Evolutionary genomics of the Neotropical Drosophila saltans species group (Diptera Drosophilidae)

Carolina Prediger

THÈSE DE DOCTORAT

NNT : 2023UPASL137

Evolutionary genomics of the Neotropical *Drosophila saltans* species group (Diptera: Drosophilidae)

Génomique évolutive du groupe d'espèces néotropical *Drosophila saltans* (Diptera : Drosophilidae)

Thèse de doctorat de l'université Paris-Saclay et de São Paulo State University

École doctorale n°577 : structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Évolution
Graduate School : Sciences de la vie et santé.
Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **UMR EGCE** (Université Paris-Saclay, CNRS, IRD), et **Department of Biology** (UNESP - São Paulo State University), sous la direction d'**Amir YASSIN**, Chargé de recherche, et la co-direction de **Lilian MADI-RAVAZZI**, Professeure associée.

Thèse soutenue à São José do Rio Preto (Brésil), le 04 décembre 2023, par

Carolina PREDIGER

Composition du Jury

Membres du jury avec voix délibérative

Didier CASANE Professeur, Université Paris Cité et Université Paris-Saclay	Président
Emmanuelle LERAT Chargée de recherche (HDR), Université de Lyon	Rapporteur & Examinatrice
Sergio Russo MATIOLI Professeur associé (equiv. HDR), Universidade de São Paulo	Rapporteur & Examinateur
Guillaume ACHAZ Professeur, Université Paris Cité	Examinateur
Cesar MARTINS Professeur, Universidade Estadual Paulista	Examinateur
Hermione E.M.C. BICUDO Professeure émérite, Universidade Estadual Paulista	Examinatrice

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Génomique évolutive du groupe d'espèces néotropical *Drosophila saltans* (Diptera : Drosophilidae)

Mots clés : Discordance Phylogénomique. Biogéographie Historique. Conflits Cyto-Nucléaires. Biais d'Usage du Code. Évolution du Génome. Tri de Lignées Incomplet.

Résumé : La région néotropicale est connue pour sa biodiversité, stimulée par la diversité des environnements qui ont favorisé l'évolution d'un large éventail d'espèces. De nombreuses études visent à découvrir les facteurs contribuant à la génération et à la préservation de la biodiversité, mais il existe une notable pénurie de recherches axées sur l'évolution génomique spécifique de la région néotropicale. Cette thèse se concentre sur le groupe d'espèces Drosophila saltans, qui comprend 23 espèces décrites, classées en cinq sous-groupes, à savoir : saltans, sturtevanti, parasaltans, cordata et elliptica. La classification repose principalement sur des caractères des organes génitaux mâles et a été confirmée à plusieurs reprises à l'aide d'autres marqueurs, mais les relations au sein de chaque sous-groupe et entre eux demeurent incertaines. En fait, les espèces du groupe saltans présentent des niveaux variables d'isolement reproductif et un motif distinct d'utilisation des codons par rapport à d'autres espèces du sous-genre Sophophora. Ces caractéristiques posent un défi dans le déchiffrement de l'évolution du groupe saltans. Pour résoudre cette question, deux aspects ont été privilégiés : les relations phylogénétiques au sein du groupe à l'aide de données génomiques et l'impact du biais d'utilisation des codons sur l'inférence phylogénétique. Dans la première partie, les relations phylogénétiques ont été reconstruites à l'aide de données génomiques nouvellement générées pour 16 espèces. L'analyse, tout en révélant un schéma cohérent des relations entre les sous-groupes, a identifié quelques conflits mineurs entre les autosomes et le chromosome X, ainsi qu'entre les génomes nucléaires et mitochondriaux. Pour quantifier le niveau d'incongruence génomique au sein de ces groupes, un nouveau test appelé 2A2B a été développé et appliqué. Le test a révélé des taux élevés de réticulation au sein du groupe saltans, avec trois points majeurs de réticulation, notamment au sein des sous-groupes sturtevanti et saltans, ainsi que le branche saltans-cordata-elliptica. Cependant, aucune preuve de réticulation n'a été trouvée dans tous les sous-groupes, comme c'était le cas avec le sous-groupe elliptica. Les modèles de réticulation ont montré des corrélations exponentielles avec les taux de spéciation et les plages géographiques ancestrales qui se chevauchent. La dernière partie de la thèse a exploré d'éventuels changements dans les modèles d'utilisation des codons dans la famille Drosophilidae. Pour cette analyse, les modèles d'utilisation des codons dans 3285 gènes uniques sur 174 génomes ont été examinés, en mettant l'accent sur le clade saltans-willistoni, car des analyses antérieures avec un nombre limité de gènes et un seul génome suggéraient un changement d'utilisation des codons dans ce clade. Le changement d'utilisation des codons a été confirmé, ce qui n'a pas été observé dans son clade sœur, Lordiphosa. Notamment, en utilisant le premier, le deuxième ou le troisième nucléotide de chaque codon dans les analyses phylogénétiques, des changements significatifs de position des branches étaient principalement évidents dans le sous-groupe saltans. Dans l'ensemble, ces résultats ont permis de résoudre les relations phylogénétiques entre les espèces du groupe saltans et d'identifier d'éventuels facteurs historiques et moléculaires ayant influencé la diversification d'un clade néotropical important mais largement méconnu. Ces découvertes renforcent notre compréhension de l'évolution de la biodiversité dans la région néotropicale et introduisent un nouveau modèle pour évaluer la base génétique de la spéciation tropicale.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Evolutionary genomics of the Neotropical *Drosophila saltans* species group (Diptera: Drosophilidae)

Keywords : Phylogenomic Discordance. Historical Biogeography. Cyto-Nuclear Conflicts. Codon Usage Bias. Genome Evolution. Incomplete Lineage Sorting

Abstract : The neotropical region is known for its biodiversity, driven by the diversity of environments that have promoted the evolution of a wide range of species. Many studies aim to discover the factors contributing to generating and sustaining biodiversity, but there is a notable scarcity of research focused on the specific genomic evolution of the neotropical region. This thesis focuses on the Drosophila saltans species group, which includes 23 described species, classified into five subgroups, namely: saltans, sturtevanti, parasaltans, cordata, and elliptica. The classification was primarily based on male genitalia characters and has been confirmed multiple times using other markers, but the relationships within and between each subgroup remain uncertain. In fact, species in the saltans group exhibit varying levels of reproductive isolation and a distinct codon usage pattern compared to other species in the subgenus Sophophora. These features pose a challenge in deciphering the evolution of the saltans group. To address this issue, two aspects were focused on: the phylogenetic relationships within the group using genomic data and the impact of codon usage bias on phylogenetic inference. In the first part, phylogenetic relationships were reconstructed using newly generated genomic data for 16 species. The analysis, while revealing a consistent pattern of relationships among the subgroups, identified some minor conflicts between autosomes and the X chromosome, as well as between nuclear and mitochondrial genomes. To quantify the level of genomic incongruence within these groups, a new test called 2A2B was developed and applied. The test revealed high rates of reticulation within the saltans group, with three major reticulation points, namely within the sturtevanti and saltans subgroups, as well as along the saltans-cordata-elliptica branch. However, no evidence of reticulation was found within all subgroups, as was the case with the elliptica subgroup. Reticulation patterns showed exponential correlations with speciation rates and overlapping ancestral geographical ranges. The last part of the thesis explored possible changes in codon usage patterns in the Drosophilidae family. For this analysis, codon usage patterns in 3285 single-copy genes across 174 genomes were investigated, with a focus on the saltans-willistoni clade, as earlier analyses with a limited number of genes and a single genome suggested codon usage change in this clade. Codon usage change was confirmed, which was not observed in its sister clade, Lordiphosa. Notably, when using the first, second, or third nucleotide of each codon in phylogenetic analyses, significant branch position changes were mainly evident in the saltans subgroup. Together, these results resolved the phylogenetic relationships among species in the saltans group and identified possible historical and molecular factors influencing the diversification of an important yet largely understudied neotropical clade. These findings enhance our understanding of how biodiversity evolves in the neotropical region and introduce a new model for evaluating the genetic basis of tropical speciation.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Título : Evolução genômica do grupo de espécies Neotropical *Drosophila saltans* (Diptera: Drosophilidae)

Palavras chaves: Discordância Filogenômica. Biogeografia Histórica. Conflitos Citonucleares. Viés na Utilização de Códons. Evolução de Genomas. Incomplete lineage sorting.

Resumo : A região neotropical é conhecida por sua biodiversidade, impulsionada pela vasta gama de ambientes que promoveram a diversificação de espécies. Muitos estudos têm como objetivo descobrir os fatores que contribuem para gerar e manter a biodiversidade, mas há uma notável escassez de pesquisas focadas na evolução de espécies neotropicais. Esta tese se concentra no grupo *saltans* de *Drosophila*, que apresenta 23 espécies descritas, classificadas em cinco subgrupos, sendo eles: *saltans*, *sturtevanti*, *parasaltans*, *cordata* e *elliptica*. A classificação foi proposta principalmente com base em caracteres da genitália masculina e foi confirmada várias vezes usando outros marcadores, mas as relações entre e dentro de cada subgrupo permanecem incertas. De fato, as espécies do grupo *saltans* exibem níveis variados de isolamento reprodutivo e um padrão distinto de uso de códons em comparação com outras espécies do subgênero *Sophophora*. Essas características representam um desafio na decifração da evolução do grupo *saltans*. Para abordar essa questão, dois aspectos foram focados, as relações filogenéticas dentro do grupo usando dados genômicos e o impacto do viés de uso de códons na inferência filogenética. Na primeira parte, as relações filogenéticas foram reconstruídas usando dados genômicos recém-gerados para 16 espécies. A análise, embora tenha revelado um padrão consistente de relações entre os subgrupos, descobriu alguns conflitos menores entre os autossomos e o cromossomo X, bem como entre os genomas nuclear e mitocondrial. Para quantificar o nível de incongruências genômicas dentro desses grupos, um novo teste chamado 2A2B foi produzido e aplicado. O teste revelou altas taxas de reticulação dentro do grupo *saltans* e com 3 pontos de maior reticulação, sendo eles dentro dos subgrupos *sturtevanti* e *saltans*, bem como no ramo *saltans-cordata-elliptica*, contudo não foi encontrada evidências de reticulação em todos os subgrupos, caso do subgrupo *elliptica*. Os padrões de reticulação mostraram correlações exponenciais com taxas de especiação e sobreposição de faixas geográficas ancestrais. A última parte da tese explorou possíveis mudanças nos padrões de uso de códons na família Drosophilidae. Para essa análise, os padrões de uso de códons em 3285 genes de cópia única em 174 genomas foram investigados, com foco no clado *saltans-willistoni*, vez que analises com poucos genes e um único genoma sugeriam mudança no uso de códons neste clado. A mudança no uso de códons foi confirmada, o que não foi observado em seu clado irmão, *Lordiphosa*. Notavelmente, ao utilizar o primeiro, segundo ou terceiro nucleotídeo de cada códon em análises filogenéticas, alterações significativas na posição dos ramos foram principalmente evidentes no subgrupo *saltans*. Em conjunto, esses resultados esclarecem significativamente as relações filogenéticas entre as espécies do grupo *saltans* e identificaram possíveis fatores históricos e moleculares que influenciaram a diversificação de um clado neotropical importante, mas em grande parte pouco estudado. Essas descobertas aprimoram nossa compreensão de como a biodiversidade evolui na região neotropical e introduzem um novo modelo para avaliar a base genética da especiação tropical.

Acknowledgments

This work was conducted with the support of the National Council for Scientific and Technological Development (CNPq), grant number 141545/2020-8, from the France Excellence Eiffel scholarship program, for which I am grateful. The Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant numbers 95/06165-1, 2014/14059-0, and 2016/11994-5, as well as the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), also supported this research, and I would like to acknowledge them here.

I would like to express my sincere appreciation to my advisers. Firstly, I would like to thank Prof. Dr. Lilian Madi-Ravazzi for entrusting me with the opportunity to work on this project. It was a significant undertaking, and I'm grateful for the trust you placed in me, even though we were strangers at the time. Secondly, I want to express my heartfelt gratitude to Dr. Amir Yassin. Working with you has been a dream come true, and your brilliance, guidance, and inspirational mentorship have been invaluable. I acknowledge the contributions of both of my advisers to my academic journey.

To my coauthors, namely, Erina, Samara, Aurélie, Lisa and Wolfgang, I thank you for your valuable contributions, also I would like to acknowledge David, for all the help with the server and software, thank you for help me in desperate times.

To my colleagues at the Laboratory of Evolutionary Biology in Insects, Dr. Bruna and M.Sc. Guilherme, you were instrumental in keeping me from giving up, and together with Natalia, Rodrigo, and Nathalia, you made the environment at UNESP much more welcoming. I cannot forget to mention Mr. Sebastião, for his tireless work in preparing culture media. Thank you very much!

To my colleagues at the Laboratoire EGCE, especially Dr. Erina, Dr. Cécile, Jérémy, who shared the office with me, my sincere thanks for welcoming me so warmly, helping me navigate all the bureaucracy as a foreigner in France, and for the raclette dinners and trips to Paris. Thank you, you will always hold a special place in my heart. I also mustn't forget Melissa, Benjamin, Loïc, Simon, Samuel, Etienne, Marie, Júlia, Antoine, Lohan, Pénélope for our conversations, invitations to various activities, and for introducing me to the French way of life, as well as for bearing with my limited fluency in French. To Beatriz and Silvie, who showed great patience with my attempts to express myself in French. Thank you for everything. To Isabelle, Veronique, Michael, and Fréderic for all the ideas, for all the behavior pole where I witnessed the power of cooperation.

I'd like to thank my friends, especially Grègory, Michel, Ayoub, Bruna, Marina, Gyovanna, and Bianca, with whom I lived at some point during my doctoral journey. Thank you for being there to listen me and also for all the amazing time we spend together.

To the friends I made in São José do Rio Preto and in Ilê-de-France, I won't mention you all by name, but please know that you were essential in helping me reach the end of this cycle. I love you all.

Finally, I want to express my gratitude to my family, my mother Vera, my father Reneu, my siblings Vanessa and Ricardo. Thank you for believing in me, often without fully understanding what I was doing. You provided the support I needed, embraced my anxieties, and pushed me when I needed that extra push. You mean everything to me. I love you all very much.

ILLUSTRATIONS LIST

1. INTRODUCTION

3. ORIGINAL ARTICLE I

4. ORIGINAL ARTICLE II

Figure 1. Investigating codon usage bias across Drosophilids reveals potential shifts in in CU evolution. The *saltans-willistoni* clade, in particular, stands out with ENC values much higher than those observed in the *Sophophora* and higher in the *Drosophila* subgenera, indicating a distinct lack of codon usage preference across all species within this clade. Notably, our analysis also uncovers a contrasting shift in ENC values during the evolutionary trajectory of the *Zaprionus* genus. In this context, *Z. bogoriensis* emerges as a striking outlier, displaying a pronounced trend towards codon bias that contradicts the broader patterns observed within its genus. Furthermore, it's worth noting that non-Drosophilids and basal clades of Drosophilidae tend to exhibit higher ENC values, suggesting a consistent trend towards a lack of codon usage bias in these lineages. The evolutionary relationship between this species is shown by a phylogenetic tree constructed using a maximum likelihood approach based on 192 single-copy genes. 59

Figure 2. Relative synonymous codon usage (RSCU) analysis shows that the codon usage pattern the neotropical *Sophophora* clade is more similar with the pattern seen for the ancestral of drosophilids, and indicate a codon usage shift in the *Zaprionus* genus. The 61 columns represent he non-stop codons. Rows correspond to the 174 genomes that have been evaluated, darker colored cells correspond to the codons of tRNAs that are favored and lighter cells the unfavored tRNAs-codons. Codons and species were clustered using hierarchical clustering by RSCU values. 63

Figure 3. Correspondence analyses of the average relative synonymous codon usage between species recover 3 major clusters, the I - neotropical *Sophophora,* II - *Sophophora – old world* and III *Lordiphosa* and *Zaprionus genera* and *Drosophila*, *Siphlodora*, *Dorsilopha* subgenera. The plot shows each of the 174 genomes examined in this study along the first two dimensions (the X and Yaxes) of a correspondence analysis. Each axis is labeled with the percent variance explained by the corresponding dimension. The the codons correspondence analysis and each codon contribution is seen in Supplementary Figure S1). 65

Figure 4. Influence of Mutational Bias on Codon Usage in *Drosophila*. The relationship between GC3 and ENC values reveals the impact of mutational bias, even in species with low Effective Number of Codons (ENC), as exemplified by *D. ironensis* (A). In contrast, species with higher ENC values, like *D. sturtevanti* (B), exhibit a less clear pattern. Notably, the percentage of genes deviating from the expected only by mutational bias (C) suggests that, despite a influence of selection, the neotropical *Sophophora* species tend to have higher ENC values due to their GC3 content approaching 50%. The reddish points indicate a higher likelihood of selection influence, while the bluish points signify a lower impact of selection on codon usage. 70

Figure 5. Comparison of topologies generated with amino-acids, first+second, first, second and third codon bases focusing in the neotropical *Sophophora* clade. 74

APPENDIX A: Supplementary material Chapter 3

Supplementary Table S1. Summary of previous competing phylogenetic hypotheses in the *saltans* group. CO = *cordata* subgroup, EL = *elliptica* subgroup, ST = *stutevanti* subgroup, PA = *parasaltans* subgroup, SA = *saltans* subgroup, aus= *D. austrosaltans*, nig = *D. nigrosaltans*, sal = *D. saltans*, pro = *D. prosaltans*, lus = *D. lusaltans*, sep = *D. septentriosaltans*, pse = *D. pseudosaltans*, stu= *D. sturtevanti*, leh = *D. lehrmanae*, mil = *D. milleri*, dac = *D. dacunhai*, nsa = *D. neosaltans*, nel = *D. neoelliptica*, ema= *D. emarginada*, OS = overall similarities, ai = measure of isolation for each interspecific cross, MP = maximum parsimonia, ML = maximum likelihood, BI =Bayesian inference. 95

TABLE LIST

ABBREVIATION LIST

Adh	Alcohol dehydrogenase
BI	Bayesian inference
COI	Cytochrome c oxidase subunit I
COII	Cytochrome c oxidase subunit II
ENC	Effective Number of Codons
GC3	Guanine-Cytosine ending codon
ILS	Incomplete lineage sorting
ITS1	Internal transcribed spacer 1
ML	Maximum likelihood
MP	Maximum parsimony
myr	million years
RSCU	Relative synonymous codon usage
Xdh	Xanthine dehydrogenase

TABLE OF CONTENTS

1 INTRODUCTION

Understanding the evolutionary relationships among species and the processes that drive their diversification is a fundamental goal in biology. Recent advancements in genomic research have shed light on the intricate patterns of evolution, revealing the significance of reticulation events (section 1.1), and also in genome evolution processes such as codon usage (section 1.2). A good model to investigate these matters is the *saltans* species group, which was chosen as the model in this thesis (section 1.3).

