifts in *Drosophila*. Molecular Phylogenetics and Evolution. 2023;178:107653.
123. Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. Zenodo [Internet]. 2023; Available from: https://zenodo.org/records/8178877
124. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
125. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. 2006;34(suppl_2):W609–12.
126. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020;37(5):1530–4.
127. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 2007;24(8):1586–91.
128. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5—A customizable platform for evolutionary hypothesis testing using phylogenies. Molecular Biology and Evolution. 2020;37(1):295–9.
129. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
130. Baril, TJ, Galbraith JD, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. Molecular Biology and Evolution. 2024;41(4):msae068.
131. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020 Apr 28;117(17):9451–7.
132. Bao Z, Eddy SR. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res. 2002;12(8):1269–76.
133. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351–8.
134. Platt RN, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biol Evol. 2016;8(2):403–10.
135. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013;30(4):772–80.
136. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
137. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 16(6):276–7.
138. SMIT, Arian FA. RepeatMasker Open 3.0. 2004;

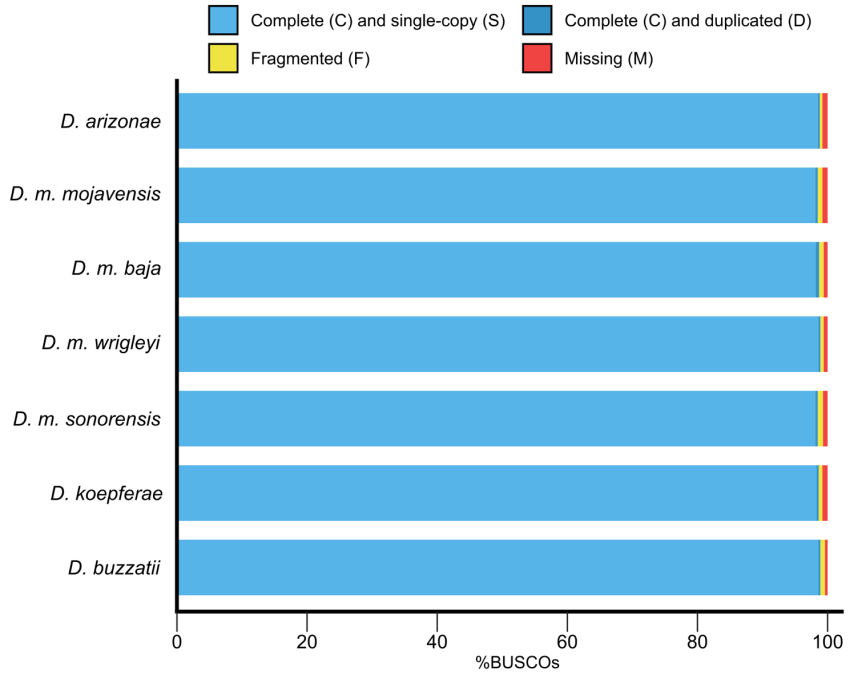
139. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44(D1):D81–9.
140. Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Hancock J, editor. Bioinformatics. 2019;35(6):1051–2.
141. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82.
142. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999;27(2):573–80.
143. Flutre T, Permal E, Quesneville H. Transposable Element Annotation in Completely Sequenced Eukaryote Genomes. In: Grandbastien MA, Casacuberta JM, editors. Plant Transposable Elements. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 [cited 2023 Feb 27]. p. 17–39. (Topics in Current Genetics; vol. 24). Available from: https://link.springer.com/10.1007/978-3-642-31842-9_2
144. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2015;btv562.
145. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
146. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8.
147. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Research. 2015;43(W1):W39–49.
148. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2019;gkz1001.
149. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
150. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
151. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

Supplementary material

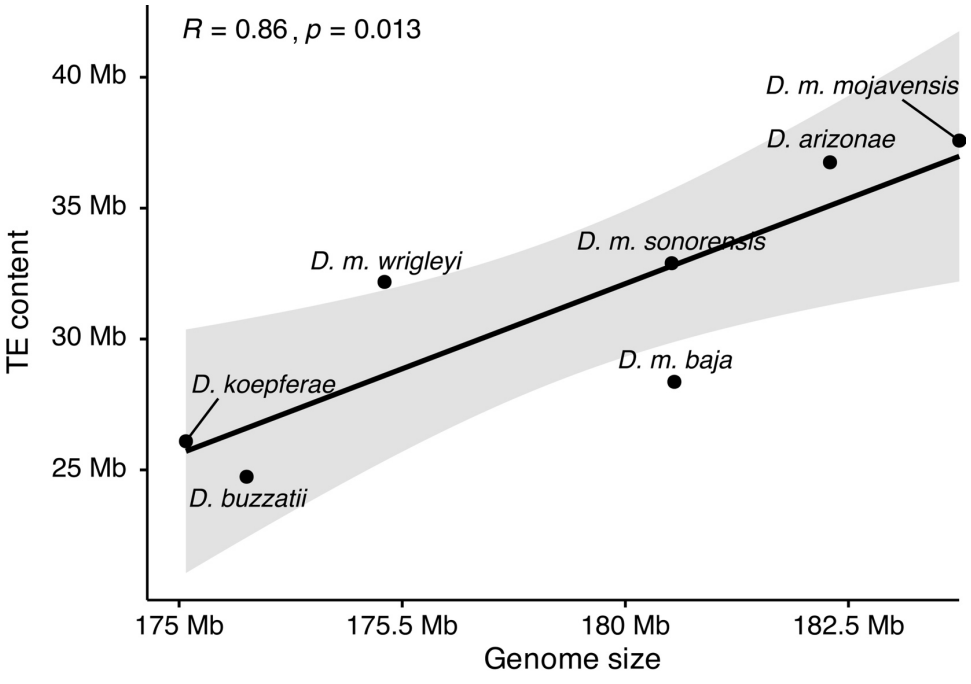
Supplementary Tables

The supplementary tables S1, S2, S3, S4, S5, S6, S7, S8, S9, S10 and S11 can be accessed in the publicly available link from Nucleic Acids Research journal: https://www.biorxiv.org/content/10.1101/2024.03.27.587021v2.supplementary-material

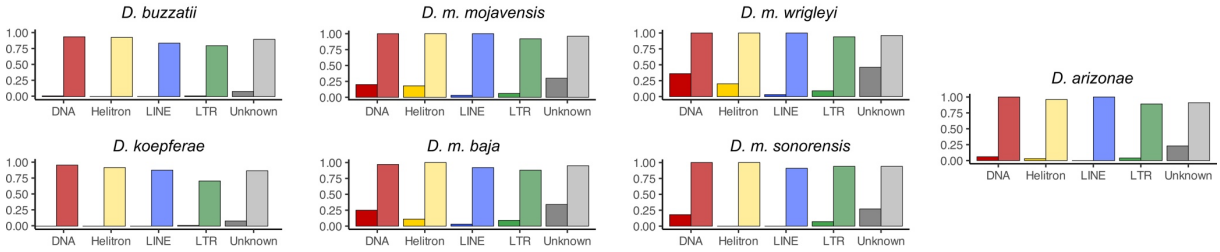
Supplementary Figures



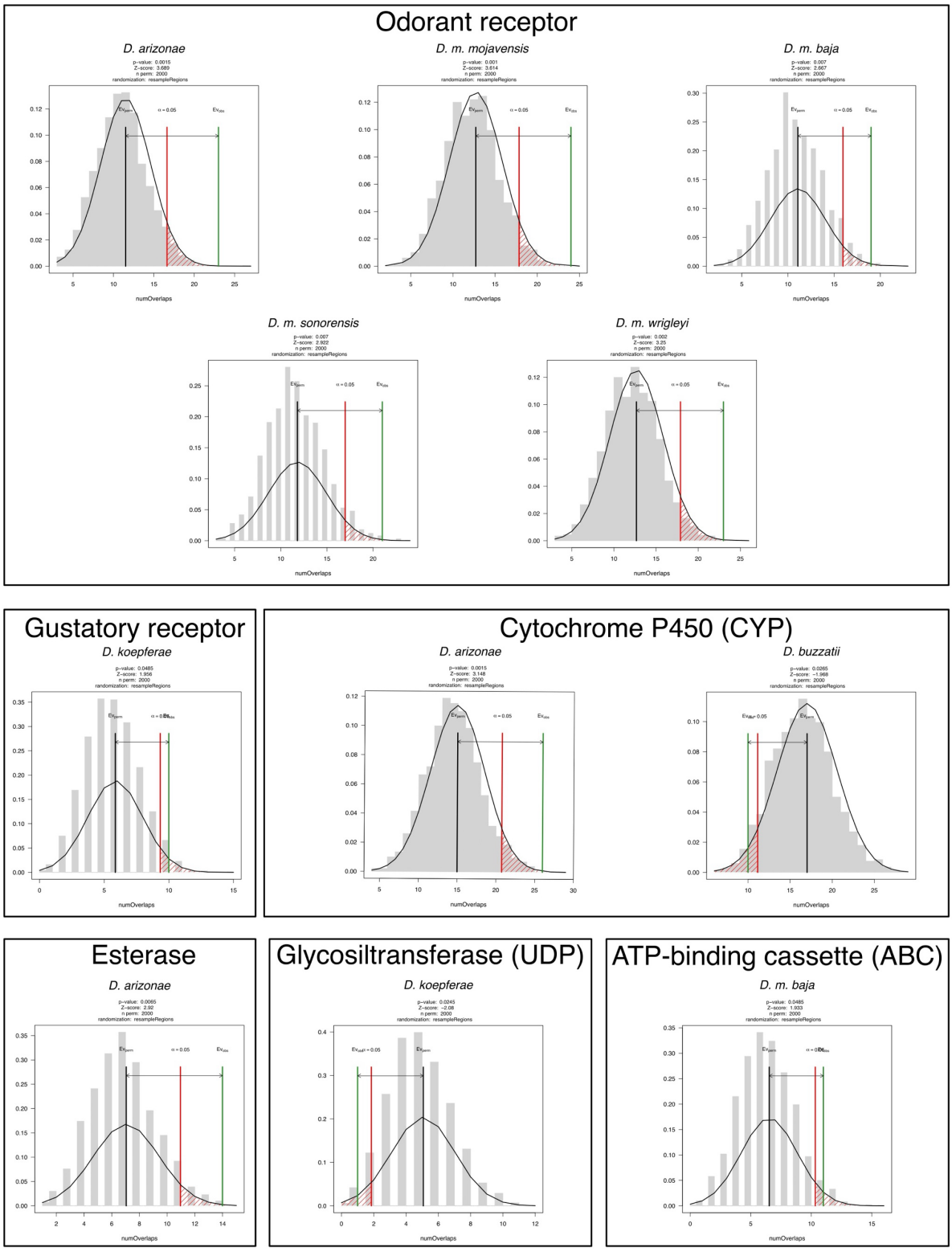
Sup. Figure 1: BUSCO score for the Nanopore genome assemblies.



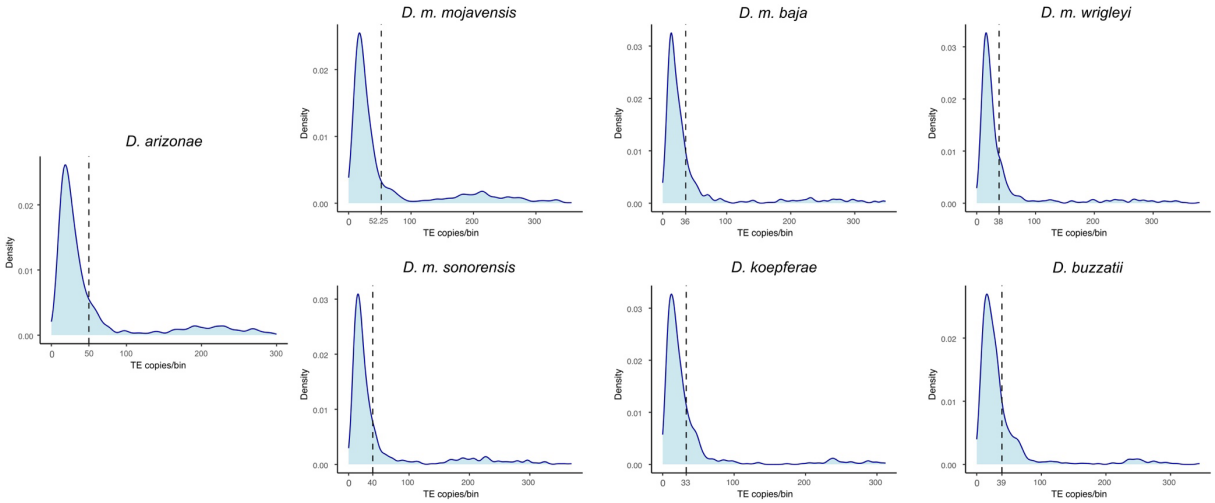
Sup. Figure 2: Correlation between TE content and genome size in the seven nanopore genomes of cactophilic *Drosophila* species.



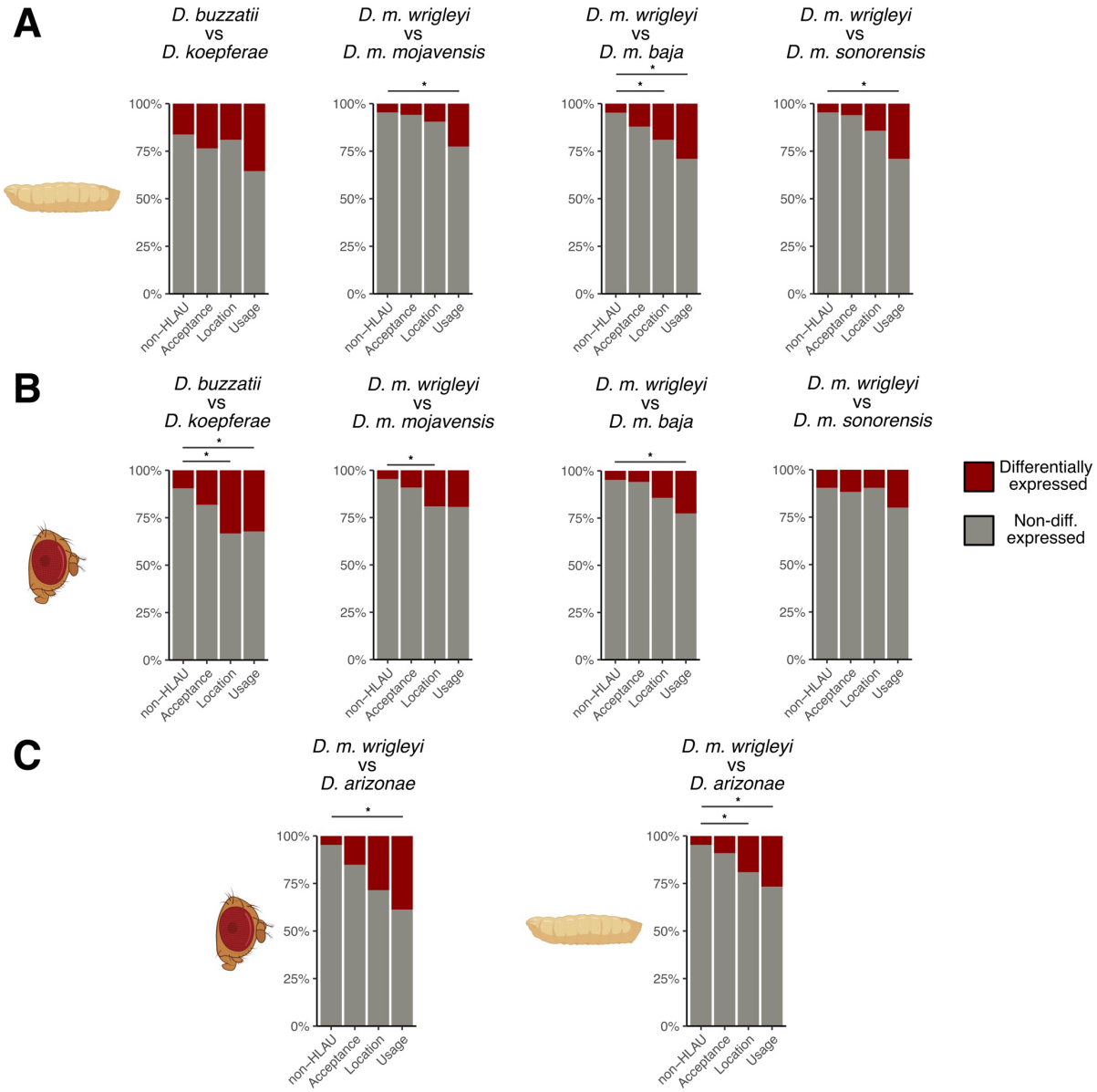
Sup. Figure 3: Intragenic TEs. For each TE order, left bar represents the proportion of TE families in this TE order with copies inserted preferentially within genes, and the right bar represents the proportion of TE families without preference for intragenic regions.



Sup. Figure 4: Histograms for permutation tests. Histograms demonstrate the frequency of HLAU genes with TEs at the 2kb upstream region. Green bar represents the observed number of genes and the respective distance from the mean is shown with a horizontal arrow.



Sup. Figure 5: Frequency of TE copies over 200kb genomic bins. The vertical blue dotted-line represents the number of insertions in the 3rd quartile of the distribution. The distribution on the right side of this line represents TE-rich scaffolds.



Sup. Figure 6: Fisher's exact test between the proportional of differentially expressed genes from location, acceptance and usage gene families, compared to non-HLAU genes (Ser/Thr Kinases and zinc-finger proteins). Bars connected by a line with star on the top represent p-value < 0.05. The pairwise analysis was performed between (A-B) species with preference for *Opuntia sp*. (*D. buzzatii* and *D. m. wrigleyi*) and columnar cacti dwellers (*D. koepferae*, *D. m. mojavensis*, *D. m. baja*, and *D. m. sonorensis*). **A**) Fisher's exact test calculated from differentially expressed genes in larvae. **B**) Fisher's exact test calculated from differentially expressed genes in head. **C**) Fisher's exact test calculated from differentially expressed genes from *D. m. wrigleyi* with preference for *Opuntia sp.* and *D. arizonae* as a non-preferential cactus.

DISCUSSION

In recent decades, RNA-seq has significantly advanced our understanding of transcriptome plasticity, which can lead to phenotypic variation between related species or individuals within a population (MARGUERAT; BÄHLER, 2010). Among the many factors influencing gene expression and the creation of novel transcript isoforms, TEs have been identified as key contributors to transcriptome plasticity. TEs play a role in gene networks by incorporating regulatory sequences or protein domains into gene transcripts. Identifying chimeric transcripts is a crucial step in understanding transcriptome plasticity, as these transcripts may be produced under ectopic conditions such as cancer, oxidative stress, or heat shock (HORVÁTH; MERENCIANO; GONZÁLEZ, 2017; OLIVEIRA et al., 2021), potentially leading to both harmful and beneficial outcomes (NICOLAU; PICAULT; MOISSIARD, 2021). Thus, investigating the extent of chimeric transcripts across individual, cell, or strain transcriptomes is vital for exploring potential exaptation/domestication events or disruptions that lead to gene loss of function.

Chimeric transcripts have been increasingly identified through various methods using RNA-seq data (BABAIAN et al., 2019; CORONADO-ZAMORA; GONZÁLEZ, 2023; MERCURI et al., 2024; PINSON et al., 2018; TREIBER; WADDELL, 2020). However, none of these approaches has been able to predict chimeras arising from TEs absent in reference genomes. In my thesis, we developed ChimeraTE, a pipeline designed to identify chimeric transcripts derived from TEs not present in the reference genome. Mode 1 of ChimeraTE is a genome-guided approach, suited for cases where the user is focused on chimeras from TEs in the reference genome or when a high-quality genome assembly for each individual, cell, or strain is available. Mode 2, in contrast, is a genome-blind approach that can predict chimeric transcripts without a reference genome, although it may miss chimeras if the TE is shorter than the read length. This novel bioinformatics pipeline offers a powerful tool for studying the effects of TE mobilization events across populations or experimental conditions. It provides valuable insights for researchers across fields such as cancer research, population transcriptomics, and adaptation studies. The implementation of ChimeraTE to support future discoveries regarding the evolutionary role of TEs and their influence on host transcriptomes. Since the publication, pipeline has nine citations and 4,920 views.

We used ChimeraTE with ovarian RNA-seq data from four *D. melanogaster* wild-type strains with genome assemblies. We found that approximately 1.12% of all genes in the ovaries

produce chimeric transcripts, aligning with findings from earlier studies using expressed sequence tags and RNA-seq in midbrain tissue (LIPATOV et al., 2005; TREIBER; WADDELL, 2020). Notably, our results showed that 50% of all TE-exonized transcripts, where TEs are embedded, originated from *roo* elements, specifically from a small region between tandem repeats in the 5' UTR and the start of the *roo* ORF. These findings suggest that these *roo* insertions may have neutral or advantageous effects, as they are maintained in gene transcripts. However, we cannot yet confirm these *roo*-exonized transcripts as cases of exaptation or domestication, due to a lack of evidence for the functional roles of these chimeras, and impact on the host fitness. Further investigation is required, especially since chimeric transcripts detected by ChimeraTE may be degraded by RNA surveillance pathways, such as nonsense-mediated decay (HUG; LONGMAN; CÁCERES, 2016), no-go decay (HARIGAYA; PARKER, 2010), and non-stop decay (VASUDEVAN; PELTZ; WILUSZ, 2002).

Following the development of ChimeraTE as a tool to identify gene-TE interactions from the transcriptome perspective, we investigated whether TEs could contribute to HS in cactophilic *Drosophila* species. Using *Drosophila* species with different cacti host preferences, we performed genomic and transcriptomic analyses to investigate how TEs might have influenced the divergence of species, through the contribution of genetic variability on genes associated with HS. We firstly assessed molecular evolution analysis on the nine gene families to explore whether they have underwent positive or relaxed selection given their critical roles in recognizing odors, metabolizing nutrients, and detoxifying host tissues (MARKOW, 2019). While a branch-site test showed no significant positive selection in these genes, the result from RELAX model (WERTHEIM et al., 2015) revealed relaxed selection in the genes *Obp56d* and *Cyp304a1*. This finding suggests that, after initial host adaptation, selective pressures diminished in certain HLAU genes, reflecting ongoing evolutionary transitions as species adapt to new hosts. This relaxation of selection might indicate a shift from strict functional constraints to more flexible, species-specific roles. It highlights the potential for evolutionary change in sensory pathways and detoxification systems, which are essential for adapting to new environments (GUILLÉN et al., 2015).

The genomic analysis demonstrated an enrichment of TEs in the promoter regions of HLAU genes, particularly in ORs, suggesting a significant role of TEs in regulatory evolution. OR genes were found to have the highest high frequency of TE insertions in the five genomes of the *mojavensis* cluster, demonstrating that such TE enrichment is likely to have happened in the common ancestor between *D. mojavensis* and *D. arizonae*. Interestingly, we identified ~300

TE families in the seven genomes under study, but *Helitron* were highly enriched in these promoter regions, accounting for ~60% of TE insertions with TFBSs associated with OR transcription. We raised the hypothesis that *Helitrons* may have contributed with TFBS to these regions, which could lead to altered gene expression in response to host environments. The enrichment of *Helitron* insertions in promoter regions across the five *Drosophila* genomes suggests that these elements might have contributed TFBS to specific gene networks. This can lead to altered expression patterns, potentially facilitating rapid adaptation to new host environments, particularly through the modulation of genes sensory perception. The ability of *Helitrons* to act as gene regulators emphasizes their potential evolutionary importance in shaping genome plasticity and adaptability in changing environments. In addition, the consistent presence of TE insertions in OR gene promoters across species might reflect a common evolutionary adaptation that facilitates host recognition and olfactory sensitivity.

Gene expression in head tissue showed significant divergence among cactophilic subspecies, particularly in detoxification-related gene families such as CYPs. Our study found that *D. m. wrigleyi*, a subspecies preferring *Opuntia sp.* cacti, had the highest number of upregulated CYP genes. This finding suggests that the divergence in host preference among subspecies is reflected in differential expression of detoxification genes, which help species cope with toxic compounds in their specific cacti hosts (GUILLÉN et al., 2015). In addition, several ORs and other chemosensory genes exhibited subspecies-specific expression patterns. These variations in the sensory transcriptome are likely a key factor in the evolution of host preference, as differential sensitivity to odors and other environmental cues can drive behavioral and ecological divergence (DEKKER et al., 2006; MATSUO et al., 2007). In the larvae tissue, the expression of detoxification genes underscores the role of gene regulation in host adaptation. Our study revealed higher levels of CYP gene expression in *D. m. wrigleyi* larvae compared to other subspecies, reinforcing the idea that detoxification pathways are critical for survival on toxic cacti. Moreover, the divergence in gene expression between larvae from the *buzzatii* and *mojavensis* clusters suggests that host adaptation is a complex, ongoing process shaped by gene regulation at different developmental stages. The high number of differentially expressed genes in larvae highlights how host preference influences both developmental pathways and detoxification mechanisms.

The identification of gene-TE chimeric transcripts in the seven cactophilic transcriptomes from the species under study was assessed with ChimeraTE, under the hypothesis that TEs contribute to transcriptome variability. We observed that TEs play a

significant role in generating chimeras from ~1.31% of genes, similarly with the proportion observed between *D. melanogaster* wild-type strains (Oliveira et al. 2023; Coronado-Zamora and González 2023). Studies have shown that TEs can splice into coding or regulatory regions of genes, forming chimeras that may have novel functions or regulatory impacts. For instance, in *Drosophila*, TE-mediated splicing events have been linked to tissue-specific expression and functional diversity (CORONADO-ZAMORA; GONZÁLEZ, 2023). *Helitrons* were the most common TEs associated with chimeric transcripts, particularly in initiating transcription. Since *Helitrons* are known to increase the mutation rate and introduce new regulatory elements that can be co-opted for host-specific adaptations (BATISTA et al., 2019; LYNCH et al., 2015), these unprecedent findings in *Drosophila* reinforce the role of *Helitrons* as *cis*-regulatory elements, contributing to gene regulation across multiple species.

We observed higher production of chimeric transcripts from *D. arizonae*, which is the only species under study without preference between *Opuntia sp.* and columnar cacti. Despite being observed in a single species, this result suggests that TEs may produce higher transcriptome diversity in species with broader ecological niches, and lower in species with more restricted niche. The higher TE content introduces more variation into the genome, either through direct insertions or by creating opportunities for gene duplication, exon shuffling, and other forms of genomic restructuring (OLIVER; GREENE, 2009). Such innovations may provide the genetic raw material for organisms to adapt to different environmental pressures, potentially facilitating occupation of a broader range of ecological niches. From the genomic perspective, preliminary studies have shown the relationship between TEs and colonization of new habitats. For instance, in the invasive species *D. suzukii*, a large increase in the TE content has been observed upon worldwide invasion (MÉREL et al., 2021). On the other hand, there is no clear changes on TE content and evolutionary dynamics of TEs when comparing generalist and specialist *Drosophila* species (FONSECA et al., 2019). Another study with flowering-breeding *Drosophila* species with different niche breadth have shown no correlation between TEs' diversity and species niche breadth (CARVALHO et al., 2023). Finally, another recent study conducted with 15 plant species from the genus *Scalesia* (Ateraceae) demonstrated mild to no variation in TE accumulation in the genomes upon shifts in ecological niche (CERCA et al., 2024). Here we demonstrate that TEs are not likely to be associated with ecological niche from the genomic perspective, through mobilization events, but from the transcriptomic perspective through transcription interaction with genes, generating chimeras. We observed not only more overall chimeras in both head and larvae tissue in *D. arizonae,* but also more

chimeras from genes associated with HS, which may have a directly relationship with evolution of host preference and niche breadth.