1.1 Reticulation events and their impact on evolutionary patterns

Traditionally, the tree-like representation of species evolution has dominated our understanding. However, several processes such as gene flow, horizontal gene transfer, and incomplete lineage sorting (ILS) do not fit within a bifurcating tree (SCORNAVACCA; DELSUC; GALTIER, 2020). Depending on the extent of these events in a species' history, the recovery of a true evolutionary dichotomous species tree may not be possible. Reticulation events can blur the boundaries between species and generate complex genetic networks. This recognition has been long accepted in plant evolutionary studies (ANDERSON; STEBBINS, 1954), but its importance is now extended to other taxonomic groups as well (HALLSTRÖM; JANKE, 2010; MORGAN et al., 2013; LI et al., 2016; SUH, 2016; MALLET; BESANSKY; HAHN, 2016; SUVOROV et al., 2022; HIME et al., 2021; OWEN; MILLER, 2022).

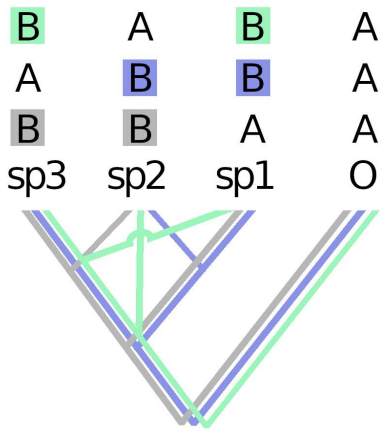
Gene flow, the transfer of genetic information between different populations or species, plays a significant role in shaping evolutionary processes. Two important concepts within gene flow are hybridization and introgression. Hybridization occurs when individuals from different populations/species mate and produce hybrid offspring. Whereas introgression refers to the transfer of genetic material from one species to another through hybridization. (TWYFORD; ENNOS, 2012). There are multiple mechanisms that prevent admixture, such as Dobzhansky–Muller hybrid incompatibilities, hybridization load and multiple forms of selection against hybrids (see MORAN et al., 2021). Moreover, different regions of the genome present more or less resistance to introgression (QVARNSTRÖM; BAILEY, 2009; ELLEGREN, 2009; SANKARARAMAN et al., 2016; SEIXAS; BOURSOT; MELO-FERREIRA, 2018; CHARLESWORTH; CAMPOS; JACKSON, 2018; MAI et al., 2020; MATUTE et al., 2020; MORAN et al., 2021; REILLY et al., 2022; SKOV et al., 2023).

In addition to the complexities introduced by reticulation events such as introgression and hybridization, another phenomenon that challenged the reconstruction of bifurcating trees

is ILS. ILS occurs when ancestral genetic variation is not completely sorted into separate lineages during speciation, and the evaluated markers fail to coalesce before the speciation events. It can result in the presence of shared ancestral polymorphisms among closely related species, leading to incongruence between gene trees and species trees (DEGNAN; ROSENBERG, 2009). ILS is affected by ancestral population size and is intensified with high speciation rate, *i.e.* multiple speciation events separated by short branches lengths.

Distinguishing between ILS and introgression can be challenging due to the similar genetic patterns they can produce. To differentiate between these processes, researchers often utilize tests that analyze bi-allelic sites within quartets comprising three species and one out-group, presuming the existing of a known bifurcating tree. Basically, when we analyze bi-allelic sites (here, A and B) in species quartets we can see three topologies, frequently referred to as AABB, ABBA and BABA in reference to the distribution of the informative bi-alelic sites found in the evaluated quartet (out group, (species 1, (species 2, species 3))) (Figure 1).

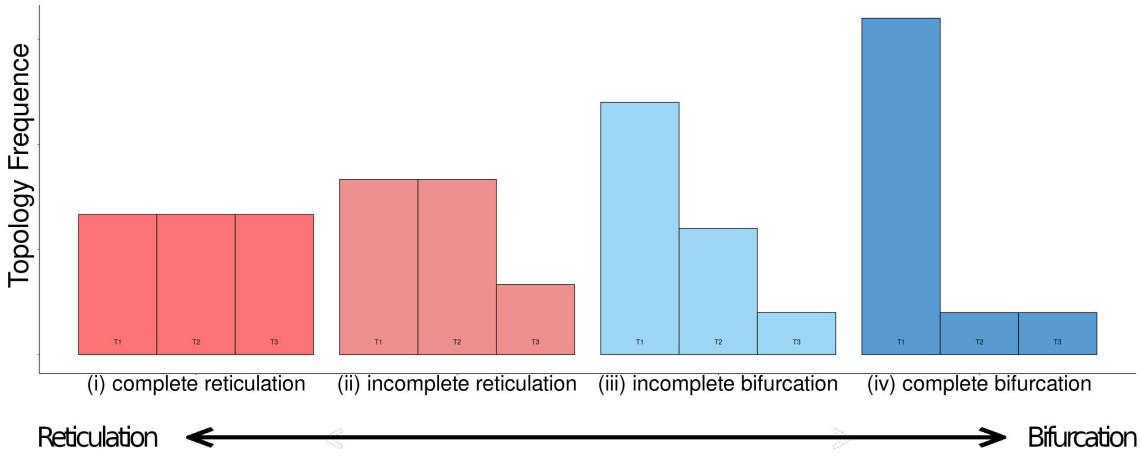
Figure 1. Potential topological structures in species quartets with an outgroup, inferred from shared genetic variation in Pairs. Three Distinct Topologies are possible: AABB (gray, species 2 and 3 as the closest relatives), ABBA (blue, species 1 and 2 as the most closely related), and BABA (green, species 1 and 3 as the nearest pair).



The distribution of frequencies of these three topologies across a genome fragment can lead to 4 possible classes (Figure 2): (i) complete reticulation, where all topologies (AABB, ABBA and BABA) have equal frequencies, this case is expected to occurs in cases of high ILS, expected in cases as adaptive radiation; (ii) incomplete reticulation, where two topologies significantly exceed the third one but do not differ significantly from each other, expected to occurs in cases of bidirectional exchange of genetic material between species/populations, as seen in cases of full hybridization; (iii) incomplete bifurcation, where

the proportions of all topologies significantly differ, frequently interpreted as a signal of introgression; and (iv) complete bifurcation, where one topology significantly exceeds the other two, with the latter having nearly equal proportions.

Figure 2. Four possible classes of topological frequency distributions across the genome. These classes offer insights into diverse evolutionary scenarios: class I, complete reticulation, expected in genome regions impacted by ILS; class II, bidirectional exchange of genetic material, interpreted as a signal of hybridization; class III, asymmetric exchange of genetic material, interpreted as introgression; and class IV, complete bifurcation, expected in events of speciation without ILS and gene flow.



One widely used test is the Patterson's D, frequently called ABBA-BABA test. This test requires the knowledge of a true-bifurcating tree (AABB) and it focuses on cases (iii) and (iv), and it is based on the fact that in the absence of gene flow and presence of ILS, the frequencies of ABBA and BABA are not significantly different, but in presence of introgression the frequency of ABBA and BABA are going to be significantly different (DURAND et al., 2011; PATTERSON et al., 2012). A different test known as HyDe offers a way to quantify admixture by examining the ratio of shared alleles, which ranges from 0 (indicating complete isolation) to 0.5 (indicating complete hybridization). HyDe is capable of addressing case (ii) mentioned earlier and employs a normal approximation method to assess the significance of the results (BLISCHAK et al., 2018; KUBATKO; CHIFMAN, 2019). Lately, a site-based test based on the χ^2 statistic has been introduced to investigate case (i) by examining the deviation of parity among the three topologies. This test employs χ^2 statistics to evaluate the significance of the observed deviations (SAYYARI; MIRARAB, 2018) However, despite these advancements, a unified test that can comprehensively assess the

prevalence of each of the four categories across the genome and a phylogenetic tree is still lacking.

1.2 Forces that shape codon usage bias

Codon usage bias, which refers to the non-random usage of synonymous codons in protein-coding genes (FIERS et al., 1976; SHARP et al., 1997; GRANTHAM et al., 1981), can pose challenges for phylogenetic inference (INAGAKI; ROGER, 2006; LI et al., 2014). Similarities with the codon usage pattern within closer related species seems to be the general pattern, considering the high phylogenetic signal that has been seen for different taxa (MILLER et al., 2017; LABELLA et al., 2019; KOKATE; TECHTMANN; WERNER, 2021). However, in certain clade-specific branches of the tree of life, similarities in codon usage may arise due to parallelism, convergence, or reversal resulting in identical character states not solely attributed to common ancestry.

Two primary explanations have been proposed to elucidate the observed non-random variation in codon usage. They rotate around the interplay of natural selection (translational selection) and neutral processes (mutational bias and genetic drift) (GRANTHAM et al., 1981; WAN et al., 2004; VICARIO; MORIYAMA; POWELL, 2007; ROTA-STABELLI et al., 2013; SUN; TAMARIT; ANDERSSON, 2017; LABELLA et al., 2019; KOKATE; TECHTMANN; WERNER, 2021). The concept of translational selection postulates that the preferential usage of certain codons is driven by the functional advantages they confer. Codon optimization, where the choice of codons aligns with the abundance of corresponding transfer RNA molecules (tRNA) in the genome, is believed to enhance translation efficiency and accuracy. This optimization has been linked to increased translation speed, accurate tRNA pairing, improved transcript stability, and the suppression of premature cleavage and polyadenylation of transcripts (BULMER, 1991; AKASHI, 1994; PARMLEY; HURST, 2007; ZHOU; WEEMS; WILKE, 2009; PRESNYAK et al., 2015; ZHOU et al., 2018).

However, in the absence of strong selection or under conditions where genetic drift predominates over selective forces, patterns of codon usage bias can be shaped by neutral processes (LABELLA et al., 2019). In such scenario, processes such as mutational biases, including biased mutation rates towards certain base compositions and GC-biased gene conversion, can influence the genome-wide mutational patterns and subsequently impact codon usage (CHEN et al., 2004). Even in the presence of selective pressures on synonymous codon sites, background substitutions driven by mutational biases may contribute to codon

preference, particularly in organisms with distinct GC compositions (SUN; TAMARIT; ANDERSSON, 2017).

It is important to highlight that explanations for codon usage bias based on natural selection and neutral processes are not mutually exclusive. Early researchers in the field recognized that codon bias likely arises from a delicate balance between selective and neutral forces (SHARP et al., 1993). The impacts of favoring certain codons over others can be significant, as changes in fitness have been observed (BALLARD; BIENIEK; CARLINI, 2019). There is evidence associating different codon usage patterns with the lifestyle of organisms, suggesting their potential role in species evolution (ARELLA; DILUCCA; GIANSANTI, 2021). Furthermore, the potential of codon usage bias in the speciation process has been proposed (RETCHLESS; LAWRENCE, 2012). Although these pieces of evidence have been primarily gathered from unicellular organisms, investigating the importance of differential codon usage throughout species evolution is an intriguing area of study.

Understanding how codon usage bias influences adaptive processes and genetic divergence underlying speciation events can provide insights into the mechanisms driving the formation of new species. Researchers can gain a better understanding of the selective pressures that shape codon usage patterns. Additionally, studying the evolutionary dynamics of codon usage over time can reveal genetic changes contributing to species divergence and specialization. Further investigations into the relationship between codon usage, species evolution, and adaptation across diverse taxa will enhance our understanding of the complex interplay between genetic variation, natural selection, and the development of biodiversity.

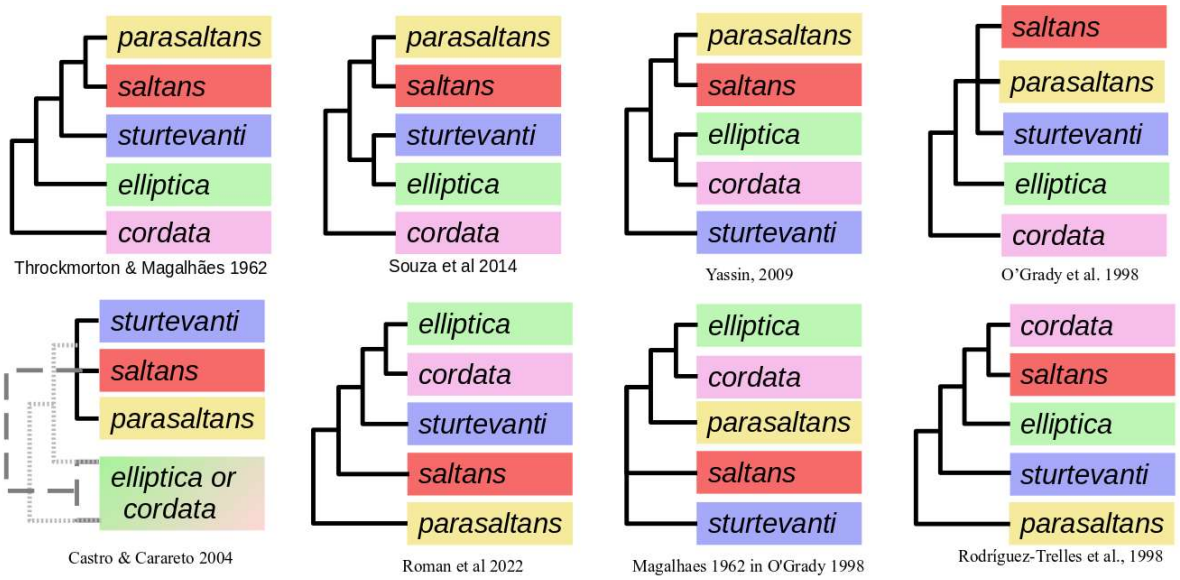
1.3 *Drosophila saltans* as model organism

The family Drosophilidae (Diptera) is composed of approximately 4600 species with wide morphological and ecological variation (TIDON; DE ALMEIDA, 2016; O'GRADY; DESALLE, 2018; BÄCHLI, 2023). These flies have many characteristics that make them excellent model organisms for various types of studies because many species have a short life cycle, well-defined developmental stages, and are easily collected and maintained in the laboratory at low cost. These characteristics are some of the reasons why *Drosophila melanogaster* has become one of the main model organisms for studies in the areas of development, genetics, and evolution (YAMAGUCHI; YOSHIDA, 2018). However, species of the *D. melanogaster* group originated in the Paleotropics (Ethiopian and Oriental biogeographic regions), with their genomes subjected to the selective pressures of these regions. Subsequently, some species of this group, such as *D. melanogaster*, *D. simulans*, *D.*

kikkawai, and more recently, *D. suzukii*, became cosmopolitan (TSACAS; DAVID, 1977; LACHAISE et al., 1988; LACHAISE; SILVAIN, 2004; MANSOURIAN et al., 2018; ØRSTED; ØRSTED, 2019; SPRENGELMEYER et al., 2020). Among the drosophilids with Neotropical evolution, we find the *D. saltans* group, which belongs to the same subgenus as *D. melanogaster* (BÄCHLI, 2023). The *saltans* group consists of 23 species divided into five subgroups: *saltans* (7 species), *cordata* (2 species), *elliptica* (4 species), *parasaltans* (2 species), and *sturtevanti* (7 species), in addition to the species *D. neoprosaltans*, which was described in 2017 and was compared to *D. prosaltans* a member of the *saltans* subgroup (MAGALHÃES; BJÖRNBERG, 1957; THROCKMORTON; MAGALHÃES, 1962; MAGALHÃES, 1962; THROCKMORTON, 1975; GUILLÍN; RAFAEL, 2017; MADI-RAVAZZI et al., 2021).

Several studies have sought to clarify the phylogenetic relationships within the *saltans* group of *Drosophila*, using different markers such as chromosome polymorphism (BICUDO, 1973b; BICUDO et al., 1978), reproductive isolation (BICUDO, 1973a, 1979; BICUDO; PRIOLI, 1978) pigmentation (THROCKMORTON; MAGALHÃES, 1962), 40 morphological characters of the body and genitalia (YASSIN, 2009), and characters of the male terminalia observed by scanning electron microscopy (SOUZA et al., 2014; ROMAN et al., 2022), as well as protein polymorphism (NASCIMENTO; BICUDO, 2002) and molecular markers like *Xdh* (RODRÍGUEZ-TRELLES; TARRÍO; AYALA, 1999a; TARRÍO; RODRÍGUEZ-TRELLES; AYALA, 2000) and the evolutionary approach reconstructed from careful analysis of transposable elements P, discarding horizontal transmission events (DE CASTRO; CARARETO, 2004), and a combination of morphological and molecular data such as the work of O'Grady et al. (1998), which combines 8 morphological characters with markers *ITS1*, *Adh*, *COI*, *COII*, and Roman et al. (2022), which combines 48 characters from literature review and electron microscopy data of male terminalia with markers *COI* and *COII* (ROMAN et al., 2022). The evolutionary relationships reconstructed by the aforementioned studies are summarized in Figure 3 and Appendix A's Supplementary Table S1.