Finally, our study provides evidence that TEs contribute to the divergence of HLAU genes between species with different host preference. The enrichment at regulatory regions of HLAU genes, and chimeric transcripts derived from different HLAU genes between species suggest that TEs contribute to HS evolution. Since divergence on these genes may have direct effect on host preference, we claim a possibility of reproductive isolation evolution driven by TEs through divergence in host preference. This process could contribute to the development of reproductive barriers and ecological speciation, as observed between *D. m. sonorensis* and *D. m. baja*, where host shifts influences mate recognition (HAVENS; ETGES, 2013; HOWARD; BLOMQUIST, 2005; MARKOW, 1991; ZOUROS; D'ENTREMONT, 1980). The potential for TEs to drive reproductive isolation through host adaptation highlights their role in the broader context of evolutionary biology, particularly in how species adapt to changing environments and new ecological niches. Overall, the results from this study provide evidence for the role of TEs in generating transcriptomic diversity and potentially facilitating host adaptation. Although our quantitative analysis regarding chimeric transcripts revealed an interesting pattern, the functional significance must be experimentally tested. Future research using RNAi and/or CRISPR methods could help determine their phenotypic impact and role in host shift.

CONCLUSIONS

This work allowed us to reach the following conclusions:

1. *Roo* elements are the main source of TE-exonized chimeras, highlighting their putative role on generating genetic diversity between populations. Half of the TE-exonized transcripts are derived from the *roo*, from a specific region at the beginning of its ORF.
2. Approximately 23% of all chimeras found in each strain analyzed in this thesis were derived from TE insertions that are absent from the reference genome. This result suggests that TEs can often disrupt genes and interact with host genes during transcription, incorporating themselves into the mRNA.
3. The high polymorphism of gene-TE chimeras observed between *D. melanogaster* wild-type strains indicates that they are a relevant source of genetic diversity in this species.
4. TEs are likely to contribute to the regulatory evolution of odorant receptor (OR) genes in *D. arizonae* and four *D. mojavensis*. *Helitron* is the most frequent order carrying TFBSs for OR transcription in these species, demonstrating its possible regulatory role.
5. TEs generate genetic diversity through gene-TE chimeric transcripts in the seven cactophilic *Drosophila* genomes analyzed in this thesis.
6. Although rare, the occurrence of the same chimeric transcript between species suggests cases of domestication/exaptation of TEs from a phylogenetic perspective.

REFERENCES

AGRAWAL, A.; EASTMAN, Q. M.; SCHATZ, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. **Nature**, v. 394, n. 6695, p. 744–751, 1998.

ALMEIDA, L. M. et al. Report of a chimeric origin of transposable elements in a bovine-coding gene. **Genetics and Molecular Research**, v. 7, n. 1, p. 107–116, 2008.

ANNALA, M. J. et al. Fusion genes and their discovery using high throughput sequencing. **Cancer Letters**, v. 340, n. 2, p. 192–200, 2013.

ARTIERI, C. G.; SINGH, R. S. Molecular evidence for increased regulatory conservation during metamorphosis, and against deleterious cascading effects of hybrid breakdown in *Drosophila*. **BMC Biology**, v. 8, n. 1, p. 26, 2010.

AYALA, F. J.; POWELL, J. R.; DOBZHANSKY, T. Polymorphisms in Continental and Island Populations of *Drosophila willistoni*. **Proceedings of the National Academy of Sciences**, v. 68, n. 10, p. 2480–2483, 1971.

AYARPADIKANNAN, S. et al. Transposable element-driven transcript diversification and its relevance to genetic disorders. **Gene**, v. 558, n. 2, p. 187–194, 2015.

BABAIAN, A. et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. **Oncogene**, v. 35, n. 19, p. 2542–2546, 2016.

BABAIAN, A. et al. LIONS: Analysis Suite for Detecting and Quantifying Transposable Element Initiated Transcription from RNA-seq. **Bioinformatics**, v. 35, n. 19, p. 3839–3841, 2019.

BABAIAN, A.; MAGER, D. L. Endogenous retroviral promoter exaptation in human cancer. **Mobile DNA**, v. 7, n. 1, p. 24, 2016.

BAKOULIS, S. et al. Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. **Nucleic Acids Research**, v. 50, n. 4, p. 2111–2127, 2022.

BARIAMI, V. et al. Gene Amplification, ABC Transporters and Cytochrome P450s: Unraveling the Molecular Basis of Pyrethroid Resistance in the Dengue Vector, *Aedes aegypti*. **PLoS Neglected Tropical Diseases**, v. 6, n. 6, p. e1692, 2012.

BATISTA, R. A. et al. The MADS-box transcription factor PHERES1 controls imprinting in

the endosperm by binding to domesticated transposons. **eLife**, v. 8, p. e50541, 2019.

BENOWITZ, K. M. et al. Contributions of cis- and trans-regulatory Evolution to Transcriptomic Divergence across Populations in the *Drosophila mojavensis* Larval Brain. **Genome Biology and Evolution**, v. 12, n. 8, p. 1407–1418, 2020.

BOURGEOIS, Y.; BOISSINOT, S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. **Genes**, v. 10, n. 6, p. 419, 2019.

BOURQUE, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. **Genome Research**, v. 18, n. 11, p. 1752–1762, 2008.

BOURQUE, G. et al. Ten things you should know about transposable elements. **Genome Biology**, v. 19, n. 1, p. 199, 2018.

CAMPBELL, P. M. et al. Developmental expression and gene/enzyme identifications in the alpha esterase gene cluster of *Drosophila melanogaster*. **Insect Molecular Biology**, v. 12, n. 5, p. 459–471, 2003.

CAO, Y. et al. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. **Genome Research**, v. 29, n. 1, p. 40–52, 2019.

CAPY, P. Taming, Domestication and Exaptation: Trajectories of Transposable Elements in Genomes. **Cells**, v. 10, n. 12, p. 3590, 2021.

CARVALHO, T. L. et al. Horizontal Transposon Transfer and Their Ecological Drivers: The Case of Flower-breeding *Drosophila*. **Genome Biology and Evolution**, v. 15, n. 5, p. evad068, 2023.

CASTILLO-CHAVEZ, C.; LEVIN, S. A.; GOULD, F. Physiological and Behavioral adaptation to varying environments: a mathematical model. **Evolution**, v. 42, n. 5, p. 986–994, 1988.

CATALÁN, A.; HUTTER, S.; PARSCH, J. Population and sex differences in *Drosophila melanogaster* brain gene expression. **BMC Genomics**, v. 13, n. 1, p. 654, 2012.

CERCA, J. et al. Genomic stability in the Galápagos Scalesia adaptive radiation: Consistent transposable element accumulation despite hybridization and ecological niche shifts. **bioRxiv**, 2024. Available at: https://www.biorxiv.org/content/10.1101/2024.09.30.614436v1. Accessed in Oct 18th, 2024.

CHAPMAN, T. et al. Sexual conflict. **Trends in Ecology & Evolution**, v. 18, n. 1, p. 41–47,

2003.

CHEN, Y.; AMREIN, H. Ionotropic Receptors Mediate *Drosophila* Oviposition Preference through Sour Gustatory Receptor Neurons. **Current Biology**, v. 27, n. 18, p. 2741- 2750.e4, 2017.

CHEUNG, S.; MANHAS, S.; MEASDAY, V. Retrotransposon targeting to RNA polymerase III-transcribed genes. **Mobile DNA**, v. 9, n. 1, p. 14, 2018.

CHUNG, H. et al. *Cis* -Regulatory Elements in the *Accord* Retrotransposon Result in Tissue-Specific Expression of the *Drosophila melanogaster* Insecticide Resistance Gene *Cyp6g1*. **Genetics**, v. 175, n. 3, p. 1071–1077, 2007.

CHUNG, H. et al. Characterization of *Drosophila melanogaster* cytochrome P450 genes. **Proceedings of the National Academy of Sciences**, v. 106, n. 14, p. 5731–5736, 2009.

CHUONG, E. B. et al. Endogenous retroviruses function as species-specific enhancer elements in the placenta. **Nature Genetics**, v. 45, n. 3, p. 325–329, 2013.

CHUONG, E. B.; ELDE, N. C.; FESCHOTTE, C. Regulatory activities of transposable elements: from conflicts to benefits. **Nature Reviews Genetics**, v. 18, n. 2, p. 71–86, 2017.

CLAYTON, E. A. et al. An atlas of transposable element-derived alternative splicing in cancer. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 375, n. 1795, p. 20190342, 2020.

CLOUDSLEY-THOMPSON, J. L. Plants and Herbivorous Animals. Em: CLOUDSLEY-THOMPSON, J. L. (Ed.). **Biotic Interactions in Arid Lands**. Berlin, Heidelberg: Springer, 1996. p. 107–126.

COLEMAN, J. M. et al. Behavioral evolution accompanying host shifts in cactophilic *Drosophila* larvae. **Ecology and Evolution**, v. 8, n. 14, p. 6921–6931, 2018.

CORONADO-ZAMORA, M.; GONZÁLEZ, J. Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across *Drosophila* natural strains. **Genome Research**, v. 33, n. 9, p. 1541–1553, 2023.

COYNE, J. A.; ORR, A. H. Speciation: A catalogue and critique of species concepts. **Philosophy of biology: an anthology**, p. 272-92, 2004.

DABORN, P. J. et al. A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. **Science**, v. 297, n. 5590, p. 2253–2256, 2002.

DANILEVSKAYA, O. N. et al. *Drosophila* telomere transposon HeT-A produces a transcript with tightly bound protein. **Proceedings of the National Academy of Sciences of the United States of America**, v. 91, n. 14, p. 6679, 1994.

DEKKER, T. et al. Olfactory Shifts Parallel Superspecialism for Toxic Fruit in *Drosophila melanogaster* Sibling, *D. sechellia*. **Current Biology**, v. 16, n. 1, p. 101–109, 2006.

DOBZHANSKY, T. Genetic nature of species differences. **The American Naturalist**, v. 71, n. 735, p. 404–420, 1937.

DOBZHANSKY, TH. et al. Genetics of Natural Populations. XXXV. A Progress Report on Genetic Changes in Populations of *Drosophila pseudoobscura* in the American Southwest. **Evolution**, v. 18, n. 2, p. 164–176, 1964.

DRÈS, M.; MALLET, J. Host races in plant–feeding insects and their importance in sympatric speciation. **Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences**, v. 357, n. 1420, p. 471–492, 2002.

DURKIN, S. M.; BALLINGER, M. A.; NACHMAN, M. W. Tissue-specific and cis-regulatory changes underlie parallel, adaptive gene expression evolution in house mice. **PLOS Genetics**, v. 20, n. 2, p. e1010892, 2024.

DUS, M. et al. Taste-independent detection of the caloric content of sugar in *Drosophila*. **Proceedings of the National Academy of Sciences**, v. 108, n. 28, p. 11644–11649, 2011.

EHRLICH, M. DNA hypomethylation in cancer cells. **Epigenomics**, v. 1, n. 2, p. 239–259, 2009.

FABRY, M. H. et al. Maternally inherited piRNAs direct transient heterochromatin formation at active transposons during early *Drosophila* embryogenesis. **eLife**, v. 10, p. e68573, 2021.

FARRÉ, D.; ENGEL, P.; ANGULO, A. Novel Role of 3'UTR-Embedded Alu Elements as Facilitators of Processed Pseudogene Genesis and Host Gene Capture by Viral Genomes. **PLOS ONE**, v. 11, n. 12, p. e0169196, 2016.

FAULKNER, G. J. et al. The regulated retrotransposon transcriptome of mammalian cells. **Nature Genetics**, v. 41, n. 5, p. 563–571, 2009.

FONSECA, P. M. et al. The mobilome of *Drosophila incompta*, a flower-breeding species: comparison of transposable element landscapes among generalist and specialist flies. **Chromosome Research**, v. 27, n. 3, p. 203–219, 2019.

FORBES, A. A. et al. Revisiting the particular role of host shifts in initiating insect speciation. **Evolution**, v. 71, n. 5, p. 1126–1137, 2017.

FRAICHARD, S. et al. Modulation of Sex Pheromone Discrimination by a UDP-Glycosyltransferase in *Drosophila melanogaster*. **Genes**, v. 11, n. 3, p. 237, 2020.

FUDGE, J. B. Transposon–gene chimeras move up as cancer epitopes. **Nature Biotechnology**, v. 41, n. 5, p. 616–616, 2023.

FUJITA, M.; TANIMURA, T. *Drosophila* Evaluates and Learns the Nutritional Value of Sugars. **Current Biology**, v. 21, n. 9, p. 751–755, 2011.

GALINDO, K.; SMITH, D. P. A Large Family of Divergent *Drosophila* Odorant-Binding Proteins Expressed in Gustatory and Olfactory Sensilla. **Genetics**, v. 159, n. 3, p. 1059–1072, 2001.

GARRIGUES, J. M. et al. Diversification of the *Caenorhabditis* heat shock response by Helitron transposable elements. **eLife**, v. 8, p. e51139, 2019.

GEORGE, J. A. et al. Genomic organization of the *Drosophila* telomere retrotransposable elements. **Genome Research**, v. 16, n. 10, p. 1231–1240, 2006.

GOMEZ-DIAZ, C. et al. The Two Main Olfactory Receptor Families in *Drosophila*, ORs and IRs: A Comparative Approach. **Frontiers in Cellular Neuroscience**, v. 12, p. 253, 2018.

GOOD, R. T. et al. The Molecular Evolution of Cytochrome P450 Genes within and between *Drosophila* Species. **Genome Biology and Evolution**, v. 6, n. 5, p. 1118–1134, 2014.

GOTEA, V.; MAKALOWSKI, W. Do transposable elements really contribute to proteomes? **Trends in Genetics**, v. 22, n. 5, p. 260–267, 2006.

GOULD, F. Role of Behavior in the Evolution of Insect Adaptation to Insecticides and Resistant Host Plants. **Bulletin of the Entomological Society of America**, v. 30, n. 4, p. 34–41, 1984.

GUILLÉN, Y. et al. Genomics of Ecological Adaptation in Cactophilic *Drosophila*. **Genome Biology and Evolution**, v. 7, n. 1, p. 349–366, 2015.

GUO, S.; KIM, J. Molecular Evolution of *Drosophila* Odorant Receptor Genes. **Molecular Biology and Evolution**, v. 24, n. 5, p. 1198–1207, 2007.