Figure 3. Previous evolutionary relationships hypothesis of the 5 subgroups of the *D. saltans* group (see Supplementary Table 1 (Appendix A) for information about markers and method used by the authors). The subgroups — *saltans*, *parasaltans*, *sturtevanti*, *elliptica*, and *cordata* — are distinctly highlighted in red, yellow, blue, green, and pink, respectively.



Due to the inconsistencies found, there are still reservations regarding the evolutionary relationships in this group. However, discrepancies between studies that use different data sources (such as molecular and morphological data) are commonly reported, especially when few characters are analyzed. Different topologies found among molecular markers can occur due to intrinsic characteristics of species evolution, such as speciation in a short period of time, which can hinder the reconstruction of the species' evolutionary history (Darwinian shortfall, DINIZ-FILHO et al., 2013). The processes that generate these incongruences are predominantly ILS, introgression, or horizontal gene transfer (SIMION et al., 2017).

A promising method to clarify the phylogeny of this group of species is the use of next-generation sequencing methodologies, as they generate a large amount of phylogenetically informative characters. Several studies have used next-generation sequencing in taxonomic groups with complex phylogenetic reconstruction and have had promising results (MAI et al., 2020). These sequencing methodologies, in addition to helping clarify the phylogenetic relationships of the groups, also contribute to the importance of reticulation in the evolutionary history of this group (see Chapter 3).

Studies indicate that neotropical species of the subgenus *Sophophora* (the clade of *saltans* and *willistoni* groups) do not show a preference for codons ending in C and G, which is reported for other studied *Drosophila* species (VICARIO; MORIYAMA; POWELL, 2007;

KOKATE; TECHTMANN; WERNER, 2021). However, the analyses of the *saltans*-*willistoni* clade were either conducted with a few genes for the *saltans* group (RODRÍGUEZ-TRELLES; TARRÍO; AYALA, 1999b; TARRÍO; RODRÍGUEZ-TRELLES; AYALA, 2000; POWELL et al., 2003; YASSIN, 2009) or only with the genome of *D. willistoni* (VICARIO; MORIYAMA; POWELL, 2007; KOKATE; TECHTMANN; WERNER, 2021). The availability of genomes from species within the *saltans* group allows for a deeper investigation of codon usage modification within this clade (see Chapter 4).

2 OBJECTIVES

2.1 Main objective

To understand the evolutionary history of the *saltans* group of *Drosophila* by inferring the phylogenetic relationships of its species using genomic data and the evolution of its shift in codon usage shift.

2.2 Specific objectives

1. To infer the phylogenetic relationships and genome evolution of 15 species belonging to the five subgroups of the *saltans* group of *Drosophila* (Chapter 3).

2. To evaluate the effects of gene flow and/or ILS in the analyzed species (genomic porosity) (Chapter 3).

3. To assess the codon usage patterns in the *saltans* group species and investigate the relevance of mutational bias, selection, and drift in the observed patterns. (Chapter 4).

3 ORIGINAL ARTICLE I

This original research was presented in the 64th Annual *Drosophila* Research Conference, Chicago, United States of America. It was submitted to the Molecular Biology and Evolution journal.

3.1 Saltational episodes of reticulate evolution in the jumping fly *Drosophila saltans* species group

Carolina Prediger[1,2], Erina A. Ferreira[2], Samara Videira Zorzato[1,3], Aurélie Hua-Van[2], Lisa Klasson[4], Wolfgang J. Miller[5], Amir Yassin[2,3*] and Lilian Madi-Ravazzi[1*]

[1]Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo, Brazil.

[2]Laboratoire Évolution, Génomes, Comportement et Écologie, CNRS, IRD, Université Paris-Saclay, Gif-sur-Yvette, France.

[3]Institut de Systématique, Évolution, Biodiversité (ISYEB), CNRS, MNHN, EPHE, Sorbonne Université, Univ. des Antilles, Paris, France.

[4]Molecular evolution, Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Sweden.

[5]Center for Anatomy and Cell Biology, Department of Cell and Developmental Biology, Medical University of Vienna, Austria.

[*] These authors contributed equally.

Correspondence: amir.yassin@universite-paris-saclay.fr; lilian.madi@unesp.br.

Running title: Phylogenomics of the *Drosophila saltans* group

Abstract

Phylogenomics revealed reticulate evolution to be widespread across taxa, but whether reticulation is due to low statistical power (soft polytomy) or true evolutionary patterns (hard polytomy) remains a field of investigation. Here, we investigate the phylogeny and quantify reticulation in the *Drosophila saltans* species group, a Neotropical clade of the subgenus *Sophophora* comprising 23 species arranged in five subgroups, namely *cordata*, *elliptica*, *parasaltans*, *saltans* and *sturtevanti*, whose relationships have long been problematic. We sequenced and assembled the genomes of 15 species. Phylogenetic analyses revealed conflicting topologies between the X chromosome, autosomes and the mitochondria. We extended the ABBA-BABA test of asymmetry in phylogenetic discordance to cases where no "true" species tree could be inferred, and applied our new test (called 2A2B) to ≥50 kb-long 1,797 syntenic blocks with conserved collinearity across neotropical *Sophophora*. High incidences of reticulation (sometimes up to 90% of the blocks) were restricted to three nodes on the tree, at the split between the *cordata-elliptica-saltans* subgroups and at the origin of the *sturtevanti* and *saltans* subgroups. By contrast, cases with asymmetric discordances, which are often interpreted as evidence for interspecific introgression, did not exceed ~5% of the blocks. Historical biogeography analysis revealed that short inter-speciational times and greater overlap of ancestral geographical ranges partly explain cases with predominant reticulation. Therefore, episodic rapid radiations have played a major role in the evolution of this largely understudied Neotropical clade.

Keywords: phylogenomic discordance; genome assembly; historical biogeography; introgression; cyto-nuclear conflicts; Neotropical speciation; *Sophophora*.

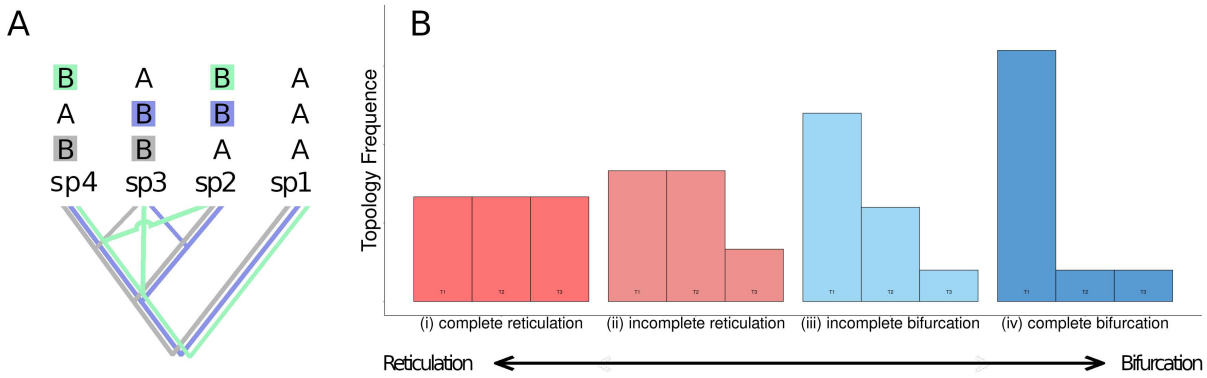
Introduction

Knowledge of phylogenetic relationships among species is a requirement for many evolutionary studies. However, it is often difficult to reconstruct well-resolved bifurcating trees for some clades. This could either be due to the lack of signal in the evaluated data, a condition known as “soft polytomy”, or due to persistent phylogenetic conflicts among datasets leading to “hard polytomies” and reticulate patterns of interspecific relationships. A plethora of biological processes could cause such conflicts, including incomplete lineage sorting (Maddison 1989; Maddison 1997; Walsh et al. 1999; Townsend et al. 2012), horizontal gene transfer, introgression and hybridization (Schrempf and Szöllősi 2020), and adaptive radiations (Glor 2010). Phylogenetic conflict also may be caused by technical errors, such as, sequencing error, contamination, wrong model selection and general lack of quality control (Philippe et al. 2011). Recent advances in genomic analyses have significantly reduced such errors and, in a wide range of taxa, increased the number of analyzed genes hence helping to resolve early conflicting topologies. However, in many other cases, whole genome analyses demonstrated persistent phylogenetic conflicts (e.g., in plants (Wickett et al. 2014; Gagnon et al. 2022), birds (Suh 2016), sponges and ctenophores (Philippe et al. 2009; Pick et al. 2010; Whelan et al. 2015; Chang et al. 2015; Simion et al. 2017), mammals (Romiguier et al. 2013; Morgan et al. 2013; Doronina et al. 2015), amphibians (Hime et al. 2021), and insects (Owen and Miller 2022)).

Of the different processes that can lead to reticulate evolution, introgression and hybridization have attracted much attention, first because they challenged long-held concept of reproductive isolation between species, and second due to the development of a number of bioinformatic tools and tests that quantify phylogenetic discordance across the genome (Durand et al. 2011; Pease and Hahn 2015; Malinsky et al. 2021). Site-based methods usually count the number of bi-allelic sites supporting each of three possible topologies in a species triplet with an outgroup (Figure 1A). Comparisons between the proportions of the three topologies can yield one of four possible outcomes (Figure 1B): (i) complete reticulation, all topologies are equally encountered; (ii) incomplete reticulation, such as in the case of full hybridization wherein two topologies significantly exceed the third one but do not significantly differ from each other; (iii) incomplete bifurcation, such as in the case of asymmetric introgression wherein the proportion of all topologies significantly differ; and (iv) complete bifurcation, one topology significantly exceeds the two others, which in their turn have nearly equal proportions. The earliest of introgression tests, Patterson's *D*, compared the two later cases (iii and iv), *i.e.* it presumed that a “true” species tree exists. A later test, HyDe,

quantifies admixture (γ) from the ratio of shared alleles with the test going from 0 (full isolation) to 0.5 (full hybridization) and therefore it can also cover case ii. The two tests differ in how they measure significance, using bootstrapping in Patterson's *D* and normal approximation in HyDe. Of late, another site-based test was developed using χ^2 to test for deviation of parity between the three topologies as in case i (Sayyari and Mirarab 2018). A unified test that can test the prevalence of each of the four categories across the genome and a phylogenetic tree is still lacking.

Figure 1. Distribution of bi-allelic patterns along the reticulation-bifurcation continuum and the 2A2B test. A) The distribution of bi-allelic sites of four species can generate three distinct topologies, BBAA with sp.3 and sp.4 as sister (gray topology), ABBA with sp.2 and sp.3 (blue topology) and BABA with sp.2 and sp.4 are most closely related (green topology). B) Based of the frequency of these topologies in a genome fragment, this fragment can be categorized in (i) complete reticulation, T1=T2=T3 (, (ii) incomplete reticulation, T1=T2>T3, (iii) incomplete bifurcation, T1>T2>T3, and (iv) complete bifurcation, T1>t2=T3, these classes are shown in red, pink, light blue and blue, respectively.



Polytomies and incongruencies have been reported for the jumping pomace fly *Drosophila saltans* species group, a clade of the subgenus *Sophophora* with 23 Neotropical species (Magalhães 1962). The group retains its name from the peculiar "jumping" habit of its larvae; *"the larva seizes its posterior end with its mouthhooks, and stretches. The hooks pull loose suddenly, the larva straightens with considerable force, and as a result is thrown several inches into the air"* (Sturtevant 1942). The group was divided into five species subgroups, namely, *saltans, parasaltans, cordata, elliptica* and *sturtevanti* subgroups, mostly on the basis of male genitalia (Magalhães and Björnberg 1957). Although the monophyly of the subgroups has been confirmed by different phylogenetic methods, the relationships among

and within them are not. Hypothesis for their evolutionary relationships have been proposed using different methods and different morphological characters (Magalhães and Björnberg 1957; Throckmorton 1962; Throckmorton and Magalhães 1962; O'Grady et al. 1998; Yassin 2009; Souza et al. 2014; Roman et al. 2022), chromosome polymorphism (Bicudo 1973a), reproductive isolation (Bicudo 1973b; Bicudo and Prioli 1978; Bicudo 1979), protein polymorphism (Nascimento and Bicudo 2002) and gene sequences (Pélandakis and Solignac 1993; O'Grady et al. 1998; Rodríguez-Trelles et al. 1999; de Castro and Carareto 2004; de Setta et al. 2007; Roman et al. 2022). The evolutionary relationships proposed are summarized in Supplementary Table S1.

Unlike other species groups in the subgenus *Sophophora*, such as the *melanogaster*, *obscura* and *willistoni* groups, genomic resources and genetic investigations in the *saltans* species group are scarce. Indeed, only four genomes have been sequenced and assembled to date (Kim et al. 2021). To bridge this gap and to test for the extent of phylogenetic conflicts, we sequenced and assembled genomes for 15 species with representatives from the five subgroups. Phylogenetic analyses using well-conserved genes resolved the evolutionary relationships among the subgroups but also highlighted conflicts between X-linked, autosomal and mitochondrial loci. To test how each of the four incongruence categories prevails across the genome, we devised a new χ^2-based test that uses pairwise comparisons of the three topologies proportions in long syntenic blocks with conserved collinearity across the neotropical *Sophophora* (Figure 1). We found reticulation levels to differ among the subgroups, in concordance with rate of speciation and historical biogeography.

Results

Short-read assembly of 17 genomes recovered 90% of BUSCO genes

We sequenced using short-read Illumina approach 17 whole genomes from 15 species collected across various locations in the Neotropical region. Genome size, estimated from 21-kmer frequency spectrum using GenomeScope 2 (Ranallo-Benavidez et al. 2020), ranged from 154.0 to 356.8 Mb. Our de novo assemblies using MaSuRCA (Zimin et al. 2013) resulted in genome lengths ranging from 177.5 to 287.7 Mb, with N50 values ranging from 2 to 92 Kb (Supplementary Table S2). To evaluate the completeness of our assembled genomes, we searched for single-copy genes (SCG) using Busco (Simão et al. 2015). We found that over 90% of the searched genes were complete for all of the genomes (Supplementary Table S2). Kim et al. (2021) assembled using both short Illumina and long Nanopore reads the

genomes of four *saltans* group species, all of which we have independently sequenced. Whereas their assemblies' contigs were much longer, with N50 ranging from 2 to 6 Mb, the BUSCO score for the same set of species did not largely differ (98% vs. 95-96% in our study; Supplementary Table S2). Their genomes were also included in subsequent phylogenetic analyses, using the assembly of *D. willistoni* as an outgroup (Kim et al. 2021).