HALLEM, E. A.; HO, M. G.; CARLSON, J. R. The Molecular Basis of Odor Coding in the *Drosophila* Antenna. **Cell**, v. 117, n. 7, p. 965–979, 2004.

HARIGAYA, Y.; PARKER, R. No-go decay: a quality control mechanism for RNA in

translation. **Wiley Interdisciplinary Reviews: RNA**, v. 1, n. 1, p. 132–141, 2010.

HAVENS, J. A.; ETGES, W. J. Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IX. Host plant and population specific epicuticular hydrocarbon expression influences mate choice and sexual selection. **Journal of Evolutionary Biology**, v. 26, n. 3, p. 562–576, 2013.

HEARD, S. B. Use of Host-Plant Trait Space by Phytophagous Insects during Host-Associated Differentiation: The Gape-and-Pinch Model. **International Journal of Ecology**, v. 2012, n. 1, p. 192345, 2012.

HEINZEN, E. L. et al. Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits. **PLOS Biology**, v. 6, n. 12, p. e1000001, 2008.

HENDRY, A. P. Ecological speciation! Or the lack thereof? **Canadian Journal of Fisheries and Aquatic Sciences**, v. 66, n. 8, p. 1383–1398, 2009.

HERNÁNDEZ-ROLDÁN, J. L. et al. Integrative analyses unveil speciation linked to host plant shift in pialia butterflies. **Molecular Ecology**, v. 25, n. 17, p. 4267–4284, 2016.

HIRSCH, C. D.; SPRINGER, N. M. Transposable element influences on gene expression in plants. **Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms**, Plant Gene Regulatory Mechanisms and Networks. v. 1860, n. 1, p. 157–165, 2017.

HOCUTT, T. A. M. G. D. Reproductive Isolation in Sonoran Desert *Drosophila*: Testing the Limits of the Rules. In: HOWARD, D. J.; BERLOCHER, S. H. (Eds.). **Endless Forms: Species and Speciation**. Oxford University Press, p. 234-244, 1998.

HOF, A. E. VAN'T et al. The industrial melanism mutation in British peppered moths is a transposable element. **Nature**, v. 534, n. 7605, p. 102–105, 2016.

HORVÁTH, V.; MERENCIANO, M.; GONZÁLEZ, J. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. **Trends in Genetics**, v. 33, n. 11, p. 832–841, 2017.

HOWARD, D. J.; BERLOCHER, S. H. **Endless Forms: Species and Speciation**. Oxford University Press, 1998.

HOWARD, R. W.; BLOMQUIST, G. J. Ecological, behavioral, and biochemical aspects of the insect hydrocarbons. **Annual Review of Entomology**, v. 50, n. Volume 50, 2005, p. 371–393, 2005.

HUG, N.; LONGMAN, D.; CÁCERES, J. F. Mechanism and regulation of the nonsense-mediated decay pathway. **Nucleic Acids Research**, v. 44, n. 4, p. 1483–1495, 2016.

HUGHES, J. F.; COFFIN, J. M. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. **Nature Genetics**, v. 29, n. 4, p. 487–489, 2001.

JORDAN, I. K. et al. Origin of a substantial fraction of human regulatory sequences from transposable elements. **Trends in Genetics**, v. 19, n. 2, p. 68–72, 2003.

KAZAZIAN, H. H. Mobile Elements: Drivers of Genome Evolution. **Science**, v. 303, n. 5664, p. 1626–1632, 2004.

KELLEY, J. L. et al. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. **Nature Communications**, v. 5, n. 1, p. 4611, 2014.

LACHAISE, D.; SILVAIN, J.-F. How two Afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster-D. simulans* palaeogeographic riddle. **Genetica**, v. 120, p. 17–39, 2004.

LAMPRECHT, B. et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. **Nature Medicine**, v. 16, n. 5, p. 571–579, 2010.

LANDE, R. The genetic covariance between characters maintained by pleiotropic mutations. **Genetics**, v. 94, n. 1, p. 203–215, 1980.

LARTER, N. K.; SUN, J. S.; CARLSON, J. R. Organization and function of *Drosophila* odorant binding proteins. **eLife**, v. 5, p. e20242, 2016.

LAVIALLE, C. et al. Paleovirology of "syncytins", retroviral *env* genes exapted for a role in placentation. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 368, n. 1626, p. 20120507, 2013.

LIPATOV, M. et al. Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. **BMC Biology**, v. 3, n. 1, p. 24, 2005.

LOCK, F. E. et al. Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. **Proceedings of the National Academy of Sciences**, v. 111, n. 34, p. E3534–E3543, 2014.

LOW, W. Y. et al. Molecular Evolution of Glutathione *S* -Transferases in the Genus *Drosophila*. **Genetics**, v. 177, n. 3, p. 1363–1375, 2007.

LU, Z. et al. The prevalence, evolution and chromatin signatures of plant regulatory elements. **Nature Plants**, v. 5, n. 12, p. 1250–1259, 2019.

LUQUE, T.; O'REILLY, D. R. Functional and phylogenetic analyses of a putative *Drosophila melanogaster* UDP-glycosyltransferase gene. **Insect Biochemistry and Molecular Biology**, v. 32, n. 12, p. 1597–1604, 2002.

LYNCH, B. T. et al. Differential pre-mRNA Splicing Alters the Transcript Diversity of Helitrons Between the Maize Inbred Lines. **G3 Genes|Genomes|Genetics**, v. 5, n. 8, p. 1703–1711, 2015.

MAGNACCA, K. N.; O'GRADY, P. M. A Subgroup Structure for the Modified Mouthparts Species Group of Hawaiian *Drosophila*. **Proceedings of The Hawaiian Entomological Society**, v. 38, p. 87–101, 2006.

MAGWIRE, M. M. et al. Successive Increases in the Resistance of *Drosophila* to Viral Infection through a Transposon Insertion Followed by a Duplication. **PLoS Genetics**, v. 7, n. 10, p. e1002337, 2011.

MAMIDALA, P. et al. RNA-Seq and molecular docking reveal multi-level pesticide resistance in the bed bug. **BMC Genomics**, v. 13, n. 1, p. 6, 2012.

MARGUERAT, S.; BÄHLER, J. RNA-seq: from technology to biology. **Cellular and Molecular Life Sciences**, v. 67, n. 4, p. 569–579, 2010.

MARKOW, T. A. Sexual Isolation among Populations of *Drosophila mojavensis*. **Evolution**, v. 45, n. 6, p. 1525–1529, 1991.

MARKOW, T. A. Host use and host shifts in *Drosophila*. **Current Opinion in Insect Science**, v. 31, p. 139–145, 2019.

MARKOW, T. A.; O'GRADY, P. M. *Drosophila* Biology in the Genomic Age. **Genetics**, v. 177, n. 3, p. 1269–1276, 2007.

MASSIE, K. R.; ANN MARKOW, T. Sympatry, allopatry and sexual isolation between *Drosophila mojavensis* and *D. arizonae*: Sympatry, allopatry and sexual isolation in *Drosophila*. **Hereditas**, v. 142, n. 2005, p. 51–55, 2006.

MATSUBAYASHI, K. W.; OHSHIMA, I.; NOSIL, P. Ecological speciation in phytophagous insects. **Entomologia Experimentalis et Applicata**, v. 134, n. 1, p. 1–27, 2010.

MATSUO, T. et al. Odorant-Binding Proteins OBP57d and OBP57e Affect Taste Perception

and Host-Plant Preference in *Drosophila sechellia*. **PLoS Biology**, v. 5, n. 5, p. e118, 2007.

MATZKIN, L. M.; MARKOW, T. A. Transcriptional Regulation of Metabolism Associated With the Increased Desiccation Resistance of the Cactophilic *Drosophila mojavensis*. **Genetics**, v. 182, n. 4, p. 1279–1288, 2009.

MCCLINTOCK, B. Chromosome Organization and Genic Expression. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 16, p. 13–47, 1951.

MCGINNIS, W.; SHERMOEN, A. W.; BECKENDORF, S. K. A transposable element inserted just 5′ to a *Drosophila* glue protein gene alters gene expression and chromatin structure. **Cell**, v. 34, n. 1, p. 75–84, 1983.

MCKENZIE, S. K.; KRONAUER, D. J. C. The genomic architecture and molecular evolution of ant odorant receptors. **Genome Research**, v. 28, n. 11, p. 1757–1765, 2018.

MERCURI, R. L. V. et al. FREDDIE: A comprehensive tool for detecting exonization of retrotransposable elements in short and long RNA sequencing data. **bioRxiv**, 2024. Available at: https://www.biorxiv.org/content/10.1101/2024.04.22.590610v1. Accessed in Nov 1st, 2024.

MÉREL, V. et al. The Worldwide Invasion of *Drosophila suzukii* Is Accompanied by a Large Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions. **Molecular Biology and Evolution**, v. 38, n. 10, p. 4252–4267, 2021.

MI, S. et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. **Nature**, v. 403, n. 6771, p. 785–789, 2000.

MILLER, W. J. et al. Molecular domestication — more than a sporadic episode in evolution. In: MCDONALD, J. F. (Ed.). **Transposable Elements and Genome Evolution**. Dordrecht: Springer Netherlands, 2000. p. 197–207.

MOHAMED, M. et al. A Transposon Story: From TE Content to TE Dynamic Invasion of *Drosophila* Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. **Cells**, v. 9, n. 8, p. 1776, 2020.

MONTELL, C. A taste of the *Drosophila* gustatory receptors. **Current Opinion in Neurobiology**, v. 19, n. 4, p. 345–353, 2009.

MONTGOMERY, E.; CHARLESWORTH, B.; LANGLEY, C. H. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. **Genetical Research**, v. 49, n. 1, p. 31–41, 1987.

MOSCHETTI, R. et al. “What You Need, Baby, I Got It”: Transposable Elements as Suppliers of Cis-Operating Sequences in Drosophila. **Biology**, v. 9, n. 2, p. 25, 2020.

NICOLAU, M.; PICAULT, N.; MOISSIARD, G. The Evolutionary Volte-Face of Transposable Elements: From Harmful Jumping Genes to Major Drivers of Genetic Innovation. **Cells**, v. 10, n. 11, p. 2952, 2021.

OAKESHOTT, J. G. et al. Evolutionary genetics of *Drosophila* esterases. **Genetica**, v. 90, n. 2–3, p. 239–268, 1993.

OHNO, S. So much "junk" DNA in our genome. In: **Evolution of Genetic Systems**. New York: Gordon and Breach, v. 23, p. 366–370, 1972.

OLIVEIRA, D. C. S. G. et al. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. **Molecular Phylogenetics and Evolution**, v. 64, n. 3, p. 533–544, 2012.

OLIVEIRA, D. S. et al. Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. **Journal of Evolutionary Biology**, v. 34, n. 4, p. 628–638, 2021.

OLIVEIRA, D. S. et al. ChimeraTE: a pipeline to detect chimeric transcripts derived from genes and transposable elements. **Nucleic Acids Research**, v. 51, n. 18, p. 9764–9784, 2023.

OLIVER, K. R.; GREENE, W. K. Transposable elements: powerful facilitators of evolution. **BioEssays**, v. 31, n. 7, p. 703–714, 2009.

ONG-ABDULLAH, M. et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. **Nature**, v. 525, n. 7570, p. 533–537, 2015.

OSADA, N.; MIYAGI, R.; TAKAHASHI, A. Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of *Drosophila melanogaster*. **Genetics**, v. 206, n. 4, p. 2139–2148, 2017.

PARDUE, M.-L.; DEBARYSHE, P. G. *Drosophila* telomeres: two transposable elements with important roles in chromosomes. Em: MCDONALD, J. F. (Ed.). **Transposable Elements and Genome Evolution**. Dordrecht: Springer Netherlands, 2000. p. 189–196.

PARISOT, N. et al. The transposable element-rich genome of the cereal pest *Sitophilus oryzae*. **BMC biology**, v. 19, n. 1, p. 241, 2021.

PFEILER, E. et al. Genetic, ecological and morphological differences among populations of

the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. **Journal of Natural History**, v. 43, n. 15–16, p. 923–938, 2009.

PINSON, M.-E. et al. CLIFinder: identification of LINE-1 chimeric transcripts in RNA-seq data. **Bioinformatics**, v. 34, n. 4, p. 688–690, 2018.

PIRIYAPONGSA, J. et al. Evaluating the protein coding potential of exonized transposable element sequences. **Biology Direct**, v. 2, n. 1, p. 31, 2007.

POWELL, J. R. **Progress and Prospects in Evolutionary Biology: The Drosophila Model**. Oxford University Press, 1997.

RAUSHER, M. D. Conditioning and genetic variation as causes of individual variation in the oviposition behaviour of the tortoise beetle, *Deloyala guttata*. **Animal Behaviour**, v. 31, n. 3, p. 743–747, 1983.

RAUSHER, M. D. The evolution of habitat preference: avoidance and adaptation. In: **Evolution of Insect Pests: Patterns of Variation**. Evolutionary Biology. Wiley, New York, 1993. p. 259–283.

REBOLLO, R.; FARIVAR, S.; MAGER, D. L. C-GATE - catalogue of genes affected by transposable elements. **Mobile DNA**, v. 3, n. 1, p. 9, 2012.

REBOLLO, R.; ROMANISH, M. T.; MAGER, D. L. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. **Annual Review of Genetics**, v. 46, n. 1, p. 21–42, 2012.

RECH, G. E. et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. **PLOS Genetics**, v. 15, n. 2, p. e1007900, 2019.

REED, L. K.; MARKOW, T. A. Early events in speciation: Polymorphism for hybrid male sterility in *Drosophila*. **Proceedings of the National Academy of Sciences**, v. 101, n. 24, p. 9009–9012, 2004.

REED, L. K.; NYBOER, M.; MARKOW, T. A. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*: DROSOPHILA MOJAVENSIS SPECIES RELATIONSHIPS. **Molecular Ecology**, v. 16, n. 5, p. 1007–1022, 2006.

REED, L. K.; NYBOER, M.; MARKOW, T. A. Evolutionary relationships of *Drosophila*

mojavensis geographic host races and their sister species *Drosophila arizonae*. **Molecular Ecology**, v. 16, n. 5, p. 1007–1022, 2007.