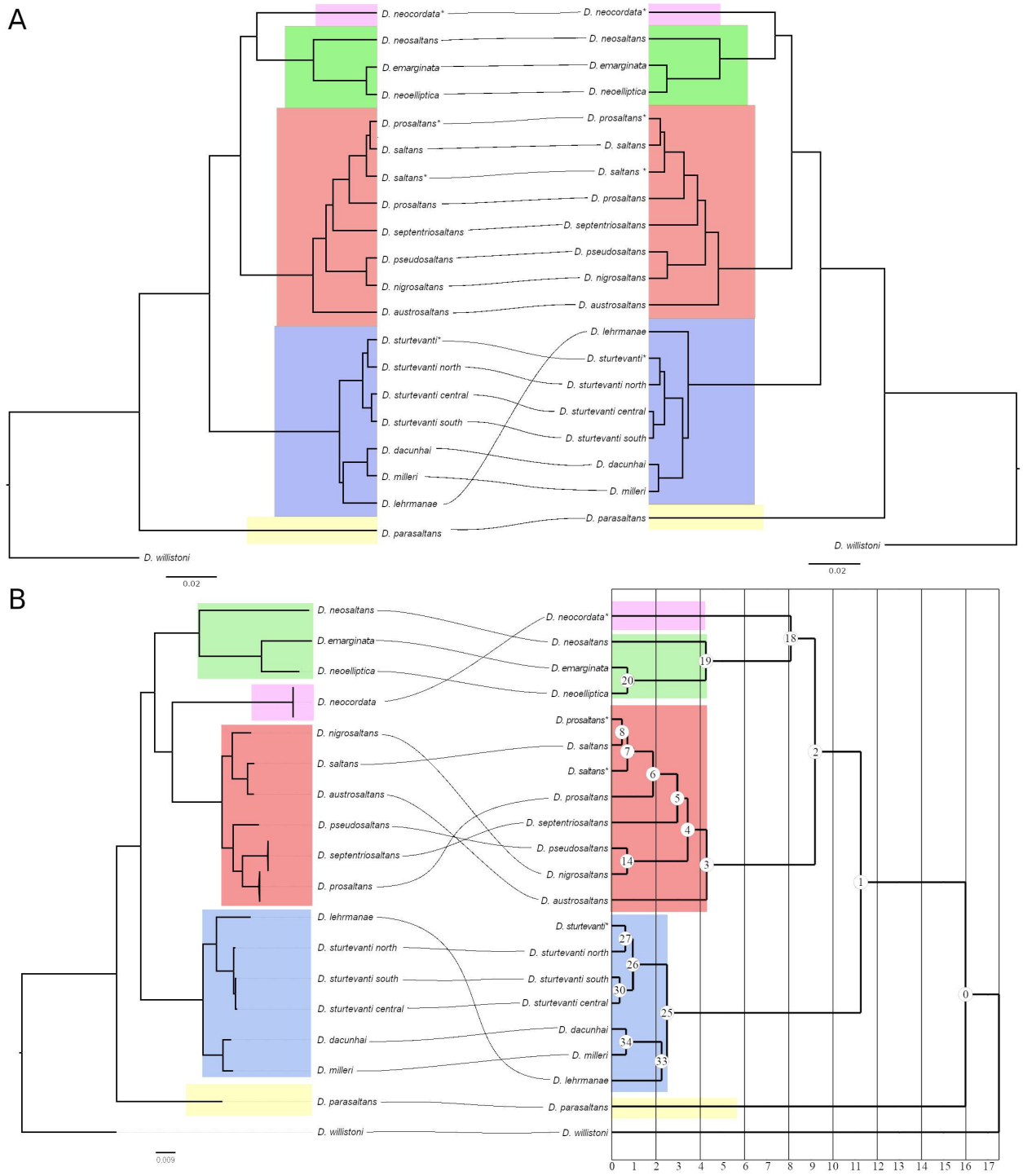
Muller elements analysis resolves relationships between the subgroups and unravels a minor X-autosomal conflict in the sturtevanti subgroup

Phylogenomic analyses were performed using 2,159 SCG shared across all species. Gene trees, inferred for each SCG using maximum-likelihood in IqTree produced 1,263 distinct topologies, with 206 of them found more than once (Supplementary Data S1). To test if SCG chromosomal position may underlie the discrepancies in gene trees, we localized each SCG to its corresponding Muller element according to the position of its *D. melanogaster* ortholog identified by Blast (Camacho et al. 2009). As a result, we generated five independent datasets, each corresponding to the Muller elements A, B, C, D, and E, comprising 337, 370, 425, 419, and 568 SCG, respectively. These datasets were then used to reconstruct the species trees using the multi-species coalescent model, and the genes within them were concatenated for Bayesian and Maximum Likelihood phylogenetic inferences.

The trees generated by the 5 data sets showed very similar topologies with well supported nodes either for the multi-species coalescent model analysis implemented in ASTRAL-III (Zhang et al. 2018), the maximum-likelihood implemented in IqTree (Nguyen et al. 2015) or Bayesian Inference implemented in BEAST (Bouckaert et al. 2019) (Supplementary Figures S1, S2 and S3). The *parasaltans* subgroup was placed as sister to all other subgroups, followed by the emergence of the *sturtevanti* subgroup. The *cordata* and *elliptica* subgroups showed a close relationship, and were sister to the *saltans* subgroup. The only discrepancy between the topologies was the placement of *D. lehrmanae*, a newly discovered species in the *sturtevanti* subgroup (Madi-Ravazzi et al. 2021). For *D. lehrmanae*, while maximum-likelihood and multi species coalescent analyses reported lack of branch support for multiple trees (Supplementary Figures S2 and S3), Bayesian inference recover well supported branches and two topologies (Figure 2A and Supplementary Figures S1). These two distinct topologies were identified among the Muller Elements forming the X chromosome (Muller elements A and D, a fusion shared by the neotropical *Sophophora*, the *saltans* and *willistoni* groups, see Sturtevant and Novitski 1941; Dobzhansky and Pavan 1943;

Cavalcanti 1948) and the Muller Elements representing autosomal chromosomes (Muller elements B, C and E).

Figure 2. Phylogenomic Conflict of X Chromosome, Autosomal, and Mitochondria. A) Comparative Analysis of autosomal topology (left, represented by Muller element B) and X-linked topology (right, represented by Muller element A) demonstrates overall agreement with minor Incongruence. B) Mitochondrial-Nuclear Disagreement highlight stronger incongruence between Mitochondrial Topology (left) and Sexual chromosome topology (right). Divergence time estimation (in million years ago, myr) for the Sexual Chromosome Topology is Provided. All posterior probabilities were equal to 1.



The published genome of *D. prosaltans* (Kim et al. 2021) did not group with the genome of this species sequenced by us, instead it grouped with *D. saltans*. The genome previously published comes from a line collected in El Salvador in 1957. According to Magalhães' (1962) detailed morphological revision of multiple geographical specimens of the *saltans* group, the sampling site of this particular strain is outside the geographical range of *D. prosaltans*, but within the expected range of *D. saltans*. Furthermore, the *D. saltans* and *D. prosaltans* lines used in our study underwent thorough morphological analyzes (Souza et al. 2014; Roman and Madi-Ravazzi 2021), indicating that the lines we used were accurately identified. Therefore, it is most likely that the previously sequenced *D. prosaltans* strain from El Salvador was misidentified and we consider it here to belong to *D. saltans*.

Mitogenomes show cytonuclear conflicts in the sturtevanti *and* saltans *subgroups*

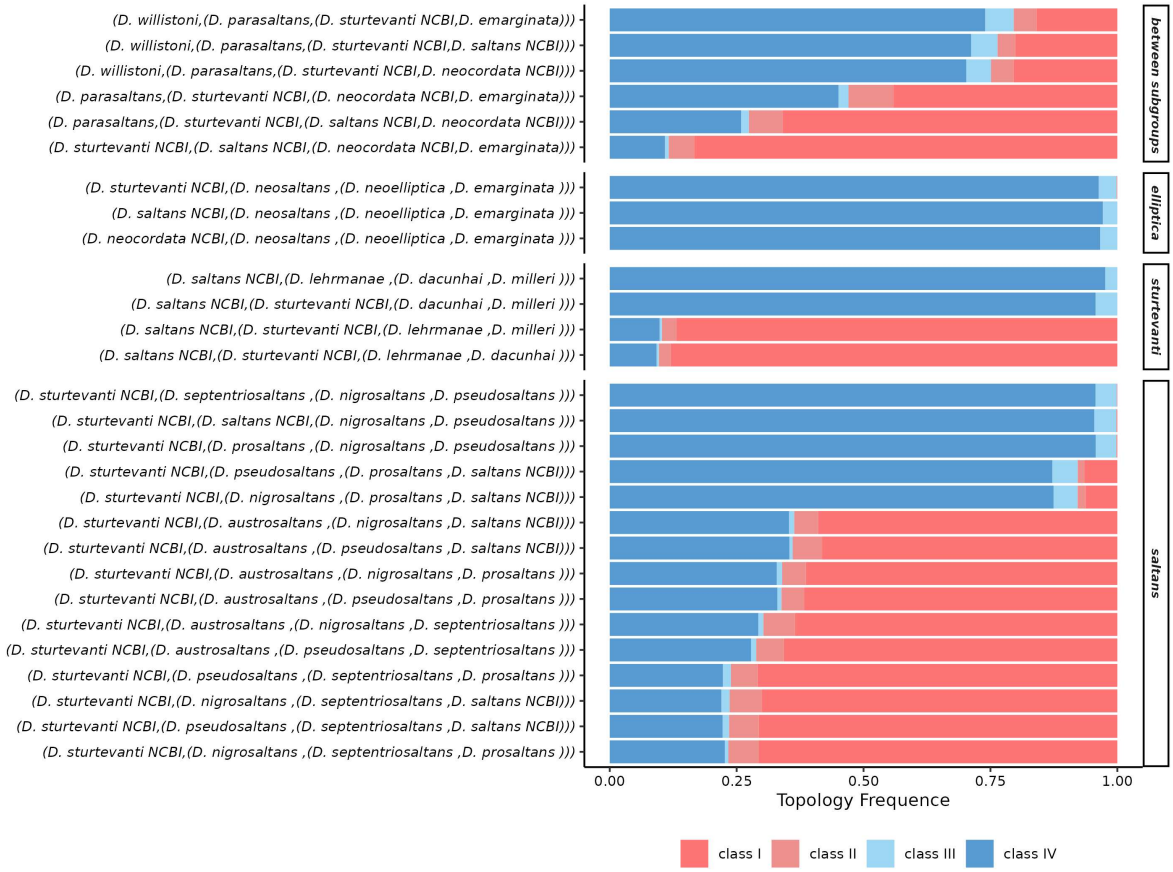
We assembled mitochondrial genomes for the 15 *saltans* group species using MitoZ (Meng et al. 2019). We did not use the previously assembled four strains since several mitochondrial scaffolds were likely removed in those assemblies (Kim et al. 2021). We conducted phylogenetic analysis on the aligned mitogenomes genes using both IqTree and MrBayes. Overall, the mitochondrial trees matched the topology of the nuclear gene trees regarding the inter-subgroup relationships. However, three major discrepancies were identified (Figure 2B). First, the position of *D. lehrmanae* within the *sturtevanti* subgroup did not agree with either the X or autosomal SCG topologies, proposing topology wherein *D. lehrmanae* is a sister species of *D. sturtevanti* (topology recover once in Multi-Species Coalesce analysis (Muller element C, Supplementary Figure S3) and Maximum likelihood (Muller element B, Supplementary Figure S2)). Second, whereas the mitochondrial tree recovered the monophyletic relationship between the *elliptica*, *cordata* and *saltans* subgroups, the position of *D. neocordata* (*cordata* subgroup) differed, being sister to the three species of the *elliptica* subgroup in the nuclear trees and to the six species of the *saltans* subgroup in the mitochondrial tree. Third, whereas nuclear trees recovered three lineages within the *saltans* subgroup, namely*, austrosaltans*, *nigrosaltans-pseudosaltans*, and *septentriosaltans-prosaltans-saltans*, only two lineages are revealed by the mitochondrial tree. Intriguingly, each of the mitochondrial clades involved one species from otherwise sister species in the nuclear trees, *i.e. D. nigrosaltans* and *D. saltans* in one clade and their respective closely-related species *D. pseudosaltans* and *D. prosaltans* in the other clade. Because *D. saltans* and *D. prosaltans* are reported as close related species in the nuclear trees and are separated in the two mitochondrial clades, the two mitotypes were called S and P, respectively. The

distribution of closely-related species between distinct mitotypic groups suggest that multiple cytoplasmic introgression events might have occurred in this subgroup (Figure 2B).

Site-specific phylogenetic analysis of syntenic blocks quantifies the extent of reticulate evolution in the saltans group

Site-specific analyses of phylogenetic discordance are highly sensitive to locus size (Martin et al. 2015; Pease and Hahn 2015). To overcome this problem, we identified 1,797 syntenic blocks ≥50 kb-long with conserved collinearity across the 15 *saltans* assemblies and *D. willistoni* (see Methods). For a four-taxon species tree with an outgroup, three topologies can possibly be obtained for each site with two alleles (A and B), namely AABB, ABBA and BABA, with the AABB topology usually refers to the true species tree (Durand et al. 2011; Patterson et al. 2012). However, to consider cases where a true species tree cannot be inferred, we designed a test for reticulation, that we call 2A2B. The test consists of comparing each pair of the three topologies using a χ^2 test, and classify each block with ≥20 evaluated sites into one of the four categories along the reticulation-bifurcation continuum given in Figure 1B. We run this test for every possible quartet (Supplementary Table S3). Whereas blocks supporting bifurcating trees (categories iii and iv) predominated in most quartets, we identified three parts on the species tree with reticulation indices (*i.e.* the proportion of blocks in categories i and ii) exceeding 70% (Figure 3).

Figure 3. The 2A2B test reveals a diminished introgression signal, while a prominent signal of reticulation evolution is evident within specific subgroups. The distribution of classes i-iv frequencies, spanning from symmetrical complete reticulation to asymmetrical bifurcation reticulation, is displayed for quartet species. A pronounced pattern of complete reticulation is apparent in the *saltans* and *sturtevanti* subgroups, whereas such a signal is absent in the *elliptica* subgroup.



At the inter-subgroup level, high incidences of reticulation were observed in any combination that involved representatives from at least two subgroups of the *cordata*, *elliptica* and *saltans* subgroups. For the *sturtevanti* subgroup, ~90% of the blocks could not resolve the relationships between *D. sturtevanti*, *D. lehrmanae* and the *dacunhai-milleri* clade, in agreement with the conflicting topologies between the X, autosomes and mitochondrial loci shown above. For the *saltans* subgroup, reticulation dominated (60-75%) in all comparisons involving *D. austrosaltans*, and representatives of the *nigrosaltans-pseudosaltans* and the *septentriosaltans-prosaltans-saltans* clades. However, not every subgroup with multiple representatives showed excess reticulation, since for the *elliptica* subgroup, almost no evidence for reticulate evolution was found whether *D. sturtevanti*, *D. neocordata* or any species of the *saltans* subgroup were used as an outgroup. Remarkably, the proportion of

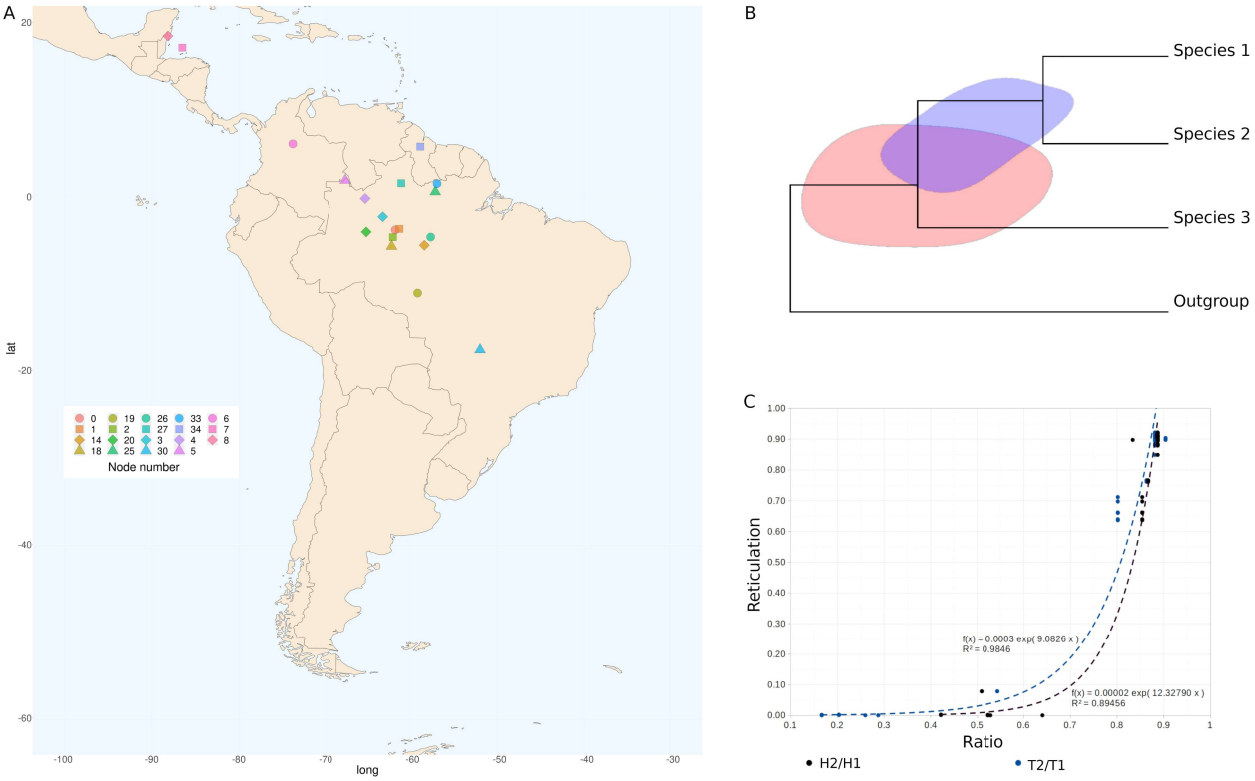
categories supporting inter-specific hybridization (ii) or introgression (iii) rarely exceeded 5% of the 50-kb long syntenic blocks (Figure 3).

Historical biogeography partly explains excess reticulation

To test if historical biogeography could explain the present incongruences, we mapped current distribution of the studied species on the Bayesian X tree. For each species, a polygon connecting the four most extreme cartesian points was drawn and the ancestral position of each point was inferred using BayesTraits (Meade and Pagel 2022) (see Methods). This approach allowed us to infer an ancestral range at each internal node of the tree. The historical biogeography supported an Amazonian origin of the *saltans* group around 16 million years (myr) ago (Figure 2B, 4A, Supplementary Table S4). Internal nodes as old as or older than 4 myr ago had ranges confined to the central or northern parts of South America. These nodes included the ancestors of all species subgroups except *sturtevanti*. Northwestern dispersal into Panama and southern Central America occurred around 3 myr ago, which correlates with the geological formation of the isthmus of Panama (O'Dea et al. 2016), and involved the ancestors of the *nigrosaltans-pseudosaltans* and *septentriosaltans-prosaltans-saltans* clades. The *sturtevanti* subgroup diversified around 2.5 myr ago in the northern parts of South America.

We tested the effects of the successiveness of speciation times on the estimated proportion of syntenic blocks with reticulated evolution patterns (*i.e.* categories i and ii). For each quartet with an (outgroup,(sp.1,(sp.2,sp.3))) topology we tested the regression of the proportion of reticulation on the ratio of the divergence time between sp.2 and sp.3 and the divergence time of the three ingroup species (hereafter T2/T1 ratio). This ratio increased as the time between successive speciation events shortened. Reticulation positively correlated with this measurement, and the regression line followed an exponential pattern (R^2 = 0.98) (Figure 4C, Supplementary Table S5). We also tested the regression of reticulation on the degree of overlap between the ancestral ranges of sp.2 and sp.3 (i.e. node 1 and 2 in Figure 4B), and of all ingroup species (hereafter H2/H1 ratio). This ratio indicates the degree of conservation of ancestral habitat and possible connectivity. A strong exponential correlation was obtained for this ratio (R^2 = 0.89) (Figure 4C).

Figure 4. Historical biogeography of the *saltans* group. A) the midpoint of the extreme geographical points for each ancestral node, reveals that the ancestral origins of all subgroups lie within the Amazonian forest, node numbers follows figure 2B. B) Illustration of the method employed to calculate the overlap of ancestral ranges of the ingroup species (H2/H1 ratio). Specifically, the geographical ranges of the ancestors, nodes 1 and 2, were inferred using BayesTraits, enabling the determination of shared and unique proportions of geographical ranges. C) Trashed blue line shows exponential relationship of reticulation in function of divergence time ratio of the three ingroup species (T2/T1 ratio) and reticulation (frequency of class i and ii). The black line depicts the exponential correlation between the overlap of ancestral ranges of the ingroup species (H2/H1 ratio) and reticulation.



Discussion

Towards a comprehensive phylogeny of the saltans species group

A large number of *Drosophila* genomes have been sequenced and used in phylogenetic analyses (Suvorov et al. 2022; Khallaf et al. 2021; Kim et al. 2021; Li et al. 2022), but studies with comprehensive sampling of nearly all species in a group remain relatively uncommon (Mai et al. 2020; Conner et al. 2021; Yusuf et al. 2022; Moreyra et al. 2023). Despite minor inconsistencies, our phylogenomic analysis of 15 species of the *Drosophila saltans* species group produced a consistent picture of the relationships between the five subgroups of this clade. All X, autosomal and mitochondrial phylogenies showed the

parasaltans subgroup as the first to diverge, followed by the *sturtevanti* subgroup, and later by a clade comprising the *cordata*, *elliptica* and *saltans* subgroups, in which the position of the *cordata* subgroup differed between nuclear and mitochondrial trees. This general picture has not been previously proposed despite the tremendous number of phylogenetic investigations of this group (Magalhães 1962; Throckmorton 1962; Throckmorton and Magalhães 1962; O'Grady et al. 1998; Rodríguez-Trelles et al. 1999; de Castro and Carareto 2004; de Setta et al. 2007; Yassin 2009; Souza et al. 2014; Roman et al. 2022), see Supplementary Table S1 for previous suggested topologies).

After establishing a coherent phylogenetic picture for the *Drosophila saltans* species group and identifying the relationships among its subgroups, the next critical step lies in expanding our sampling efforts. While our analysis has shed light on the intricate evolutionary dynamics within this clade, further sampling holds the potential to provide a more comprehensive understanding into this complex evolutionary history. For example, the inclusion of *D. subsaltans*, *D. lusaltans*, *D. cordata*, and *D. rectangularis* through whole-genome sequencing promises to provide insight into unresolved phylogenetic questions raised from previously published observations on reproductive isolation and morphology (Magalhães 1962; Bicudo and Prioli 1978). These questions include the monophyly and positioning of the *parasaltans* and *cordata* subgroups. Additionally, the inclusion of the insular species *D. lusaltans* which presents low reproductive isolation (Bicudo 1973b), can bring new insights into the reticulation evolution. These prospects for heightened sampling efficacy and its potential to unlock further dimensions of the *saltans* subgroup's evolution are explored in greater detail in Supplementary Document S1.

The *saltans* subgroup showed the most dramatic signal of cyto-nuclear discordance and reticulated evolution. Bicudo (1973a) investigated reproductive isolation among the seven then described species of this subgroup, and in a remarkably partial agreement with our nuclear phylogenomic trees, she concluded that *D. pseudosaltans*, *D. nigrosaltans* and *D. austrosaltans* showed more basal relationships compared to *D. lusaltans*, *D. septentriosaltans*, *D. prosaltans* and *D. saltans*. Indeed, nearly all crosses among the last four species produce fertile females with some even producing fertile females and males (Bicudo 1973b). This behavioral porosity largely agrees with the high incidence of reticulate evolution we report here for this subgroup.

Two widespread species of the *saltans* subgroup, *D. saltans* and *D. prosaltans*, show a peculiar geographical disjunction. The discrimination between strains belonging to each species has long been erroneous (Dobzhansky 1944; Mayr and Dobzhansky 1945; Spassky

1957; Magalhães 1962) and we showed here that their misidentification persists even in the genomic era (Suvorov et al. 2022; Kim et al. 2021). Interestingly, Bicudo (1973a) provided evidence for reproductive reinforcement between these two sister species; sympatric populations in their junction zone in Costa Rica demonstrated stronger reproductive isolation than allopatric populations of both species. We have only included one to a few geographical lines from each species and a broader sampling to investigate the extent of their reproductive isolation and genome porosity is strongly needed.

Intra- and inter-genomic conflicts impact the inference of phylogenetic patterns in the saltans *group*

Concatenation helped recovering a sexual versus autosome conflict, similar to the one described by Mai et al. (2020) for the *nasuta* subgroup. Like these authors, this conflict was limited to a single part of the tree, *i.e.* the relationship of *D. pulau* to *D. sulfurigaster sulfuricaster* and *D. s. bilimbata* in the *nasuta* group and the placement of *D. lehrmanae* in the *sturtevanti* subgroup. The peculiarities of sexual chromosomes such as the lower effective number, different recombination and mutation rates, the greater exposition to natural selection when found in hemizygosity, leads to higher rates of adaptive evolution of sexual-linked genes compared with autosomal genes (*i.e.* faster-X evolution) and also to the disproportional accumulation of genes related to reproductive isolation and Dobzhanski-Muller hybrid incompatibilities (*i.e.* Haldane's rule). Altogether, those characteristics are thought to be responsible for the resistance to hybridization in the sexual chromosomes (Ellegren 2009; Qvarnström and Bailey 2009; Sankararaman et al. 2016; Charlesworth et al. 2018; Seixas et al. 2018; Mai et al. 2020; Matute et al. 2020; Moran et al. 2021; Reilly et al. 2022; Skov et al. 2023; but see David et al. 2022).

The second conflict regards a significant disagreement between mitochondrial (mtDNA) and nuclear data. Discordance between nuclear and mitochondrial genomes is a well documented phenomenon in the tree of life as highlighted by (Toews and Brelsford 2012). Several characteristics of mtDNA, such as being haploid and uniparentally inherited, resulting in a fourfold reduction in effective population size when compared with autosomal chromosome loci, affect its evolution. Cytoplasmic introgression has long been recognized in *Drosophila* (Solignac et al. 1986; Ballard 2000; Llopart et al. 2014). In a recent population study within the *willistoni* group, multiple mitochondrial introgressions were observed in *D. paulistorum* populations. These included an ancient introgression with a highly divergent mitochondrial type, followed by more recent events. While nuclear-mitochondrial

incompatibilities likely posed challenges, the study also suggested two possible alternatives to overcome these challenges: a selective advantage provided by the mitochondrial type it self. Or a non-selective factor, such as *Wolbachia, a* bacteria known to modify the reproduction of its host, could facilitate a mitochondrial type fixation (Baião et al. 2023). Although, interesting results have been report from population approaches, conflicts between nuclear and mitochondrial genomes have not been addressed in recent phylogenomic analyses in the Drosophilidae (Mai et al. 2020; Khallaf et al. 2021; Suvorov et al. 2022; Yusuf et al. 2022). The disagreement was particularly evident for the *saltans* subgroup, where it was most likely of recent origins, separating species that have diverged only 0.7 myr ago, *i.e. D. nigrosaltans* and *D. pseudosaltans*. Remarkably, the two mitotypes P and S do not correlate with the degree of reproductive isolation inferred by Bicudo (1973b), contrary to nuclear tree, indicating that cytoplasmic introgression in the *saltans* subgroup did not contribute to the evolution of reproductive isolation in this clade.

Syntenic blocks also allowed a quantification of the degree of reticulate evolution. Of the three subgroups for which multiple species were sequenced, the *saltans* subgroup had the highest incidence of reticulation. For all subgroups, the degree of reticulation correlated negatively with the time between successive speciation events and positively with the degree of range conservatism. Indeed, reticulation is expected to increase with fast speciation increasing incomplete lineage sorting and/or range overlap promoting either gene flow or the selective retention of habitat-associated alleles (Avise and Robinson 2008; Degnan and Rosenberg 2009; Feng et al. 2022). In the *saltans* subgroups, multiple large chromosomal inversions are known to be shared among closely-related species (Dobzhansky and Pavan 1943; Cavalcanti 1948; Bicudo 1973a; Bicudo et al. 1978) and evidence for balancing on ancestral inversion has been demonstrated in a number of cases (Bicudo 1973a). Whether the high degree of reticulation in the *saltans* subgroup are associated with large ancestral inversions potentially absent in other bifurcating clades would require the future generation of chromosome-level assemblies for multiple *saltans* group species.

Large syntenic blocks distinguish soft from hard polytomies in the saltans *group*

There is no consensus in current phylogenomic analysis between concatenating and partitioning approaches. Whereas the former approach increases the power, *i.e.* providing a total evidence, it also introduces bias due to the non-independence of linked loci and in some cases it cannot be computationally feasible to analyze whole genomes. Alternatively, multi-locus-coalescent (MLC) approaches that partition the data into presumably independent and

neutral loci have been proposed. Those last approaches have broadly been applied in the investigation of phylogenetic discordances, mostly in studies inferring asymmetric introgressions. The definition of independent loci widely differs between studies with an impact on discordance estimate. For example, in a study of 155 genomes covering a wide range of drosophilid lineages, Suvorov et al. (2022) limited their MLC analyses on highly conserved single-copy protein-coding genes. However, their discordance estimates were highly sensitive to the length of the analyzed single genes as well as by the slightest relaxation of selective pressures, e.g., the exclusion of 5% of loci with the highest non-synonymous to synonymous ratio (dN/dS) led to a decrease of nearly 50% of discordance cases. An alternative approach is to align reads from multiple species to a well annotated genome, hence creating pseudo-references genome wherein different nucleotides replace their orthologous sites for each species. This approach was used by Mai et al. (2020) in the study of the *D. nasuta* subgroup, a clade of 12 species that diverged ~3 myr ago (Suvorov et al. 2022). These authors defined loci in terms of 500-kb long windows for phylogenetic reconstruction and 50-kb long windows for discordance analyses. The 500-kb windows were either analyzed separately or concatenated according to chromosomal arm (Muller's element). Whereas such an approach would increase the signal, it also introduces biases due to paralogy, misalignments or absence of collinearity among species. Besides, this approach is highly sensitive to the choice of the reference genome (Valiente-Mullor et al. 2021; Rick et al. 2023).

We combined here both approaches. First, we based our phylogenetic analysis on conserved single-copy protein-coding genes like Suvorov et al. (2022), but like Mai et al. (2020) we concatenated those genes according to their Muller elements. Second, we inferred phylogenetic discordance using large ≥50 kb-long windows like Mai et al. (2020), but unlike these authors we did not infer pseudo-references and defined our windows on large syntenic blocks that conserved their collinearity across neotropical *Sophophora*. Both approaches helped us to define signals of reticulate evolution that were not homogeneously distributed across the subgroups.(Mai et al. 2020; Khallaf et al. 2021; Suvorov et al. 2022; Yusuf et al. 2022).

Perhaps the most striking outcome of our synteny-based analysis is the low incidence of interspecific introgression compared to recent analyses across the genus *Drosophila* suggesting introgression to be widespread (Suvorov et al. 2022). Whether this discrepancy is due to the size of the analyzed loci or reflect genuine differences between the *saltans* group and other *Drosophila* clades would require the extension of the 2A2B test to these clades. Early phylogenetic studies in *Drosophila* suggested radiation episodes to be the most common

evolutionary patterns in drosophilids (Throckmorton 1975). If hard polytomes are widespread, currently common introgression analyses based on the assumption of true bifurcating species trees may be misled. Given the ever growing evidence for introgression in other animal and plant clades, we strongly recommend the application of phylogenetic discordance tests in large syntenic blocks in these organisms as well to distinguish introgression from rapid radiation events.

Materials and Methods

Sample collection, whole genome sequencing and assembly

We performed whole genome pool sequencing on female flies from 15 different species from the *saltans* group, as well as three populations of *D. sturtevanti*. The specimens used for sequencing were obtained from one or multiple strains, and detailed information regarding the number of individuals and their collection locations can be found in Supplementary Table S6. For all species except *D. neocordata*, which had its DNA extraction from ovaries and genome assembly described in BAIÃO et al. 2023, DNA was extracted following the manufacturer's instructions using the Promega DNeasy Kit. We conducted whole genome sequencing using the Illumina Hi-seq platform. The resulting genomes were then assembled using the Maryland Super Read Cabog Assembler (MaSuRCA) (Zimin et al. 2013), which utilizes both the Bruijn graph and overlap-layout-consensus (OLC) methods to generate super-reads. To assess the assembly's completeness, we searched for SCG using default parameters in Busco5 (Waterhouse et al. 2018) with the diptera_odb10 database (Kuznetsov et al. 2023).

Phylogenomics: Nuclear genes

In addition to the sequenced flies, we also utilized the reference genomes of *D. saltans*, *D. neocordata*, *D. prosaltans* and *D. sturtevanti* published by Kim et al. (2021) (assembly numbers ASM1890357v1, ASM1890361v1, ASM1815127v1 and ASM1815037v1, respectively) in our downstream analysis. For phylogenomics analysis, SCG searches were carried out using 3,285 SCG from diptera_odb10 database on Busco5 (Waterhouse et al. 2018). SCG present in all species were kept and aligned using the L-INS-i method implemented on MAFFT (Katoh and Standley 2013) (mafft --localpair --maxiterate 1000 –adjustdirection).

Genomic data of different species of *Drosophila* support the ancient proposition that genes tend to be situated within the same Muller element across multiple species, suggesting

that natural selection has maintained a low rate of gene transposition between chromosomal arms (see SCHAEFFER, 2018). Taking this gene linkage into account, we reconstructed five independent datasets (Muller elements A-E), each comprising all SCG found in the respective Muller element. To achieve this, we performed a tBlastn search against the *D. melanogaster*, and subsequently we concatenated genes found within the same Muller element. Phylogenetic trees were then constructed using maximum likelihood and Bayesian methods implemented in the softwares IQ-TREE (Nguyen et al. 2015) and BEAST (Bouckaert et al. 2019), respectively. Additionaly, maximum-likelihood trees were generated for each gene, the output tree from each Muller element data-set were used to reconstruct to species trees, using multi-species coalesce model implemented in ASTRAL-III (Zhang et al. 2018).

Phylogenomics: Mitochondrial Genome

Mitochondrial genomes were assembled and annotated with MitoZ (Meng et al. 2019), with the Megahit assembler (Li et al. 2015). In order to ensure the exclusion of nuclear-embedded mitochondrial DNA sequences within the assembly, a strategic approach was taken. Considering that mitochondrial reads are found in higher frequency than nuclear-mitochondrial DNA sequences, the read subsampling were set to 0.5 gigabases (--data_size_for_mt_assembly 0.5). The genes obtained from the mitochondrial genome were aligned using the MAFFT alignment tool with the --auto parameter due to the close similarity between sequences. Subsequently, the aligned genes were concatenated into a dataset for phylogenetic analysis. The concatenated dataset served as the basis for reconstructing phylogenetic trees using both Maximum Likelihood (ML) implemented in IQ-TREE (Nguyen et al. 2015) and Bayesian Inference (BI) in BEAST (Bouckaert et al. 2019).

Quantifying reticulation: 2A2B test

The 20 genomes of the *saltans* group, the 16 sequenced here and 4 published by Kim et al. (2021) were preliminary annotated with Miniprot (miniprot -Iut16, (Li 2023)). The primary objective of this annotation was to accurately map proteins from the robust and reliable genome annotation of *D. willistoni*. After the protein mapping, the predicted gene loci were assessed to identify syntenic blocks present in the neotropical *Sophophora* (comprising *D. willistoni* and *saltans* group). The identification of these blocks was based on gene order and orientation, achieved using an in-house Perl script. First, this script compares the scaffolds' genes order and orientation between the references genome *D. saltans* and *D. sturtevanti*, the synteny block were defined when all the genes were fond in same order and

orientation for both species. The identified collinear blocks were than searched for the remaining genomes. Blocks with missing data, i.e. missing gene for one or more species were subsequently removed, and the remaining blocks were subjected to a size-based filtering with threshold of 50kb. The selected blocks were subjected to alignment using the Mafft (Katoh and Standley 2013). The resulting alignments were integral to the subsequent analysis, which aims to measure reticulation evolution, in the new 2A2B test.