RIVERON, J.; BOTO, T.; ALCORTA, E. Transcriptional basis of the acclimation to high environmental temperature at the olfactory receptor organs of *Drosophila melanogaster*. **BMC Genomics**, v. 14, n. 1, p. 259, 2013.

R'KHA, S.; CAPY, P.; DAVID, J. R. Host-plant specialization in the *Drosophila melanogaster* species complex: a physiological, behavioral, and genetical analysis. **Proceedings of the National Academy of Sciences**, v. 88, n. 5, p. 1835–1839, 1991.

ROQUIS, D. et al. Genomic impact of stress-induced transposable element mobility in *Arabidopsis*. **Nucleic Acids Research**, v. 49, n. 18, p. 10431–10447, 2021.

ROSS, C. L.; MARKOW, T. A. Microsatellite variation among diverging populations of *Drosophila mojavensis*. **Journal of Evolutionary Biology**, v. 19, n. 5, p. 1691–1700, 2006.

RUIZ, A.; HEED, W. B. Host-Plant Specificity in the Cactophilic *Drosophila mulleri* Species Complex. **The Journal of Animal Ecology**, v. 57, n. 1, p. 237, 1988.

RUIZ, A.; HEED, W. B.; WASSERMAN, M. Evolution of the Mojavensis Cluster of Cactophilic *Drosophila* with Descriptions of Two New Species. **Journal of Heredity**, v. 81, n. 1, p. 30–42, 1990.

RUNDLE, H. D.; NOSIL, P. Ecological speciation. **Ecology Letters**, v. 8, n. 3, p. 336–352, 2005.

RYTZ, R.; CROSET, V.; BENTON, R. Ionotropic Receptors (IRs): Chemosensory ionotropic glutamate receptors in *Drosophila* and beyond. **Insect Biochemistry and Molecular Biology**, v. 43, n. 9, p. 888–897, 2013.

SÁNCHEZ-GRACIA, A.; ROZAS, J. Divergent evolution and molecular adaptation in the *Drosophila* odorant-binding protein family: inferences from sequence variation at the OS-E and OS-F genes. **BMC Evolutionary Biology**, v. 8, n. 1, p. 323, 2008.

SCHLUTER, D. Evidence for Ecological Speciation and Its Alternative. **Science**, v. 323, n. 5915, p. 737–741, 6 fev. 2009.

SELA, N. et al. Characteristics of Transposable Element Exonization within Human and Mouse. **PLoS ONE**, v. 5, n. 6, p. e10907, 2010.

SHAH, N. M. et al. Pan-cancer analysis identifies tumor-specific antigens derived from

transposable elements. **Nature Genetics**, v. 55, n. 4, p. 631–639, 2023.

SINZELLE, L.; IZSVÁK, Z.; IVICS, Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. **Cellular and Molecular Life Sciences**, v. 66, n. 6, p. 1073–1093, 2009.

SLOTKIN, R. K.; MARTIENSSEN, R. Transposable elements and the epigenetic regulation of the genome. **Nature Reviews Genetics**, v. 8, n. 4, p. 272–285, 2007.

SOREK, R. The birth of new exons: Mechanisms and evolutionary consequences. **RNA**, v. 13, n. 10, p. 1603–1608, 2007.

SOREK, R.; AST, G.; GRAUR, D. Alu-Containing Exons are Alternatively Spliced. **Genome Research**, v. 12, n. 7, p. 1060–1067, 2002.

SPALLER, T. et al. TRE5-A retrotransposition profiling reveals putative RNA polymerase III transcription complex binding sites on the *Dictyostelium* extrachromosomal rDNA element. **PLOS ONE**, v. 12, n. 4, p. e0175729, 2017.

STAFFORD, J. W. et al. Integration of Taste and Calorie Sensing in *Drosophila*. **Journal of Neuroscience**, v. 32, n. 42, p. 14767–14774, 2012.

STENNETT, M. D.; ETGES, W. J. Premating Isolation Is Determined by Larval Rearing Substrates in Cactophilic *Drosophila mojavensis*. III. Epicuticular Hydrocarbon Variation Is Determined by Use of Different Host Plants in *Drosophila mojavensis* and *Drosophila arizonae*. **Journal of Chemical Ecology**, v. 23, n. 12, p. 2803–2824, 1997.

SUN, H.; BUCHON, N.; SCOTT, J. G. *Mdr65* decreases toxicity of multiple insecticides in *Drosophila melanogaster*. **Insect Biochemistry and Molecular Biology**, v. 89, p. 11–16, 2017.

THORNBURG, B. G.; GOTEA, V.; MAKAŁOWSKI, W. Transposable elements as a significant source of transcription regulating signals. **Gene**, Genome and RNA: Expression and Functions. v. 365, p. 104–110, 2006.

THROCKMORTON, L. The *virilis* species group. **The Genetics and Biology of Drosophila**, v. 3, p. 227–296, 1982.

TREIBER, C. D.; WADDELL, S. Transposon expression in the *Drosophila* brain is driven by neighboring genes and diversifies the neural transcriptome. **Genome Research**, v. 30, n. 11, p. 1559–1569, 2020.

VASUDEVAN, S.; PELTZ, S. W.; WILUSZ, C. J. Non-stop decay—a new mRNA

surveillance pathway. **BioEssays**, v. 24, n. 9, p. 785–788, 2002.

VERTACNIK, K. L.; LINNEN, C. R. Evolutionary genetics of host shifts in herbivorous insects: insights from the age of genomics. **Annals of the New York Academy of Sciences**, v. 1389, n. 1, p. 186–212, 2017.

VIEIRA, F. G.; SÁNCHEZ-GRACIA, A.; ROZAS, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. **Genome Biology**, v. 8, n. 11, p. R235, 2007.

VIHERVAARA, A.; DUARTE, F. M.; LIS, J. T. Molecular mechanisms driving transcriptional stress responses. **Nature Reviews Genetics**, v. 19, n. 6, p. 385–397, 2018.

VOLFF, J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. **BioEssays**, v. 28, n. 9, p. 913–922, 2006.

WALSH, B. D. **On the phytophagic varieties and phytophagic species**. Proceedings of the Entomological Society of Philadelphia, 1864.

WANG, J. et al. MIR retrotransposon sequences provide insulators to the human genome. **Proceedings of the National Academy of Sciences**, v. 112, n. 32, p. E4428–E4437, 2015.

WANG, T. et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. **Proceedings of the National Academy of Sciences**, v. 104, n. 47, p. 18613–18618, 2007.

WERTHEIM, J. O. et al. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. **Molecular Biology and Evolution**, v. 32, n. 3, p. 820–832, 2015.

WHITEMAN, N. K.; PIERCE, N. E. Delicious poison: genetics of *Drosophila* host plant preference. **Trends in Ecology & Evolution**, v. 23, n. 9, p. 473–478, 2008.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, 2007.

WILLINGHAM, A. T.; KEIL, T. A tissue specific cytochrome P450 required for the structure and function of *Drosophila* sensory organs. **Mechanisms of Development**, v. 121, n. 10, p. 1289–1297, 2004.

WU, C. et al. Insect ATP-Binding Cassette (ABC) Transporters: Roles in Xenobiotic Detoxification and Bt Insecticidal Activity. **International Journal of Molecular Sciences**, v. 20, n. 11, p. 2829, 2019.

YE, M. et al. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. **Proceedings of the National Academy of Sciences**, v. 117, n. 14, p. 7905–7916, 2020.

YOUNUS, F. et al. Identification of candidate odorant degrading gene/enzyme systems in the antennal transcriptome of *Drosophila melanogaster*. **Insect Biochemistry and Molecular Biology**, v. 53, p. 30–43, 2014.

YU, Y. et al. Dynamic Regulation of Alternative Splicing by Silencers that Modulate 5′ Splice Site Competition. **Cell**, v. 135, n. 7, p. 1224–1236, 2008.

ZOUROS, E.; D'ENTREMONT, C. J. Sexual Isolation Among Populations of *Drosophila mojavensis*: Response to Pressure from a Related Species. **Evolution**, v. 34, n. 3, p. 421–430, 1980.

APPENDICES

Received: 15 May 2020 | Revised: 27 October 2020 | Accepted: 6 December 2020

DOI: 10.1111/jeb.13762

RESEARCH PAPER

Journal of Evolutionary Biology ESEB WILEY

Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*

Daniel Siqueira de Oliveira[1] | **Marcos Trindade Rosa**[2] | **Cristina Vieira**[1] | **Elgion L. S. Loreto**[3]

[1]Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, 1- Université de Lyon, Université Lyon 1, CNRS, Villeurbanne, France

[2]PPG Biodiversidade Animal, Universidade Federal de Santa Maria, Santa Maria, RS, Brazil

[3]Dep de Bioquímica e Biologia Molecular, Universidade Federal de Santa Maria, Santa Maria, RS, Brazil

Correspondence
Elgion L. S. Loreto, Departamento de Bioquímica e Biologia Molecular, CCNE, Universidade Federal de Santa Maria, Av. Roraima, 1000, 97105-900 Santa Maria, RS, Brazil.
Email: elgion@base.ufsm.br

Funding information
This study was supported by research grants and fellowships from Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq (421255/2018-8), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), PRONEX-FAPERGS (16/2551 – 0000 499-4) and University of Lyon, IDEX de Lyon grant to DSO, ANR GREEN grant to CV.

Abstract

It has been shown that stressors are capable of activating transposable elements (TEs). Currently, there is a hypothesis that stress activation of TEs may be involved in adaptive evolution, favouring the increase in genetic variability when the population is under adverse conditions. However, TE activation under stress is still poorly understood. In the present study, we estimated the fraction of differentially expressed TEs (DETEs) under ionizing radiation (144, 360 and 864 Gy) and oxidative stress (dioxin, formaldehyde and toluene) treatments. The stress intensity of each treatment was estimated by measuring the number of differentially expressed genes, and we show that several TEs families are activated by stress whereas others are repressed. The proportion of DETEs was positively related to stress intensity. However, even under the strongest stress, only a small fraction of TE families were activated (9.28%) and 17.72% were repressed. Considering all treatments together, the activated proportion was 19.83%. Nevertheless, as several TEs are incomplete or degenerated, only 10.55% of *D. melanogaster* mobilome is, at same time, activated by the stressors and able to transpose or at least code a protein. Thus, our study points out that although stress activates TEs, it is not a generalized activation process, and for some families, the stress induces repression.

KEYWORDS
genome evolution, mobilome, stress, transposon activation

1 | INTRODUCTION

It is currently well known that a significant fraction of eukaryotic genomes is made up of transposable elements (TEs). After its discovery by Barbara McClintock in the 1940s, several hypotheses were developed to explain the role of these elements in organisms, going from the suggestion that they are "selfish DNA" or "junk DNA," to the hypothesis that they are considered a raw material for evolution, being a major source of genetic variability (Biémont & Vieira, 2006; Hua-Van et al., 2011; Cosby et al., 2019). However, TE mobilization can have strong deleterious effects. When transposition occurs in the germ line, it can lead to sterility, as in *Drosophila* hybrid dysgenesis, or to reduced viability, as in hypermutable strains (Kidwell, 1983; Kocur et al., 1986). Furthermore, TE mobilization occurring in somatic tissues has been associated with ageing, cancer and degenerative diseases (De Cecco et al., 2013; Jang et al., 2019; Sedivy et al., 2013). Several mechanisms that control TE mobilization have evolved, such Piwi-interacting RNA (piRNA) and small interfering RNA (siRNA) regulation, DNA methylation, histone modification and chromatin remodelling (Horváth et al., 2017). Nevertheless, several

communications biology

ARTICLE

https://doi.org/10.1038/s42003-021-02095-0 OPEN

Genome-wide bioinformatic analyses predict key host and viral factors in SARS-CoV-2 pathogenesis

Mariana G. Ferrarini[1,11], Avantika Lal[2,11], Rita Rebollo[1], Andreas J. Gruber[3], Andrea Guarracino[4], Itziar Martinez Gonzalez[5], Taylor Floyd[6], Daniel Siqueira de Oliveira[7], Justin Shanklin[8], Ethan Beausoleil[8], Taneli Pusa[9], Brett E. Pickett[8✉] & Vanessa Aguiar-Pulido[10✉]

The novel betacoronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) caused a worldwide pandemic (COVID-19) after emerging in Wuhan, China. Here we analyzed public host and viral RNA sequencing data to better understand how SARS-CoV-2 interacts with human respiratory cells. We identified genes, isoforms and transposable element families that are specifically altered in SARS-CoV-2-infected respiratory cells. Well-known immunoregulatory genes including *CSF2, IL32, IL-6* and *SERPINA3* were differentially expressed, while immunoregulatory transposable element families were upregulated. We predicted conserved interactions between the SARS-CoV-2 genome and human RNA-binding proteins such as the heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and eukaryotic initiation factor 4 (eIF4b). We also identified a viral sequence variant with a statistically significant skew associated with age of infection, that may contribute to intracellular host–pathogen interactions. These findings can help identify host mechanisms that can be targeted by prophylactics and/or therapeutics to reduce the severity of COVID-19.

[1] University of Lyon, INSA-Lyon, INRA BF2I, Villeurbanne, France. [2] NVIDIA Corporation, Santa Clara, CA, USA. [3] Department of Biology, University of Konstanz, Konstanz, Germany. [4] Centre for Molecular Bioinformatics, Department of Biology, University Of Rome Tor Vergata, Rome, Italy. [5] Amsterdam UMC, Amsterdam, The Netherlands. [6] Center for Neurogenetics, Weill Cornell Medicine, Cornell University, New York, NY, USA. [7] Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS UMR 5558, Villeurbanne, France. [8] Brigham Young University, Provo, UT, USA. [9] Luxembourg Centre for Systems Biomedicine, Belvaux, Luxembourg. [10] Department of Computer Science, University of Miami, Coral Gables, FL, USA. [11] These authors contributed equally: Mariana G. Ferrarini, Avantika Lal. ✉email: brett_pickett@byu.edu; vanessa@cs.miami.edu

cells

MDPI

Article

Transposable Element Expression and Regulation Profile in Gonads of Interspecific Hybrids of *Drosophila arizonae* and *Drosophila mojavensis wrigleyi*

Cecília Artico Banho [1,2], Daniel Siqueira Oliveira [1,2], Annabelle Haudry [2], Marie Fablet [2,3], Cristina Vieira [2,*] and Claudia Marcia Aparecida Carareto [1,*]

[1] Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University (Unesp), São José do Rio Preto 15054-000, SP, Brazil; artico.banho@unesp.br (C.A.B.); daniel.siqueira@unesp.br (D.S.O.)
[2] Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR 5558, F-69622 Villeurbanne, France; annabelle.haudry@univ-lyon1.fr (A.H.); marie.fablet@univ-lyon1.fr (M.F.)
[3] Institut Universitaire de France (IUF), F-75231 Paris, France
* Correspondence: cristina.vieira@univ-lyon1.fr (C.V.); claudia.carareto@unesp.br (C.M.A.C.)