All combination of quartets species were evaluated for test reticulation, bi-allelic non degenerated sited shared by pairs were searched in every synteny blocks. Collinear blocks that presented at least 20 informative sites between the evacuated quartets were kept. Bi-alellic sites shared between pairs of species quartets can generate 3 topologies, AABB (species 2 and 3 closely related), ABBA (species 1 and 2 closely related), and BABA (species 1 and 3 closely related) as shown in Figure 1A. For each synteny block, the occurrences of these three topologies were counted, and three different χ^2-based tests were conducted. First, the Patterson's D, this measure quantifies the difference in allele sharing between species pairs. It provides insights into whether a ABBA or BABA topology is more prevalent. It is calculated as the difference in allele sharing normalized by the total allele sharing as in Equation 1.

$$D1 = \frac{\left(\sum ABBA - \sum BABA\right)^2}{\sum ABBA + \sum BABA} \quad (1),$$

The two other test were D2 and D3 (Equation 2 and 3) focus on discordant allele-sharing patterns (AABB vs. ABBA and AABB vs. BABA, respectively). They help identify cases where allele sharing between species pairs deviates from what's expected under a simple divergence model. Significant values for D1 or D2 might indicate that certain alleles are more shared between species pairs than expected.

$$D2 \frac{\left(\sum AABB - \sum BABA\right)^2}{\sum AABB + \sum BABA} (2), \text{ and}$$

$$D3 = \frac{\left(\sum AABB - \sum ABBA\right)^2}{\sum AABB + \sum ABBA} (3)$$

Afterward, based on how these three topologies were distributed within each synteny block and considering the significance of the three tests, the collinear blocks were grouped into one of four categories (see figure 1B). Class I comprises complete reticulation, the frequencies of the three topologies do not deviate from neutral expectation, i.e. they are equal (D1, D2 and D3 are not significant), high frequencies of this class are caused by incomplete lineage sorting. Class II, comprises the cases in which two topologies do not significantly differentiate between them and are more frequent than the third topology (i.e. two D tests are significant), high frequencies of this class are expected in cases where hybridization had happen. Class iii, incomplete bifurcating, comprises the cases in with the frequency of all topologies are significantly different. Blocks classified in class iii show asymmetric introgression signals. Class iv comprise the cases in each one topology is much more frequent than the two alternatives ones, and the minor topologies are not significantly different from each other. High frequency of class iv is expected under the complete lineage sorting, and it is seen when one topology frequency greater outweighs two the alternative ones, which do not different between them. After each block classification, the overall genome porosity between the quartets were evaluated.

Historical biogeography

To determine the sampling locations of the evaluated species, we conducted searches in TaxoDros (https://www.taxodros.uzh.ch/search/class.php). Additionally, we incorporated sampling sites that we ourselves had conducted. It is important to note that the accuracy of our species identification was confirmed through BarCode verification. After inspection of the geographical points and manual correction, we identified the most northern, southern, western, and eastern points for each species. We used those points to reconstruct the ancestral geographical extremes in BayesTraits (Meade and Pagel 2022). This analysis was carried out using the GEO model with the phylogenetic tree generated reconstructed with the Muller element A (XL chromosome arm), 1.000.000 of MCMC and 25% burn-in. The divergence times were estimated using this tree under Bayesian inference. The calibration point used was the split between *D. willistoni* (17.5 myr), as estimated by Suvorov et al. (2022).

To assess the relationship between reticulation ratio and speciation ratio, we employed specific calculations. The reticulation ratio, indicating the frequency of syntenic blocks in class i and ii, was computed for groups of four species. Similarly, the speciation (T2/T1) ratio was also calculated using quartets, it is determined by the divergence time between the ancestor of species 2 and species 3 in relation to the divergence time of the species 1. We also

evaluated the relationship between reticulation and ancestral connectivity between the species. To do that, we utilized the ancestral geographical extremes and determined the predicted overlap area using the polygon R package. Finally, we computed a ratio according to Equation 4:

$$H2/H1 ratio = \frac{2 \times H}{E_{A1} + E_{A2}} \quad (4)$$

Here, "H" represented the shared geographical area, "E" is the exclusive geographical area of ancestral 1 (A1) and 2 (A2). The fit for linear and exponential regressions between Reticulation and T2/T1 ratio and between Reticulation and H2/H1 ratio were calculated.

Acknowledgments

We would like to thank David Ogereau for his assistance and insights in genome assembly for this study. We thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the financial support to L.M.R. (Number processes: 95/06165-1, 2014/14059-0 and 2016/ 11994-5) enabling us to collect and sequence many strains used here. We extend our thanks to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their support, specifically grant number 141545/2020-8, as well as the France Excellence Eiffel Scholarship Program for funding C.P.'s PhD scholarship. *Illumina* sequencing of *D. neocordata* was performed at the SNP&SEQ Technology Platform in Uppsala, Sweden, which is part of the Swedish National Genomics Infrastructure and Science for Life Laboratory. SNP&SEQ is supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. This project was partly supported by a grant from The Swedish research council VR (2014-4353) to L.K. We want to thank the Nouragues research field station (managed by CNRS), which benefits from "Investissement d'Avenir" grants managed by Agence Nationale de la Recherche (AnaEE France ANR-11-INBS-0001; Labex CEBA ANR-10-LABX-25-01). The project was partly funded by the Austrian Science Fund FWF grant P28255-B22 to W.J.M..

References

Avise JC, Robinson TJ. 2008. Hemiplasy: A New Term in the Lexicon of Phylogenetics. Kubatko L, editor. *Systematic Biology* 57:503–507.

Baião GC, Schneider DI, Miller WJ, Klasson L. 2023. Multiple introgressions shape mitochondrial evolutionary history in *Drosophila paulistorum* and the *Drosophila willistoni* group. *Molecular Phylogenetics and Evolution* 180.

Ballard JWO. 2000. Comparative genomics of mitochondrial DNA in members of the Drosophila melanogaster subgroup. *Journal of Molecular Evolution* 51:48–63.

Bicudo HEMC. 1973a. Chromosomal polymorphism in the *saltans* group of *Drosophila* I. The *saltans* subgroup. *Genetica* 44:520–552.

Bicudo HEMC. 1973b. Reproductive isolation in the *saltans* group of *Drosophila*. I. The *saltans* subgroup. *Genetica* 44:313–329.

Bicudo HEMC. 1979. Reproductive isolation of the *saltans* group of *Drosophila*. IV. the *sturtevanti* subgroup. *Revista Brasileira de Genética* II:247–258.

Bicudo HEMC, Hosaki MK, Machado J, Marques MCN. 1978. Chromosomal polymorphism in the *saltans* group of *Drosophila* II. Further study on *D. prosaltans*. *Genetica* 48:5–15.

Bicudo HEMC, Prioli AJ. 1978. Reproductive isolation in the *saltans* group of *Drosophila* II. The *parasaltans* subgroup. *Genetica* 48:17–22.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 15:1–28.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10:1–9.

de Castro JP, Carareto CMA. 2004. P elements in the *saltans* group of *Drosophila*: A new evaluation of their distribution and number of genomic insertion sites. *Molecular Phylogenetics and Evolution* 32:383–387.

Cavalcanti AGL. 1948. Geographic variation of chromosome structure in *Drosophila prosaltans*. *Genetics* 33:529–536.

Chang ES, Neuhof M, Rubinstein ND, Diamant A, Philippe H, Huchon D, Cartwright P. 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences* 112:14912–14917.

Charlesworth B, Campos JL, Jackson BC. 2018. Faster-X evolution: Theory and evidence from *Drosophila*. *Molecular Ecology* 27:3753–3771.

Conner WR, Delaney EK, Bronski MJ, Ginsberg PS, Wheeler TB, Richardson KM, Peckenpaugh B, Kim KJ, Watada M, Hoffmann AA, et al. 2021. A phylogeny for the *Drosophila montium* species group: A model clade for comparative analyses. *Molecular Phylogenetics and Evolution* 158:107061.

David JR, Ferreira EA, Jabaud L, Ogereau D, Bastide H, Yassin A. 2022. Evolution of

assortative mating following selective introgression of pigmentation genes between two *Drosophila* species. *Ecology and Evolution* 12:e8821.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* 24:332–340.

Dobzhansky T. 1944. Experiments on Sexual Isolation in *Drosophila*: III. Geographic Strains of *Drosophila Sturtevanti*. *Proceedings of the National Academy of Sciences* 30:335–339.

Dobzhansky TG, Pavan C. 1943. Studies on Brazilian species of *Drosophila*. *Boletim da Faculdade de Filosofia, Ciências e Letras da Universidade de São Paulo. Biologia Geral.* 36:7–72.

Doronina L, Churakov G, Shi J, Brosius J, Baertsch R, Clawson H, Schmitz J. 2015. Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria, Carnivora). *Molecular Biology and Evolution* 32:3194–3204.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol* 28:2239–2252.

Ellegren H. 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics* 25:278–284.

Feng S, Bai M, Rivas-González I, Li C, Liu S, Tong Y, Yang Haidong, Chen G, Xie D, Sears KE, et al. 2022. Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell* 185:1646-1660.e18.

Gagnon E, Hilgenhof R, Orejuela A, McDonnell A, Sablok G, Aubriot X, Giacomin L, Gouvêa Y, Bragionis T, Stehmann JR, et al. 2022. Phylogenomic discordance suggests polytomies along the backbone of the large genus *Solanum*. *Am J Bot*.

Glor RE. 2010. Phylogenetic Insights on Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics* 41:251–270.

Guillín ER, Rafael V. 2017. Cinco especies nuevas del género *Drosophila* (Diptera, Drosophilidae) en la provincia de Napo, Ecuador. *Iheringia - Serie Zoologia* 107:1–12.

Hime PM, Lemmon AR, Lemmon ECM, Prendini E, Brown JM, Thomson RC, Kratovil JD, Noonan BP, Pyron RA, Peloso PLV, et al. 2021. Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life. *Systematic Biology* 70:49–66.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780.

Khallaf MA, Cui R, Weißflog J, Erdogmus M, Svatoš A, Dweck HKM, Valenzano DR,

Hansson BS, Knaden M. 2021. Large-scale characterization of sex pheromone communication systems in *Drosophila*. *Nat Commun* 12:4165.

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'agostino ERR, Pelaez J, et al. 2021. Highly contiguous assemblies of 101 drosophilid genomes. *eLife* 10:1–32.

Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* 51:D445–D451.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674–1676.

Li F, Rane RV, Luria V, Xiong Z, Chen J, Li Z, Catullo RA, Griffin PC, Schiffer M, Pearce S, et al. 2022. Phylogenomic analyses of the genus *Drosophila* reveals genomic signals of climate adaptation. *Molecular Ecology Resources* 22:1559–1581.

Li H. 2023. Protein-to-genome alignment with miniprot.Valencia A, editor. *Bioinformatics* 39:btad014.

Llopart A, Herrig D, Brud E, Stecklein Z. 2014. Sequential adaptive introgression of the mitochondrial genome in Drosophila yakuba and Drosophila santomea. *Molecular Ecology* 23:1124–1136.

Maddison W. 1989. Reconstructing character evolution on polytomous cladograms. *Cladistics* 5:365–377.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.

Madi-Ravazzi L, Roman BE, Cesar K, Alevi C, Prediger C, Yassin A, Wolfgang J. 2021. Integrative taxonomy and a new species description in the *sturtevanti* subgroup of the *Drosophila saltans* group (Diptera: Drosophilidae). 4980:269–292.

Magalhães LE de. 1962. Notes on the taxonomy, morphology, and distribution of the *saltans* group of *Drosophila*, with description of four new species. *The University of Texas Publication*:135–154.

Magalhães LE de, Björnberg AJS. 1957. Estudo da genitália masculina de *Drosophila* do grupo *saltans* (Díptera). *Revista Brasileira de Biologia* 17:435–450.

Mai D, Nalley MJ, Bachtrog D, Wright S. 2020. Patterns of genomic differentiation in the *Drosophila nasuta* species complex. *Molecular Biology and Evolution* 37:208–220.

Malinsky M, Matschiner M, Svardal H. 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources* 21:584–595.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* 32:244–257.

Matute DR, Comeault AA, Earley E, Serrato-Capuchina A, Peede D, Monroy-Eklund A, Huang W, Jones CD, Mackay TFC, Coyne JA. 2020. Rapid and Predictable Evolution of Admixed Populations Between Two *Drosophila* Species Pairs. *Genetics* 214:211–230.

Mayr E, Dobzhansky Th. 1945. Experiments on Sexual Isolation in *Drosophila*. *Proceedings of the National Academy of Sciences* 31:75–82.

Meade A, Pagel M. 2022. Ancestral State Reconstruction Using BayesTraits. *Methods Mol Biol* 2569:255–266.

Meng G, Li Y, Yang C, Liu S. 2019. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research* 47:e63.

Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. 2021. The genomic consequences of hybridization.Wittkopp PJ, editor. *eLife* 10:e69016.

Moreyra NN, Almeida FC, Allan C, Frankel N, Matzkin LM, Hasson E. 2023. Phylogenomics provides insights into the evolution of cactophily and host plant shifts in *Drosophila*. *Molecular Phylogenetics and Evolution* 178:107653.

Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O'Connell MJ. 2013. Heterogeneous models place the root of the placental mammal phylogeny. *Mol Biol Evol* 30:2145–2156.

Mourão CA, Bicudo HEMDC. 1967. Duas novas especies do grupo *saltans* (drosophilidae: diptera). *Papeis Avulsos de Zoologia* 20:123–134.

Nascimento AP, Bicudo HEMC. 2002. Esterase patterns and phylogenetic relationships of *Drosophila* species in the *saltans* subgroup (*saltans* group). *Genetica* 114:41–51.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32:268–274.

O'Dea A, Lessios HA, Coates AG, Eytan RI, Restrepo-Moreno SA, Cione AL, Collins LS, de Queiroz A, Farris DW, Norris RD, et al. 2016. Formation of the Isthmus of Panama. *Science Advances* 2:e1600883.

O'Grady PM, Clark JB, Kidwell MG. 1998. Phylogeny of the *Drosophila saltans* species

group based on combined analysis of nuclear and mitochondrial DNA sequences. *Molecular Biology and Evolution* 15:656–664.

Owen CL, Miller GL. 2022. Phylogenomics of the Aphididae: Deep relationships between subfamilies clouded by gene tree discordance, introgression and the gene tree anomaly zone. *Systematic Entomology* 47:470–486.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. *Genetics* 192:1065–1093.

Pease JB, Hahn MW. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology* 64:651–662.

Pélandakis M, Solignac M. 1993. Molecular phylogeny of *Drosophila* based on ribosomal RNA sequences. *Journal of Molecular Evolution* 37:525–543.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology* 9.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Current Biology* 19:706–712.

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alié A, Morgenstern B, Manuel M, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Molecular Biology and Evolution* 27:1983–1987.

Qvarnström A, Bailey RI. 2009. Speciation through evolution of sex-linked genes. *Heredity* 102:4–15.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 11:1432.

Reilly PF, Tjahjadi A, Miller SL, Akey JM, Tucci S. 2022. The contribution of Neanderthal introgression to modern human traits. *Current Biology* 32:R970–R983.

Rick JA, Brock CD, Lewanski AL, Golcher-Benavides J, Wagner CE. 2023. Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. :2022.03.10.483737. Available from: https://www.biorxiv.org/content/10.1101/2022.03.10.483737v2

Rodríguez-Trelles F, Tarrío R, Ayala FJ. 1999. Molecular evolution and phylogeny of the *Drosophila saltans* species group Inferred from the *Xdh* Gene. *Molecular Phylogenetics and Evolution* 13:110–121.

Roman BE, Madi-Ravazzi L. 2021. Male terminalia morphology of sixteen species of the *Drosophila saltans* group Sturtevant (Diptera, Drosophilidae). *Zootaxa* 5061:523–544.

Roman BE, Santana DJ, Prediger C, Madi-Ravazzi L. 2022. Phylogeny of *Drosophila saltans* group (Diptera: Drosophilidae) based on morphological and molecular evidence. *Plos One* 17:e0266710.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular Biology and Evolution* 30:2134–2144.

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology* 26:1241–1247.

Sayyari E, Mirarab S. 2018. Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes* 9:132.

Schaeffer SW. 2018. Muller "Elements" in *Drosophila* : How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. *Genetics* 210:3–13.

Schrempf D, Szöllősi G. 2020. The sources of phylogenetic conflicts. In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. chapter. 3.1, p. 3.1:1-3.1:23. Available from: https://hal.inria.fr/PGE

Seixas FA, Boursot P, Melo-Ferreira J. 2018. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. *Genome Biology* 19:91.

de Setta N, Loreto ELS, Carareto CMA. 2007. Is the evolutionary history of the O-type P element in the *saltans* and *willistoni* groups of *Drosophila* similar to that of the canonical P element? *Journal of Molecular Evolution* 65:715–724.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol* 27:958–967.

Skov L, Coll Macià M, Lucotte EA, Cavassim MIA, Castellano D, Schierup MH, Munch K. 2023. Extraordinary selection on the human X chromosome associated with archaic admixture. *Cell Genomics* 3:100274.

Solignac M, Monnerot M, Mounolou JC. 1986. Mitochondrial DNA evolution in the

melanogaster species subgroup of Drosophila. *J Mol Evol* 23:31–40.

Souza TAJ, Noll FB, Bicudo HEMC, Madi-Ravazzi L. 2014. Scanning electron microscopy of male terminalia and its application to species recognition and phylogenetic reconstruction in the *Drosophila saltans* group. *PLoS ONE* 9.

Spassky B. 1957. Morphological differences between sibling species of Drosophila. In: Genetics of Drosophila. Vol. 5721.

Sturtevant AH. 1942. The classification of the genus *Drosophila*, with descriptions of nine new species. *The University of Texas Publication* 4213:7–51.

Sturtevant AH, Novitski E. 1941. The homologies of the chromosome elements in the genus *Drosophila*. *Genetics* 26:517–541.

Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zoologica Scripta* 45:50–62.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Wadell P, Lang M, Courtier-Orgogozo V, et al. 2022. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Current Biology* 32:111–123.

Tarrío R, Rodríguez-Trelles F, Ayala FJ. 1998. New *Drosophila* introns originate by duplication. *Proc Natl Acad Sci U S A* 95:1658–1662.