Abstract: Interspecific hybridization may lead to sterility and/or inviability through differential expression of genes and transposable elements (TEs). In *Drosophila*, studies have reported massive TE mobilization in hybrids from interspecific crosses of species presenting high divergence times. However, few studies have examined the consequences of TE mobilization upon hybridization in recently diverged species, such as *Drosophila arizonae* and *D. mojavensis*. We have sequenced transcriptomes of *D. arizonae* and the subspecies *D. m. wrigleyi* and their reciprocal hybrids, as well as piRNAs, to analyze the impact of genomic stress on TE regulation. Our results revealed that the differential expression in both gonadal tissues of parental species was similar. Globally, ovaries and testes showed few deregulated TEs compared with both parental lines. Analyses of small RNA data showed that in ovaries, the TE upregulation is likely due to divergence of copies inherited from parental genomes and lack of piRNAs mapping to them. Nevertheless, in testes, the divergent expression of genes associated with chromatin state and piRNA pathway potentially indicates that TE differential expression is related to the divergence of regulatory genes that play a role in modulating transcriptional and post-transcriptional mechanisms.

Keywords: repleta group; hybrids; transposable elements; expression

Citation: Banho, C.A.; Oliveira, D.S.; Haudry, A.; Fablet, M.; Vieira, C.; Carareto, C.M.A. Transposable Element Expression and Regulation Profile in Gonads of Interspecific Hybrids of *Drosophila arizonae* and *Drosophila mojavensis wrigleyi*. *Cells* **2021**, *10*, 3574. https://doi.org/10.3390/cells10123574

Academic Editor: Laura Fanti

Received: 9 November 2021
Accepted: 13 December 2021
Published: 18 December 2021

1. Introduction

Transposable elements (TEs) are repetitive DNA sequences that move from one place to another in the genome and between genomes. Because of their ability to mobilize, these elements play an important role in creating genetic variability and, consequently, in genome evolution and adaptation [1–3]. TE activation can be induced by environmental stress [4] and/or by genomic shocks, such as hybridization [5–8]. The disruption of genome stability following hybridization is attributed to the divergence of regulatory sequences and/or to the content of TEs [9]. Several studies have reported the role of TEs on reproductive isolation, which may be due to their ability to modify regulatory networks and gene expression and lead to structural rearrangements [1,10]. The effects of TEs on intraspecific or interspecific hybrids may be beneficial or deleterious, depending on the species. In *Drosophila* intraspecific hybrids, some TEs are responsible for hybrid dysgenesis syndrome, which is characterized by gonadal atrophy and generally affects offspring obtained from crosses in only one direction [11–16]. In interspecific hybrids obtained from highly divergent parental species, massive TE expression and mobilization have been observed [7,9,17–21]. This phenomenon was mainly associated with the lack of small interfering RNA from the class of the Piwi-RNA (piRNA) that are able to control TE mobilization and expression,

GBE

Impact of Heat Stress on Transposable Element Expression and Derived Small RNAs in *Drosophila subobscura*

Alejandra Bodelón[1], Marie Fablet [2,3], Daniel Siqueira de Oliveira[2,4], Cristina Vieira [2], and Maria Pilar García Guerreiro [1,*]

[1]Grup de Genòmica, Bioinformática i Biologia Evolutiva, Departament de Genètica i Microbiologia (Edifici C), Universitat Autònoma de Barcelona, Barcelona, Spain

[2]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Villeurbanne, France

[3]Institut universitaire de France, Paris, France

[4]Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University (Unesp), São Paulo, Brazil

*Corresponding author: E-mail: mariapilar.garcia.guerreiro@uab.es.

Accepted: October 09, 2023

Abstract

Global warming is forcing insect populations to move and adapt, triggering adaptive genetic responses. Thermal stress is known to alter gene expression, repressing the transcription of active genes, and inducing others, such as those encoding heat shock proteins. It has also been related to the activation of some specific transposable element (TE) families. However, the actual magnitude of this stress on the whole genome and the factors involved in these genomic changes are still unclear. We studied mRNAs and small RNAs in gonads of two *Drosophila subobscura* populations, considered a good model to study adaptation to temperature changes. In control conditions, we found that a few genes and TE families were differentially expressed between populations, pointing out their putative involvement in the adaptation of populations to their different environments. Under heat stress, sex-specific changes in gene expression together with a trend toward overexpression, mainly of heat shock response-related genes, were observed. We did not observe large changes of TE expression nor small RNA production due to stress. Only population and sex-specific expression changes of some TE families (mainly retrotransposons), or the amounts of siRNAs and piRNAs, derived from specific TE families were observed, as well as the piRNA production from some piRNA clusters. Changes in small RNA amounts and TE expression could not be clearly correlated, indicating that other factors as chromatin modulation could also be involved. This work provides the first whole transcriptomic study including genes, TEs, and small RNAs after a heat stress in *D. subobscura*.

Key words: *Drosophila*, stress, transposable elements, piRNAs, RNA-seq, heat shock stress.

Significance

Global warming provokes intense heat waves affecting the organism genomes. Usually, heat stress alters gene expression, but the effect on transposable element (TE) activity and their control mechanisms, involving small RNAs, is not clear. Here, we studied how the thermal stress affects the gonadal transcriptome of *Drosophila subobscura*, and we found that changes on the expression of specific TE families were not always coupled with their derived small RNAs, indicating that other factors should also be involved. This work provides the first whole genome expression study in *D. subobscura*, including TE expression after a heat stress, and provides a framework for future studies on the thermal effects on the epigenome and their consequences for organisms.

G3, 2024, 14(3), jkad299
https://doi.org/10.1093/g3journal/jkad299
Advance Access Publication Date: 29 December 2023
Investigation

A comprehensive evolutionary scenario for the origin and neofunctionalization of the *Drosophila* speciation gene *Odysseus* (*OdsH*)

William Vilas Boas Nunes,[1,2] Daniel Siqueira Oliveira ,[1,2] Guilherme de Rezende Dias,[3] Antonio Bernardo Carvalho ,[3] Ícaro Putinhon Caruso,[1] Joice Matos Biselli,[1] Nathalie Guegen,[4] Abdou Akkouche,[4] Nelly Burlet,[2] Cristina Vieira ,[2] Claudia M.A. Carareto [1,*]

[1]Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University (Unesp), 2265 Cristóvão Colombo Street, 15054-000 São José do Rio Preto, Brazil
[2]Laboratoire de Biométrie et Biologie Evolutive UMR 5558, Université de Lyon, Université Lyon 1, CNRS, Bât. Grégor Mendel, 43 Boulevard 11 Novembre 1918, 69622 Villeurbanne, France
[3]Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, CCS sl A2-075, 373 Carlos Chagas Filho Avenue, 21941-971 Rio de Janeiro, Brazil
[4]Faculté de Médecine, iGReD, Université Clermont Auvergne, CNRS, INSERM, 4 Bd Claude Bernard, 63000 Clermont-Ferrande, France

*Corresponding author: Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University (Unesp), 2265 Cristóvão Colombo Street, 15054-000 São José do Rio Preto, Brazil. Email: claudia.carareto@unesp.br

Odysseus (*OdsH*) was the first speciation gene described in *Drosophila* related to hybrid sterility in offspring of mating between *Drosophila mauritiana* and *Drosophila simulans*. Its origin is attributed to the duplication of the gene *unc-4* in the subgenus *Sophophora*. By using a much larger sample of Drosophilidae species, we showed that contrary to what has been previously proposed, *OdsH* origin occurred 62 MYA. Evolutionary rates, expression, and transcription factor–binding sites of *OdsH* evidence that it may have rapidly experienced neofunctionalization in male sexual functions. Furthermore, the analysis of the OdsH peptide allowed the identification of mutations of *D. mauritiana* that could result in incompatibility in hybrids. In order to find if *OdsH* could be related to hybrid sterility, beyond *Sophophora*, we explored the expression of *OdsH* in *Drosophila arizonae* and *Drosophila mojavensis*, a pair of sister species with incomplete reproductive isolation. Our data indicated that *OdsH* expression is not atypical in their male-sterile hybrids. In conclusion, we have proposed that the origin of *OdsH* occurred earlier than previously proposed, followed by neofunctionalization. Our results also suggested that its role as a speciation gene might be restricted to *D. mauritiana* and *D. simulans*.

Keywords: gene duplication; *unc-4*; homeodomain; transcription factor; *OdsH* expression; Drosophilidae

Introduction

Odysseus (*OdsH*) was the first so-called speciation gene characterized in *Drosophila*, specifically between *Drosophila mauritiana* and *Drosophila simulans* (Ting *et al.* 1998). The role of *OdsH* within the male hybrid sterility was attributed to the introgression of a sequence from *D. mauritiana* encompassing *OdsH* into the *D. simulans* genome (Perez *et al.* 1993; Perez and Wu 1995; Ting *et al.* 1998). The atypical expression of *OdsH* at the apical testis region was observed in these hybrids, which was not observed for fertile hybrids and parental species (Sun *et al.* 2004). The origin of the *OdsH* gene is proposed to have arisen by duplication of the *unc-4* gene, a conserved gene in Metazoa located *in tandem* with *OdsH* (Ting *et al.* 2004). The gene *OdsH*, the duplicated copy, is expressed in spermatocytes in species of the *melanogaster* subgroup and acts as a transcription factor binding to heterochromatic regions (Ting *et al.* 2004; Bayes and Malik 2009). Meanwhile, *unc*-4, the parental gene, is a transcription factor associated with motor neuron and proprioceptor developmental pathways in *Drosophila melanogaster* (Tabuchi *et al.* 1998; Lacin and Truman 2016; Lacin *et al.* 2019, 2020), similar to its conserved single-copy ortholog, which acts on motor neuron and optical sensorial cell development in *Caenorhabditis elegans* (Miller *et al.* 1992; Fox *et al.* 2005; Marques *et al.* 2019).

Both genes, *unc-4* and *OdsH*, encode homologous DNA-binding homeodomains, phylogenetically classified in the Paired-like class (Winnier *et al.* 1999; Copley 2005). The *OdsH* homeodomain has a high amino acid substitution rate in species of the *melanogaster* subgroup, corresponding to a higher divergence between the domains from *unc-4* between *Drosophila* and evolutionarily distant species, such as *C. elegans* (Ting *et al.* 2004). As expected for duplicated genes, the faster evolution of the *unc-4* paralog was associated with the acquisition of novel functions in the testis and with the speciation process (Ting *et al.* 1998, 2004).

Since the *OdsH* duplicate has been proposed to be a new gene in the *Sophophora* subgenus (Ting *et al.* 2004) and is associated with speciation in this clade, we would not expect to see this gene further in *Drosophila* phylogeny. However, searches in orthology databases GenTree (Shao *et al.* 2019) and OrthoDB (Kuznetsov *et al.* 2023) indicated the presence of *OdsH* duplicate in the ancestral node of the *Drosophila* genus, highlighting that its origin might be older than previously thought. We have thus asked the following questions: (1) how

Received on 22 November 2023; accepted on 20 December 2023

Transposable element expression with variation in sex chromosome number: insights into a toxic Y effect on human longevity

Jordan Teoli[1,2,3,4]*, Miriam Merenciano[2], Marie Fablet[2,5], Anamaria Necsulea[2], Daniel Siqueira-de-Oliveira[2], Alessandro Brandulas-Cammarata[6,7], Audrey Labalme[8], Hervé Lejeune[9], Jean-François Lemaitre[2], François Gueyffier[2], Damien Sanlaville[8], Claire Bardel[1,2], Cristina Vieira[2]†, Gabriel AB Marais[2,4,10,11]†, Ingrid Plotton[1,3,9]†

[1] Laboratoire de Biochimie et Biologie Moléculaire, Centre de Biologie et Pathologie Est, Hospices Civils de Lyon; Bron, 69500, France.
[2] Universite Claude Bernard Lyon 1, UMR 5558, CNRS, Laboratoire de Biométrie et Biologie Evolutive; Villeurbanne, 69310, France.
[3] Université Claude Bernard Lyon 1, Institut Cellule Souche et Cerveau (SBRI), Unité INSERM 1208, Centre de Recherche INSERM; Bron, 69500, France.
[4] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto; Vairão, 4485-661, Portugal.
[5] Institut universitaire de France.
[6] Department of Ecology and Evolution, University of Lausanne; Lausanne, 1015, Switzerland.
[7] SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[8] Service de Génétique, Hospices Civils de Lyon; Bron, 69500, France.
[9] Service de médecine de la reproduction, Hôpital Femme-Mère-Enfant, Hospices Civils de Lyon; Bron, 69500, France.
[10] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto; Porto, 4099-002, Portugal.
[11] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão; Vairão, 4485-661, Portugal.

†Equal contributions as senior authors

*Corresponding author
Correspondence: jordan.teoli@chu-lyon.fr

ABSTRACT

Why women live longer than men is still an open question in human biology. Sex chromosomes have been proposed to play a role in the observed sex gap in longevity, and the Y male chromosome has been suspected of having a potential toxic genomic impact on male longevity. It has been hypothesized that in aging individuals, TE repression is diminished, which could lead to detrimental effects (e.g. somatic mutations, perturbed gene expression) and to an acceleration of the aging process. As the Y chromosome is typically enriched in transposable elements (TE), this could explain why the presence of a Y

chromosome is associated with shorter longevity. Using transcriptomic data from humans with atypical karyotypes, we found an association between TE expression and the presence and number of Y chromosomes. These findings are consistent with the existence of a toxic Y effect on men's longevity.

Keywords: transposable elements, lifespan, sex chromosomes, toxic Y effect, transcriptomics, differential expression analysis

Peer Community Journal

Section: Genomics

Research article

Published
2024-09-10

Cite as
Rita Rebollo, Pierre Gerenton, Eric Cumunel, Arnaud Mary, François Sabot, Nelly Burlet, Benjamin Gillet, Sandrine Hughes, Daniel S. Oliveira, Clément Goubert, Marie Fablet, Cristina Vieira and Vincent Lacroix (2024) *Identification and quantification of transposable element transcripts using Long-Read RNA-seq in Drosophila germline tissues*, Peer Community Journal, 4: e89.

Correspondence
cristina.vieira@univ-lyon1.fr
vincent.lacroix@univ-lyon1.fr

Peer-review
Peer reviewed and recommended by PCI Genomics, https://doi.org/10.24072/pci.genomics.100250

Identification and quantification of transposable element transcripts using Long-Read RNA-seq in Drosophila germline tissues

Rita Rebollo[1], Pierre Gerenton[2,3], Eric Cumunel[2,3], Arnaud Mary[2,3], François Sabot[4], Nelly Burlet[2], Benjamin Gillet[5], Sandrine Hughes[5], Daniel S. Oliveira[2,6], Clément Goubert[7], Marie Fablet[2,8], Cristina Vieira[2,3], and Vincent Lacroix[2,3]

Volume **4** (2024), article e89

https://doi.org/10.24072/pcjournal.457

Abstract

Transposable elements (TEs) are repeated DNA sequences potentially able to move throughout the genome. In addition to their inherent mutagenic effects, TEs can disrupt nearby genes by donating their intrinsic regulatory sequences, for instance, promoting the ectopic expression of a cellular gene. TE transcription is therefore not only necessary for TE transposition *per se* but can also be associated with TE-gene fusion transcripts, and in some cases, be the product of pervasive transcription. Hence, correctly determining the transcription state of a TE copy is essential to apprehend the impact of the TE in the host genome. Methods to identify and quantify TE transcription have mostly relied on short RNA-seq reads to estimate TE expression at the family level while using specific algorithms to discriminate copy-specific transcription. However, assigning short reads to their correct genomic location, and genomic feature is not trivial. Here we retrieved full-length cDNA (TeloPrime, Lexogen) of *Drosophila melanogaster* gonads and sequenced them using Oxford Nanopore Technologies. We show that long-read RNA-seq can be used to identify and quantify transcribed TEs at the copy level. In particular, TE insertions overlapping annotated genes are better estimated using long reads than short reads. Nevertheless, long TE transcripts (> 4.5 kb) are not well captured. Most expressed TE insertions correspond to copies that have lost their ability to transpose, and within a family, only a few copies are indeed expressed. Long-read sequencing also allowed the identification of spliced transcripts for around 107 TE copies. Overall, this first comparison of TEs between testes and ovaries uncovers differences in their transcriptional landscape, at the subclass and insertion level.

[1]INRAE, INSA Lyon, BF2I, UMR203, Villeurbanne, France, [2]Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR5558, Villeurbanne, France., [3]ERABLE team, Inria, Lyon Rhone-Alpes, Villeurbanne, France, [4]DIADE unit, Univ Montpellier, Cirad, IRD, Montpellier, France, [5]Institut de Génomique Fonctionnelle de Lyon (IGFL), CNRS UMR 5242 , Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1 , Lyon, France., [6]São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, SP, Brazil., [7]Human Genetics, McGill University, Montreal, QC, Canada, [8]Institut Universitaire de France (IUF), Paris, France.

ChimeraTE: Differences between Mode 1 and Mode 2

ChimeraTE Mode 1 and Mode 2 use different alignment strategies and downstream approaches to detect chimeric transcripts (see Methods). In order to test whether these differences can lead to distinct outputs, we compared chimeras from Mode 1 and Mode 2 by using the same RNA-seq libraries from four *D. melanogaster* wild-type strains. Taking all strains together, Mode 2 uncovered 171 chimeric transcripts generated by 165 out of the 327 genes (50.46%) identified by Mode 1, representing 25% of all TE-initiated transcripts detected by Mode 1; ~52% of TE-exonized; ~29% of TE-terminated (Figure S1.1A). These results indicate that ChimeraTE Mode 2 had low sensitivity (27.39%) to detect chimeric transcripts from TE insertions near genes. For chimeras derived from TE inside genes, 17.23% of them were detected by the transcriptome assembly approach in Mode 2, showing the relevance of performing this optional analysis. However, it must be considered that *--assembly* performed by Mode 2 is time-consuming, as well as hardware-consuming (Table S1.1).

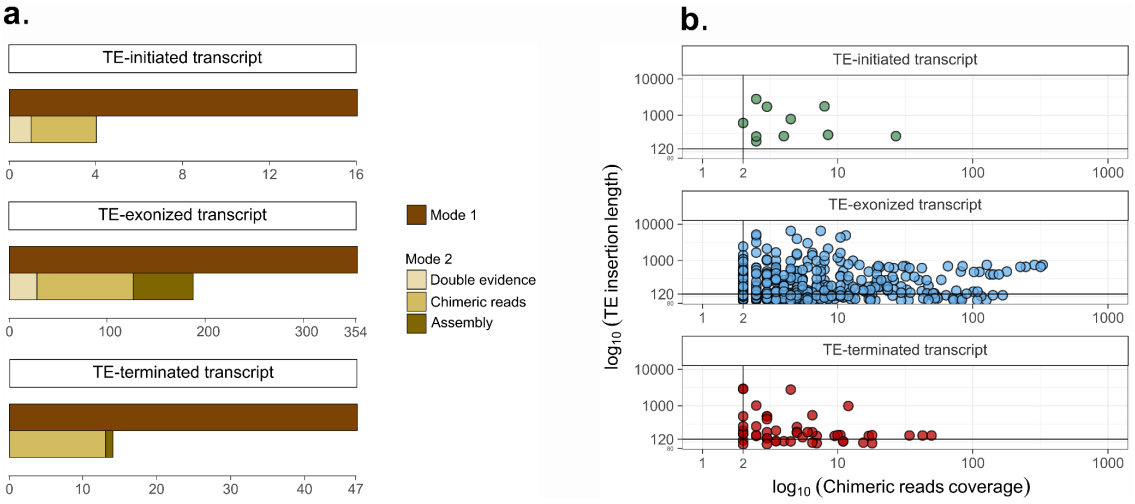


Figure S1.1: **A**) Total chimeric transcripts detected by ChimeraTE mode 1 and ChimeraTE Mode 2. The three brown boxes depict the type of evidence used by Mode 2 to support the chimeric transcripts (see Methods). Mode 2 had more efficiency to detect chimeric transcripts derived from TEs inside exons than near genes. **B**) Chimeric transcripts found by Mode 1, but not by Mode 2. 47.55% of all chimeric transcripts detected only by Mode 1 have TEs shorter than the read length (120 nt), and 32.51% of chimeras with TEs longer than reads have low chimeric reads coverage (10 chimeric reads). These factors explain the differences between results found by both Modes.

Species	Reads	Mode 1	Mode 2
		Time	***Time***
dmgoth101	36M	01h:50min	11h:25min
dmgoth63	40M	01h:55min	14h:20min
dmsj23	35M	01h:30min	11h:35min
dmsj7	42M	02h:00min	13h:35min
H. sapiens	58M	1d 3h:00min	3d13h:40min
A. thaliana	45M	02h:32min	19h:10min
Poecilia reticulata	49M	3h:20min	1d 10h:05min

Table S1.1: Running time of ChimeraTE Mode 1 and Mode 2 (with --assembly option), in the four *D. melanogaster* wild-type strains, human, A. thaliana, and *P. reticulata*. All species were analyzed with two RNA-seq replicates. The analysis was performed with --threads 32 cores and 64Gb RAM (--ram 64 for Mode 2 --assembly).

In both ChimeraTE Modes, the main evidence used to detect chimeric transcripts is the presence of chimeric reads, which are paired-end reads spanning between TE and gene sequences. In Mode 1, at least 50% of one read (default parameter) from the read pair must align against the TE insertion, whereas in Mode 2 the whole read must align against a TE copy. Therefore, the alignment method performed by Mode 2 does not allow the detection of chimeric transcripts derived from TE copies shorter than the read size. Hence, we investigated whether chimeric transcripts found with Mode 1, but not with Mode 2, are generated by TE insertions shorter than the sequenced reads. From 237 chimeric transcripts predicted only by Mode 1, we found that 77 (~32%) have TEs shorter than the read size (120 nt), making it impossible to detect them with Mode 2 (Figure S1.1B). It is important to highlight that TEs longer than reads may have splice sites, generating chimeric transcripts with a small TE-derived fragment, not being detected by Mode 2 as well. Furthermore, we observed from Mode 2 results that 121 (~57%) chimeric transcripts from TEs longer than the reads have coverage lower than 10 chimeric reads. We hypothesized that Mode 1 may have substantially more TE- aligned reads than Mode 2, due to the use of strain-specific TE insertions and by counting split read alignment.

ChimeraTE Mode 1 is dependent on a reliable genome annotation for both genes and TEs, contrary to Mode 2. Indeed, we found in total, seven chimeric transcripts detected only with Mode 2 (Table S1.2), which are derived from genes that were not annotated in the wild-type genomes. We compared them with the *dm6* genome, and we found all seven have the predicted TE family inserted near/inside the gene, reinforcing these annotations (Figure S1.2).

For instance, the chimeric transcript *CG3164-McClintock* was detected by Mode 2 with double evidence in the four wild-type strains, and it is not annotated in any of the four genomes, perhaps due to its location in the telomeric region of the 2L chromosome (Figure S1.2G).

dmgoth101		**dmgoth63**		**dmsj23**		**dmsj7**	
Gene	**TE**	**Gene**	**TE**	**Gene**	**TE**	**Gene**	**TE**
FBgn0287182	transib2	FBgn0003016	jockey	FBgn0286778	HMS-Beagle	FBgn0042127	roo
FBgn0287183	roo	FBgn0287182	transib2	FBgn0287183	rover	FBgn0287182	transib2
FBgn0288229	McClintock	FBgn0288229	McClintock	FBgn0288229	McClintock	FBgn0288229	McClintock
FBgn0286778	roo	FBgn0286778	HMS-Beagle	FBgn0287478	roo		
		FBgn0287478	roo				

Table S1.2: Chimeric transcripts found by Mode 2, but not by Mode 1, due to lack of genomic annotation.

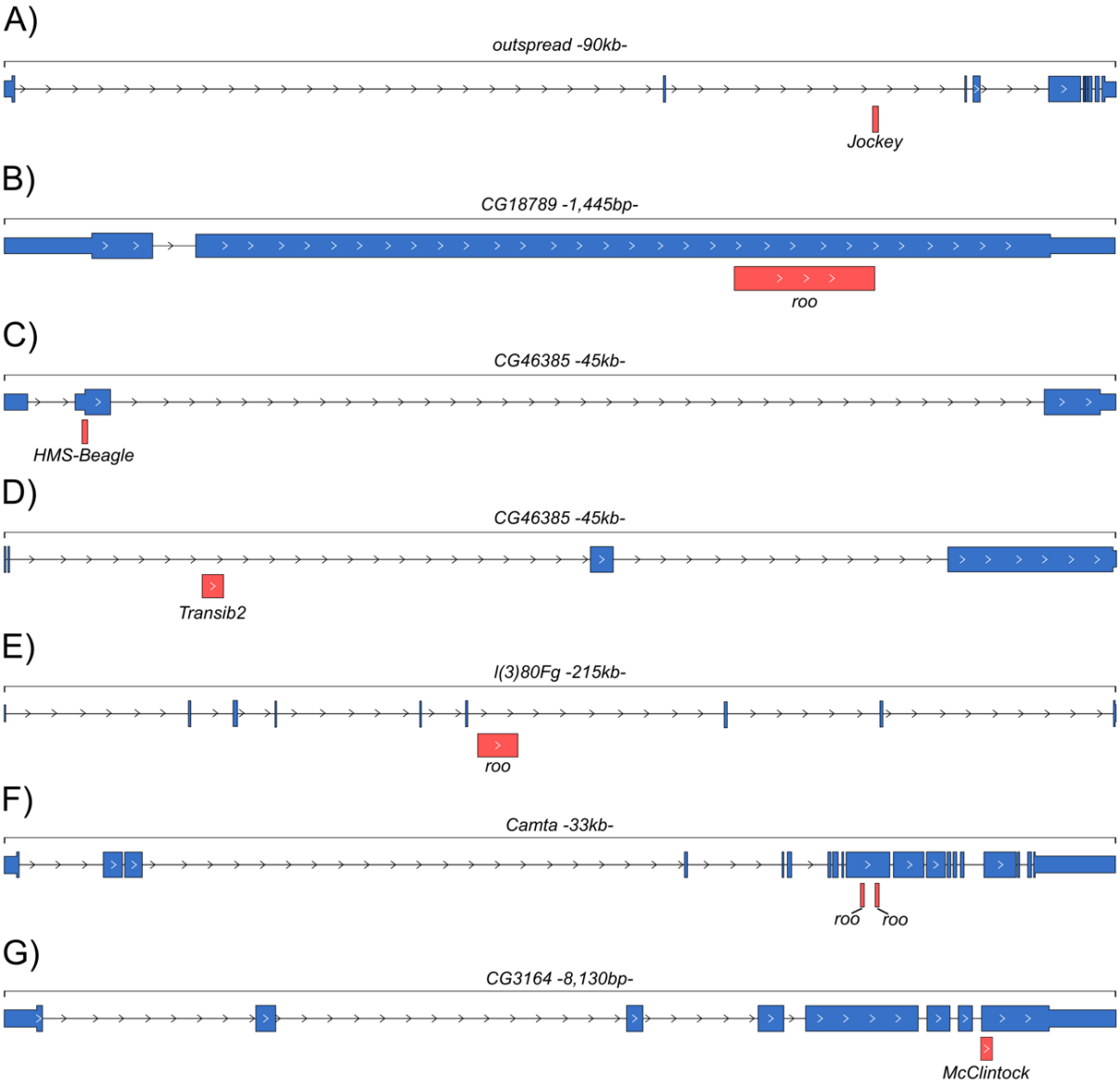


Figure S1.2: Gene and TE annotation from *dm6* genome for chimeric transcripts identified by Mode 2 corresponding to genes that were not annotated in the wild-type genome assemblies. Blue boxes: UTRs and CDSs; Red boxes: TE insertion identified by ChimeraTE as the TE family generating the chimeric transcript.