Throckmorton LH. 1962. The problem of phylogeny in the genus *Drosophila*. *University of Texas Publications* 6205:207–343.

Throckmorton LH. 1975. The phylogeny, ecolopy, and geography of Drosophila. *Handbook of Genetics, Vol 3.*:421–469.

Throckmorton LH, Magalhães LE. 1962. Changes with evolution of pteridine accumulations in species of the *saltans* group of the genus *Drosophila*. *University of Texas Publications* 6205:489–505.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology* 21:3907–3930.

Townsend JP, Su Z, Tekle YI. 2012. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. *Systematic Biology* 61:835–849.

Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N, Mejía L, Ruiz-Hueso P, González-Candelas F. 2021. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLOS Computational Biology* 17:e1008678.

Vilela CR, Bächli G. 1990. Taxonomic studies on Neotropical species of seven genera of Drosophilidae (Diptera). *Mitteilungen der Schweizerischen Entomologischen Gesellschaft* 63:1–332.

Walsh HE, Kidd MG, Moum T, Friesen VL. 1999. Polytomies and the power of phylogenetic inference. *Evolution* 53:932–937.

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* 35:543–548.

Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences* 112:5773–5778.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* 111:E4859–E4868.

Yassin A. 2009. Phylogenetic relationships among species subgroups in the *Drosophila saltans* group (Diptera: Drosophilidae): Can morphology solve a molecular conflict. *Zoological Research* 30:225–232.

Yusuf LH, Tyukmaeva V, Hoikkala A, Ritchie MG. 2022. Divergence and introgression among the *virilis* group of *Drosophila*. *Evolution Letters* 6:537–551.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29:2669–2677.

4 ORIGINAL ARTICLE II

This original research had part of its partial results presented in the 63rd Annual *Drosophila* Research Conference and at 66th Brazilian Congress of Genetics (virtual participation in both events). And it is planed to be submitted to the Journal of Molecular Evolution.

4.1 Ancestral state relaxation and contrasting trends in codon usage across 174 *Drosophila* species

Carolina Prediger[1,2]; Lilian Madi-Ravazzi[1,*]; Amir Yassin[2,3,*]

[1]Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo, Brazil.

[2]Laboratoire Évolution, Génomes, Comportement et Écologie, CNRS, IRD, Université Paris-Saclay, Gif-sur-Yvette, France.

[3]Institut de Systématique, Évolution, Biodiversité (ISYEB), CNRS, MNHN, EPHE, Sorbonne Université, Univ. des Antilles, Paris, France.
* These authors contributed equally.

Correspondence: amir.yassin@universite-paris-saclay.fr; lilian.madi@unesp.br.
Running title: Codon usage in *Drosophila*

Abstract

We investigate codon usage bias evolution in the Drosophilidae family, with a specific focus on the neotropical *Sophophora* clade composed of the *saltans* and *willistoni* groups. The lack of codon usage bias has been described in this clade with few genes or only one genome. We obtained genome assemblies for 197 drosophilids, including 27 genomes representing 23 species from the neotropical *Sophophora,* the *saltans* and *willistoni* groups. After filtering for genome completeness, we analyzed a final dataset consisting of 174 genomes and selected 3285 single copy genes for the evaluation of codon usage bias. Our results revealed varying degrees of codon usage bias among the species, with average effective number of codons (ENC) ranging from 40.61 to 53.92. The *saltans*-*willistoni* clade displayed a remarkable lack of codon usage preference, similar to outgroup taxa. We further examined the relative synonymous codon usage (RSCU) and identified specific codon preferences within different drosophilid species. Additionally, we detected significant differences in codon usage in *Zaprionus bogoriensis* compared to other species within the genus. Phylogenetic signal analysis indicated a strong association between codon usage patterns and evolutionary relationships, suggesting the influence of genetic drift. Furthermore, we found mutational bias to play a crucial role in determining the observed codon preferences. Interestingly, our results indicated a primary shift favoring codons ending in C and G in the early branches of Drosophilid evolution, followed by a reversal in the neotropical *Sophophora* clade. We also identified a distinct, yet opposite codon usage shift within the *Zaprionus* genus. We discuss the different causes that had potentially led to contrasting codon usage bias trends in the Drosophilidae.

Key words: Mutational bias, codon usage, genome evolution, *Drosophila saltans* group, *Drosophila willistoni* group, Neotropical *Drosophila*

Introduction

The degeneration of the genetic code and the implications of synonymous and non-synonymous mutations play crucial roles in living organisms and are fundamental concepts in the neutralist school (KIMURA, 1968). Although synonymous mutations do not directly alter the primary structure of proteins, it is evident that codons are not randomly utilized. This uneven usage of synonymous codons encoding the same amino acid, known as codon usage bias, has long been studied by researchers (FIERS et al., 1976; GRANTHAM et al., 1981; SHARP et al., 1997). The unequal usage of synonymous codons encoding the same amino acid has been widely observed across genes and organisms, raising questions about the forces driving this bias. Two primary explanations have been proposed to elucidate the observed variation in codon usage: natural selection (translational selection) and neutral processes (mutational bias and genetic drift). Natural selection suggests that certain codons are preferentially used due to the functional advantages they confer, such as enhanced translation efficiency, accurate tRNA pairing, improved transcript stability, and regulation of gene expression. On the other hand, neutral processes can shape codon usage in the absence of strong selection, with mutational biases and genetic drift influencing genome-wide mutational patterns (GRANTHAM et al., 1981; WAN et al., 2004; VICARIO; MORIYAMA; POWELL, 2007; ROTA-STABELLI et al., 2013; SUN; TAMARIT; ANDERSSON, 2017; BALLARD; BIENIEK; CARLINI, 2019; LABELLA et al., 2019; BALLARD; BIENIEK; CARLINI, 2019; KOKATE; TECHTMANN; WERNER, 2021). It is worth noting that the explanations based on natural selection and neutral processes are not mutually exclusive. Early researchers recognized that codon bias likely results from a delicate balance between selective and neutral forces (SHARP et al., 1993). Codon usage bias has been associated with fitness changes (BALLARD; BIENIEK; CARLINI, 2019) and linked to the lifestyle of organisms (ARELLA; DILUCCA; GIANSANTI, 2021), and even its potential role speciation event has been suggested (RETCHLESS; LAWRENCE, 2012).

Regarding the drosophilids, it has been shown that many species tend to favor synonymous mutation in the codons ending with G or C (VICARIO; MORIYAMA; POWELL, 2007; KOKATE; TECHTMANN; WERNER, 2021). However, this pattern is not seen within the neotropical species of the subgenus *Sophophora*, namely the *saltans* and *willistoni* groups. Previous analyses were conducted with a few genes (POWELL et al., 2003; RODRÍGUEZ-TRELLES; TARRÍO; AYALA, 1999b; TARRÍO; RODRÍGUEZ-TRELLES;

AYALA, 2000; YASSIN, 2009) or focused solely on the genome of *D. willistoni (KOKATE; TECHTMANN; WERNER, 2021; VICARIO; MORIYAMA; POWELL, 2007)*.

The study of codon usage bias provides valuable insights into the selective pressures that shape genetic variation and contribute to the development of biodiversity. Drosophilids, a well-studied clade of insects in the field of genetics, offer a unique opportunity to investigate this phenomenon. With over 150 genomes already sequenced and available, Drosophilids provide a rich dataset for understanding codon usage patterns. In this study, our main focus is on the neotropical *Sophophora* clade, as previous research has indicated a significant shift in codon usage within this group.

Results and discussion

Ortholog Identification

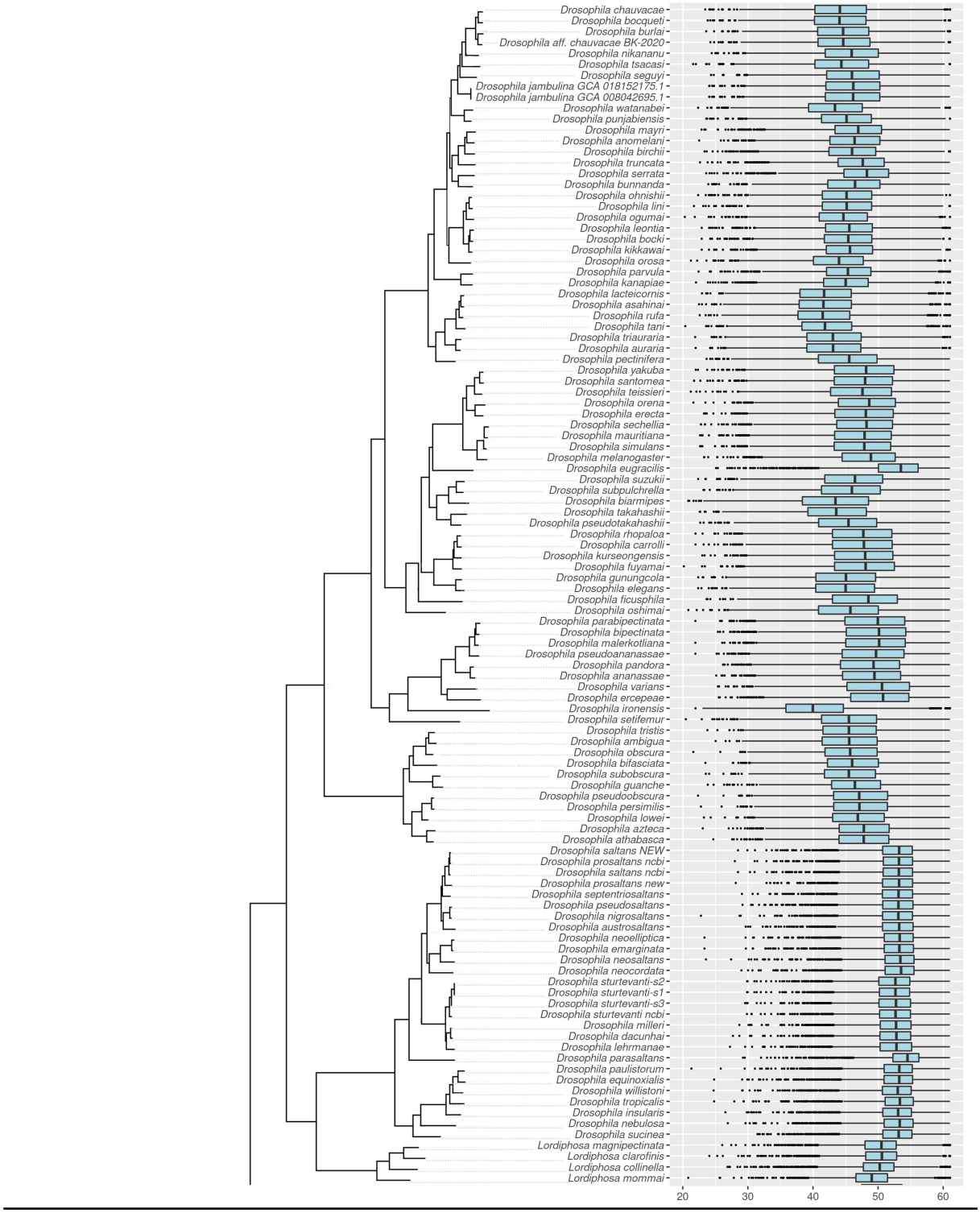
In our study, we focused on obtaining genome assemblies for Acalyptratae, with particular attention to the Drosophilidae family. A total of 197 genomes were acquired and assessed, with a primary emphasis on the neotropical *Sophophora* clade. These genomes underwent a thorough completeness filtering process, applying stringent BUSCO score criteria of 90% or higher. As a result, we refined the dataset to consist of 174 genomes. Within this dataset, 27 genomes from 23 species are member of *saltan-willistoni* clade, covering nearly 50% of the neotropical *Sophophora*. All SCG were subsequently employed to examine codon usage patterns within the Drosophilidae family. Within this dataset, we employed a total of 3,285 Single-Copy Genes (SCGs) as search queries. This search led to the recovery of 556,876 SCG, with each species yielding at least of 2,959 genes (Supplementary Table S1).This comprehensive dataset formed the basis for our subsequent analysis of codon usage patterns in Drosophilidae.

Overview of codon Usage Bias in Drosophilidae

Aiming to obtain a general picture of codon usage bias in Drosophidae, the effective number of codons (ENC) were calculated. ENC provide insights into the preference or lack thereof in gene codon usage. It varies from 20 to 61, and a value of 20 indicates complete bias, where each amino acid is carried by a unique tRNA molecule in a gene. Conversely, a value of 61 indicates complete unbiased selection of codons, where all synonymous codons are used equally. The ENC averages ranged from 40.61 to 53.92 (Supplementary Table S2), indicating varying degrees of codon usage bias among the species. Notably, strong codon usage biases average were observed in species such as *D. ironensis* (*melanogaster* group,

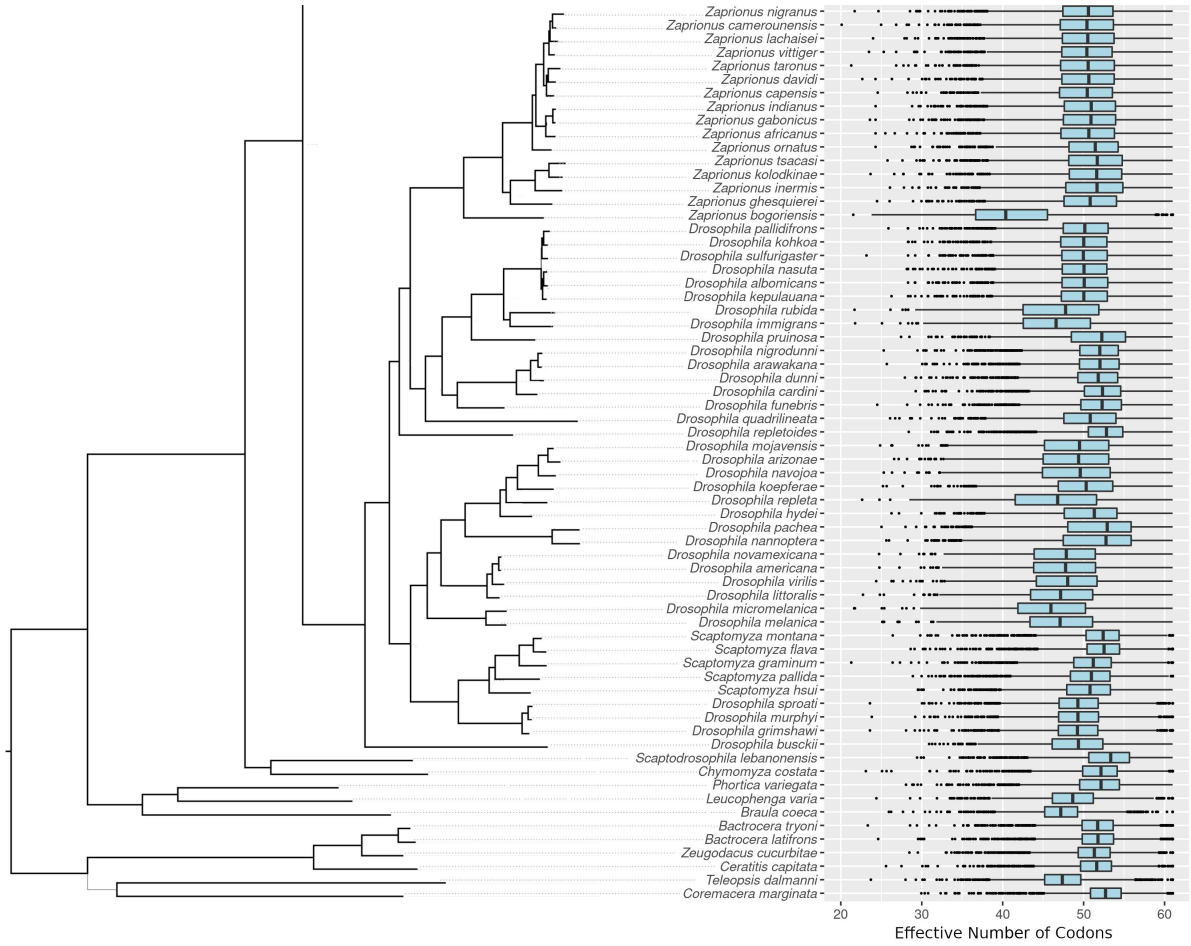
ananassae subgroup), *Z. bogoriensis*, and multiple species of the *montium* group (*D. rufa*, *D. asahinai*, *D. lacteicornis* and *D. tani*). Conversely, species within the *saltans-willistoni* clade exhibited a remarkable lack of codon usage preference. This observation confirms that the previously described pattern, based on one genome (VICARIO; MORIYAMA; POWELL, 2007; KOKATE; TECHTMANN; WERNER, 2021) or few genes (RODRÍGUEZ-TRELLES; TARRÍO; AYALA, 1999b, 1999a; YASSIN, 2009), is a conserved feature across the neotropical *Sophophora* clade. Some early diverged Drosophilidae, such as *Scaptodrosophila lebanonensis* (average ENC = 52.84) and *Phortica variegata* (51,76), displayed high ENC values, indicating low codon usage bias in the ancestral, however *Braula coeca*, early divergent line of Drosophilidae, presented moderated ENC averages (47.17) ((Supplementary Table S2, Figure1).