ChimeraTE Mode 1 and Mode 2 detected 327 and 324 genes producing chimeric transcripts respectively, of which 165 were found by both methods. In Mode 2, chimeras with TE-derived sequences smaller than the read length are not detected. However, Mode 2 can detect TEs that are absent from the reference genome, along with low-frequency TE insertions of the wild-type strains at the population level, which are not present in the assembled Nanopore genome. Since low-frequency Nanopore reads are discarded during the genome assembly (54), Mode 1 is not able to detect them, whereas Mode 2 can. Furthermore, Mode 1 detects chimeric transcripts derived from TEs inside or +/- 3 kb from genes. TEs located farther away from genes and involved in chimeric transcripts are not detected. Such chimeras have been reported in *D. melanogaster* (31), as well as TEs acting as distal *cis*-regulatory elements (70, 71). Therefore, although we could consider them as potential false positives all cases in which the TE was not found inside/near a gene in the manual curation, we speculate that part of these cases found only by Mode 2 might be either from low-frequency TEs in the pool of individuals sequenced, or from TEs located far from genes.

Replicability and coverage of chimeric transcripts in both ChimeraTE modes

The ability to detect TE expression is an important factor to identify chimeric reads. The different strategies of alignment to identify chimeric reads in Mode 1 and Mode 2 cause differences in the sensitivity of chimeric transcript detection (Figure 2.1A). We investigated whether such differences might be associated with the detection of TE-derived reads. We found that Mode 1 is more efficient than Mode 2 in detecting reads aligned against TE insertions (Figure S2.2), as well as for chimeric read detection (Figure 2.1A). Indeed, the proportion of chimeric reads in Mode 1 and Mode 2 is 0.34% and 0.04% of the total library sizes, respectively. The power of chimeric read detection from the two Modes is different because of the alignment strategies used. Despite the differences, both modes show significant positive correlations between the library size 45and the number of chimeric reads, as well as the number of TE-aligned reads and chimeric reads (Figure S2.3).

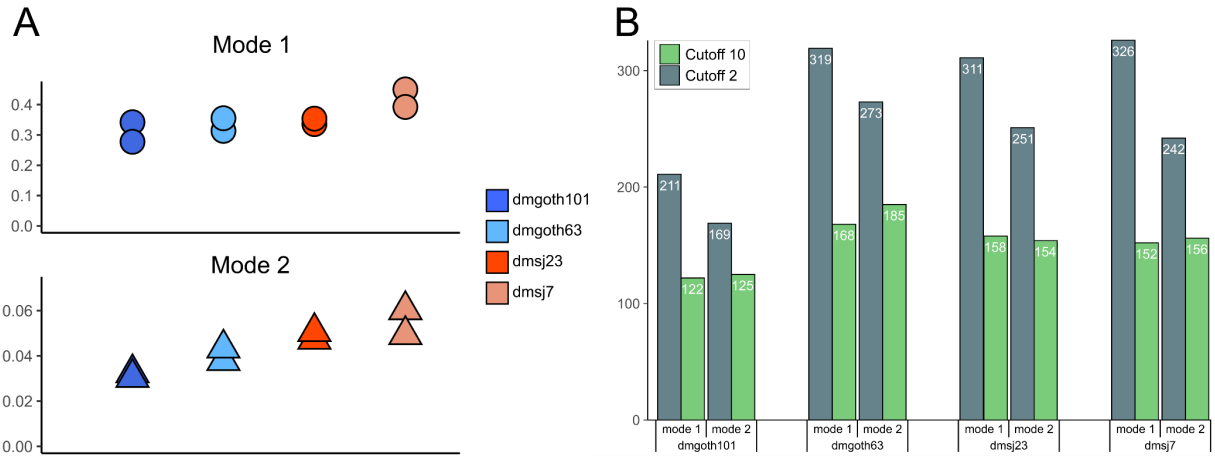


Figure S2.1: **A**) Proportion of reads from RNA-seq libraries comprised by chimeric reads on both replicates. **B**) Chimeric transcripts detected by both modes of ChimeraTE. The bars represent chimeras found in both RNA-seq replicates in the four *D. melanogaster* wild-type strains, applying two chimeric read cutoffs: 2 and 10.

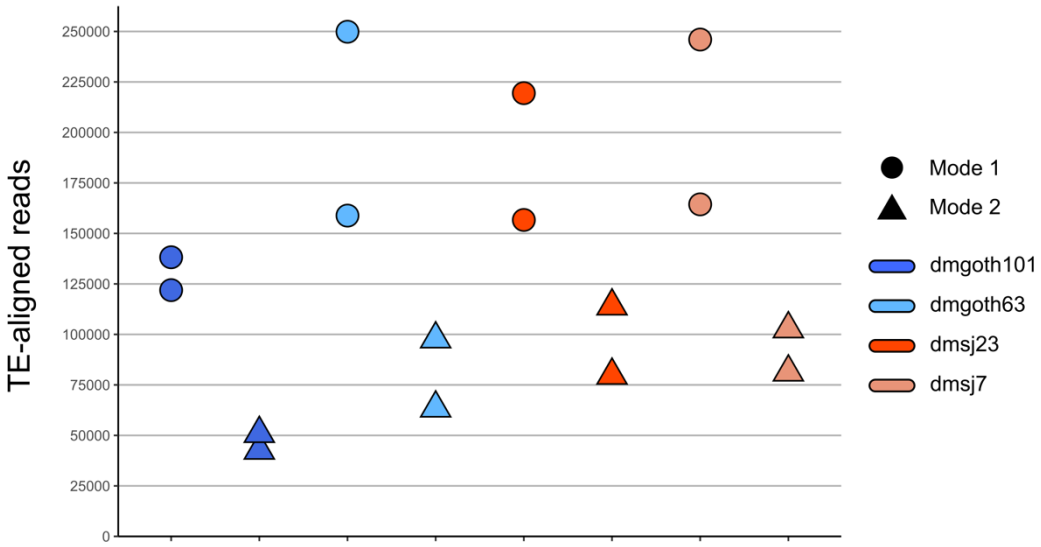


Figure S2.2: The total number of TE-aligned reads between both ChimeraTE Modes, in all strains and their respective replicates.

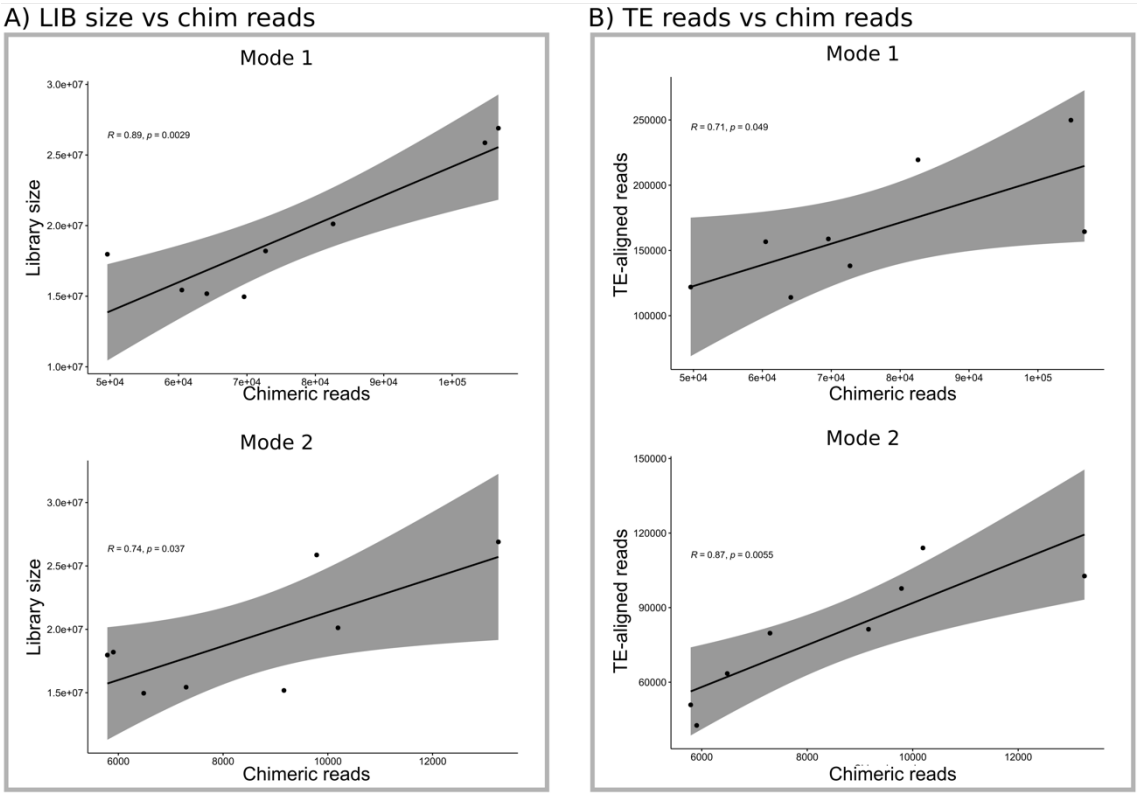


Figure S2.3: Positive Person correlations between: A) RNA-seq libraries size and the total of chimeric reads detected in Mode 1 (top) and Mode 2 (bottom); B) TE-aligned reads and the total of chimeric reads detected in in Mode 1 (top) and Mode 2 (bottom).

To quantify ChimeraTE replicability between RNA-seq samples in both modes, as well as the impact of changes in the coverage thresholds, we performed a comparative analysis of chimeric transcripts using two thresholds of chimeric reads, 2 and 10. Overall, Mode 1 finds 291 and 150 chimeric transcripts per strain in both replicates (Figure 2.1B), using thresholds 2 and 10 respectively, whereas Mode 2 obtains 233 and 155. These results show that by increasing the chimeric read thresholds from 2 to 10, there is a decrease of 48.45% and 33.48% in the number of detected chimeric transcripts in Mode 1 and Mode 2. Therefore, even when found in both replicates, a substantial amount of chimeric transcripts is detected with low chimeric read coverage by both modes.

Artifactual chimeric reads are largely reduced with RNA-seq replicability

In both Mode 1 and Mode 2, only chimeric transcripts found in both RNA-seq replicates were considered true chimeras. Chimeric transcripts found in only one replicate may exist due to pervasive transcription in one of the replicates, the lack of read coverage in one of the replicates to predict it, or they could be artifacts of cDNA library preparation, such as template switching and hybridization of templates followed by chimeric elongation (72-74). In addition, although comprising low rates with Illumina's traditional bridge amplification, artifactual

chimeric reads might be generated due to index hopping in multiplexed sequencing (37,75). To quantify cases that may be artifacts, we aligned the RNA-seq from the four wild-type strains against their masked genomes to identify paired-end reads where each mate maps to genes from different chromosomes, or were aligned with an insert size > 500kb, and therefore are artifacts. It is important to highlight that these alignments can be inflated because they do not differentiate artifactual reads produced during sequencing, from biological fusion gene products derived from the breakage and re-joining from different chromosomes, chromosomal rearrangements (76-77) and also to reads mapping to different paralog genes or genes from the same gene family sharing high similarity. In replicate 1 and replicate 2, we found ~50,522 and ~42,303 chimeric reads representing fusions of genes that are in different chromosomes, and ~7,509, ~5,334 farther away in the genome, with at least one chimeric read (Sup. Table S3.1) Such fusions were composed in average by ~5,966 genes from different chromosomes and ~3,234 genes farther away. However, we found only 498 (8.35%) of the artifactual chimeras from different chromosomes, and 212 (6.56%) genes farther away in both RNA-seq replicates with >= 2 chimeric reads (Table S3.1). In addition, these fusion genes generated by artifacts have high expression level (average FPKM = 722 per strain), reinforcing the hypothesis that high expressed genes are more likely to produce artifacts (72). Therefore, artifactual chimeric reads present in more than one replicate, and derived from genes and TEs, might have even lower prevalence, since the proportion of RNA-seq reads derived from TEs is very low (Figure S2.1A) in comparison to genes. Thus, in order to significantly reduce false positives, it is strongly recommended to use RNA-seq replicates, since artifactual chimeric reads exist in high frequency in highly expressed genes but only in one RNA-seq replicate. On the other hand, the proportion of the same chimeras found in more than one RNA-seq replicate with >= 2 chimeric reads is low (212 chimeras).

	Replicate 1				**Replicate 2**				**Both replicates**			
	Diff. chr.		***> 500kb***		***Diff. chr.***		***> 500kb***		***Diff. chr.***		***> 500kb***	
	Reads	*Genes*	*Reads*	*Genes*	*Reads*	*Genes*	*Reads*	*Genes*	*Reads*	*Genes*	*Reads*	*Genes*
dmgoth101	99714	8009	18718	6155	19495	4430	1719	1523	6308	665	1422	256
dmgoth63	48539	6057	4361	2894	33556	5856	3558	2744	7301	314	1169	169
dmsj23	40764	6613	5844	3601	31227	6038	4534	2990	5151	489	1865	294
dmsj7	13072	3112	1113	1022	84934	7621	11525	4943	6950	527	842	129
Average	50522	5947	7509	3418	42303	5986	5334	3050	6427	498	1324	212

Table S3.1: Amount of artifactual chimeric reads (mates mapped in different chromosomes; or farther away > 500kb) in the two replicates, and presence of the same artifacts in both RNA-seq replicates. Green cells represent the average of genes found in both replicates with >= 2 chimeric reads.