Figure 1. Investigating codon usage bias (CUB) across Drosophilids reveals potential shifts in in CUB evolution. The *saltans-willistoni* clade, in particular, stands out with ENC values much higher than those observed in the *Sophophora* and higher in the *Drosophila* subgenera, indicating a distinct lack of codon usage preference across all species within this clade. Notably, our analysis also uncovers a contrasting shift in ENC values during the evolutionary trajectory of the *Zaprionus* genus. In this context, *Z. bogoriensis* emerges as a striking outlier, displaying a pronounced trend towards codon bias that contradicts the broader patterns observed within its genus. Furthermore, it's worth noting that non-Drosophilids and basal clade of Drosophilidae tend to exhibit higher ENC values, suggesting a consistent trend towards a lack of codon usage bias in these lineages. The evolutionary relationship between this species is shown by a phylogenetic tree constructed using a maximum likelihood approach based on 192 single-copy genes.


Drosophila chauvacae
Drosophila bocqueti
Drosophila burlai
Drosophila aff. chauvacae BK-2020
Drosophila nikananu
Drosophila tsacasi
Drosophila seguyi
Drosophila jambulina GCA 018152175.1
Drosophila jambulina GCA 008042695.1
Drosophila watanabei
Drosophila punjabiensis
Drosophila mayri
Drosophila anomelani
Drosophila birchii
Drosophila truncata
Drosophila serrata
Drosophila bunnanda
Drosophila ohnishii
Drosophila lini
Drosophila ogumai
Drosophila leontia
Drosophila bocki
Drosophila kikkawai
Drosophila orosa
Drosophila parvula
Drosophila kanapiae
Drosophila lacteicornis
Drosophila asahinai
Drosophila rufa
Drosophila tani
Drosophila triauraria
Drosophila auraria
Drosophila pectinifera
Drosophila yakuba
Drosophila santomea
Drosophila teissieri
Drosophila orena
Drosophila erecta
Drosophila sechellia
Drosophila mauritiana
Drosophila simulans
Drosophila melanogaster
Drosophila eugracilis
Drosophila suzukii
Drosophila subpulchrella
Drosophila biarmipes
Drosophila takahashii
Drosophila pseudotakahashii
Drosophila rhopaloa
Drosophila carrolli
Drosophila kurseongensis
Drosophila fuyamai
Drosophila gunungcola
Drosophila elegans
Drosophila ficusphila
Drosophila oshimai
Drosophila parabipectinata
Drosophila bipectinata
Drosophila malerkotliana
Drosophila pseudoananassae
Drosophila pandora
Drosophila ananassae
Drosophila varians
Drosophila ercepeae
Drosophila ironensis
Drosophila setifemur
Drosophila tristis
Drosophila ambigua
Drosophila obscura
Drosophila bifasciata
Drosophila subobscura
Drosophila guanche
Drosophila pseudoobscura
Drosophila persimilis
Drosophila lowei
Drosophila azteca
Drosophila athabasca
Drosophila saltans NEW
Drosophila prosaltans ncbi
Drosophila saltans ncbi
Drosophila prosaltans new
Drosophila septentriosaltans
Drosophila pseudosaltans
Drosophila nigrosaltans
Drosophila austrosaltans
Drosophila neoelliptica
Drosophila emarginata
Drosophila neosaltans
Drosophila neocordata
Drosophila sturtevanti-s2
Drosophila sturtevanti-s1
Drosophila sturtevanti-s3
Drosophila sturtevanti ncbi
Drosophila milleri
Drosophila dacunhai
Drosophila lehrmanae
Drosophila parasaltans
Drosophila paulistorum
Drosophila equinoxialis
Drosophila willistoni
Drosophila tropicalis
Drosophila insularis
Drosophila nebulosa
Drosophila sucinea
Lordiphosa magnipectinata
Lordiphosa clarofinis
Lordiphosa collinella
Lordiphosa mommai
20
30
40
50
60

continue

continue



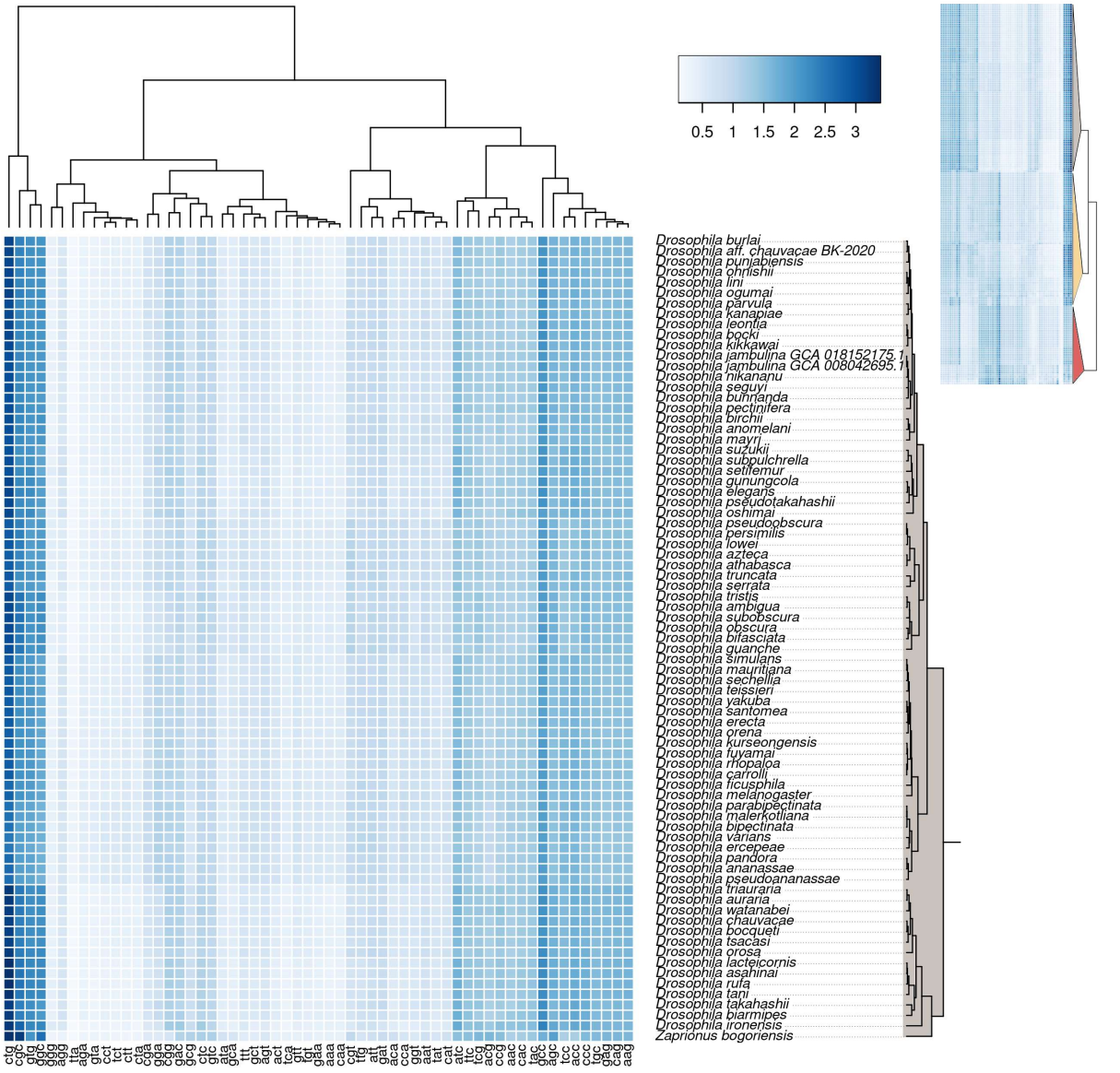
To visualize the differences in codon preferences among Drosophilids, we used relative synonymous codon usage (RSCU) and constructed a hierarchical clustering based on average RSCU values (refer to Supplementary Table S2 and Figure 2). The evaluation of the heatmap (Figure 2) This heatmap revealed three primary clusters: The first cluster, depicted in red in Figure 2, consists of the neotropical *Sophophora*, the outgroup species, and the basal drosophilids. Notably, the *willistoni-saltans* groups, exhibited codon usage patterns related to those of the outgroups. For instance, they moderately preferred AGT for serine instead of AGC, and they favored TTG for leucine over CTG, as well as CGT for arginine instead of CGC (Supplementary Table S2 and Figure 2). The second cluster comprises the *Dorsiphola*, *Siphlodora*, and *Drosophila* subgenera, along with the Hawaiian *Drosophila* and *Zaprionus* genus (shown in sandy beige, Figure 2). The third cluster includes *Sophophora* - old world and *Zaprionus bogoriensis*.

Within the *saltans* subgroup, an intriguing pattern emerged, with *D. parasaltans* closely resembling the codon usage of the *willistoni* group. This suggests variations in genetic

coding preferences within the *saltans* group may have evolved after the split *D. parasaltans* branched off. For example, *D. parasaltans* shares more similarities in codon usage with the *willistoni* group, particularly regarding glutamine codons. Additionally, an interesting observation was the differential usage of multiple codons in *Zaprionus bogoriensis*, which carries various amino acids using different tRNAs compared to other species within the genus. This divergence in codon usage is a novel finding. (Figure 2, Supplementary Table S2). In summary, this analysis separated Drosophilids into three major clusters, reflecting distinct codon usage patterns and highlighting intriguing differences within and between species or subgroups.

Figure 2. Relative synonymous codon usage (RSCU) analysis shows that the codon usage pattern the neotropical *Sophophora* clade is more similar with the pattern seen for the ancestral of drosophilids, and indicate a codon usage shift in the *Zaprionus* genus. The 61 columns represent he non-stop codons. Rows correspond to the 174 genomes that have been evaluated, darker colored cells correspond to the codons of tRNAs that are favored and lighter cells the unfavored tRNAs-codons. Codons and species were clustered using hierarchical clustering by RSCU values.

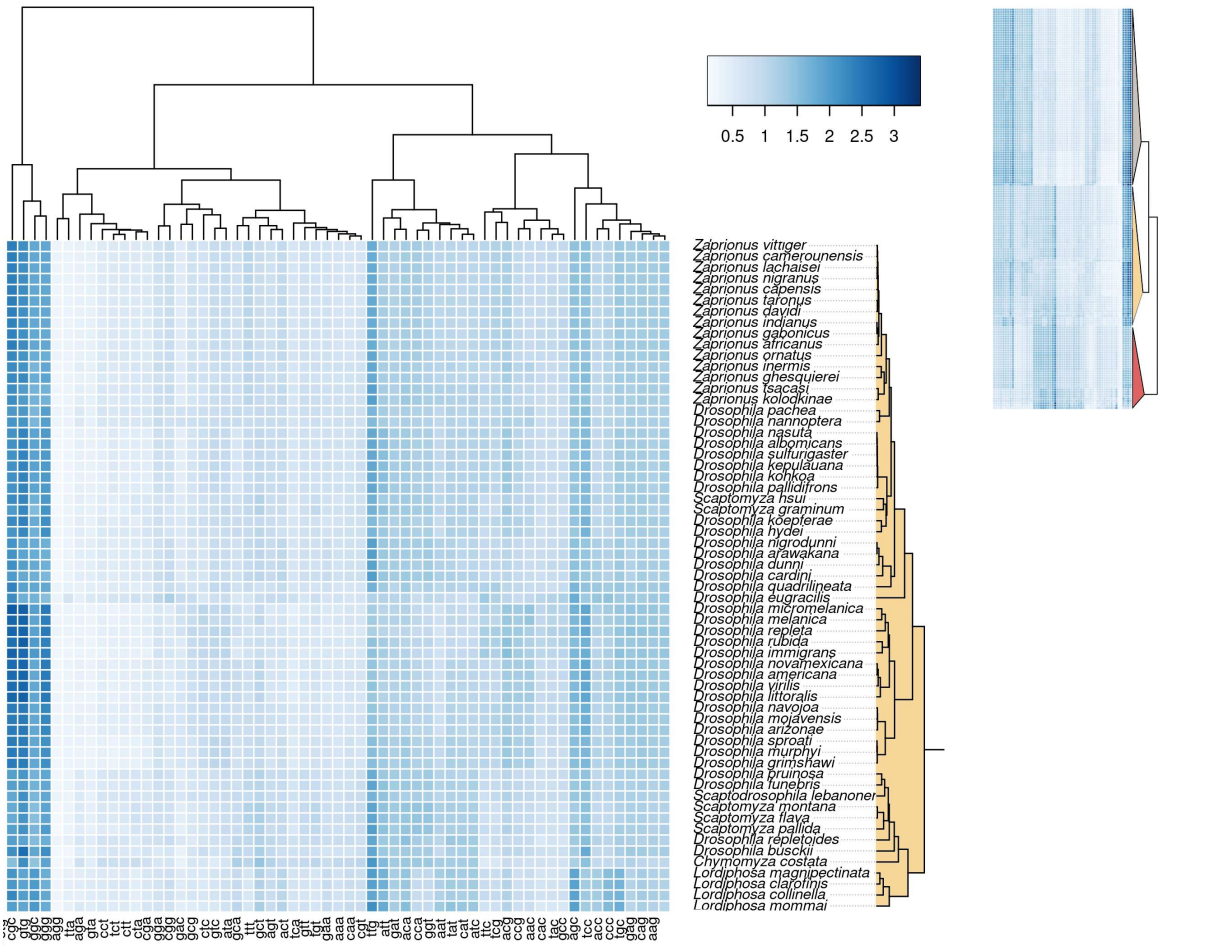
A



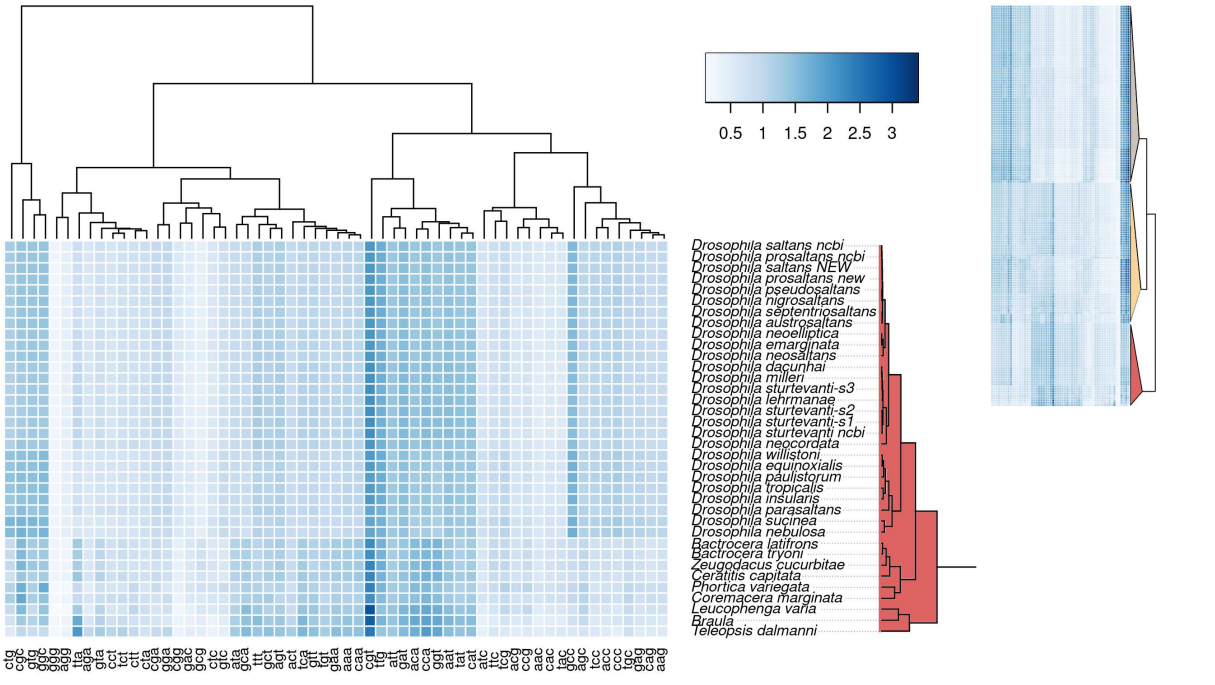
continue

continue

B


0.5 1 1.5 2 2.5 3
Zaprionus vittiger
Zaprionus carherounensis
Zaprionus lachaisei
Zaprionus nigranus
Zaprionus capensis
Zaprionus taronus
Zaprionus davidi
Zaprionus indianus
Zaprionus gabonicus
Zaprionus africanus
Zaprionus ornatus
Zaprionus inermis
Zaprionus ghesquierei
Zaprionus tsacasi
Zaprionus kolodkinae
Drosophila pachea
Drosophila hannoptera
Drosophila nasuta
Drosophila albomicans
Drosophila sulfurigaster
Drosophila kepulauana
Drosophila kohkoa
Drosophila pallidifrons
Scaptomyza hsui
Scaptomyza graminum
Drosophila koepferae
Drosophila hydei
Drosophila nigrodunni
Drosophila arawakana
Drosophila dunni
Drosophila cardini
Drosophila quadrilineata
Drosophila eugracilis
Drosophila micromelanica
Drosophila melanica
Drosophila repleta
Drosophila rubida
Drosophila immigrans
Drosophila novamexicana
Drosophila americana
Drosophila virilis
Drosophila littoralis
Drosophila navojoa
Drosophila mojavensis
Drosophila arizonae
Drosophila sproati
Drosophila murphyi
Drosophila grimshawi
Drosophila bromosa
Drosophila funebris
Scaptodrosophila lebanonen
Scaptomyza montana
Scaptomyza flava
Scaptomyza pallida
Drosophila repletoides
Drosophila busckii
Chymomyza costata
Lordiphosa magnipectinata
Lordiphosa clarofinis
Lordiphosa collinella
Lordiphosa mommai

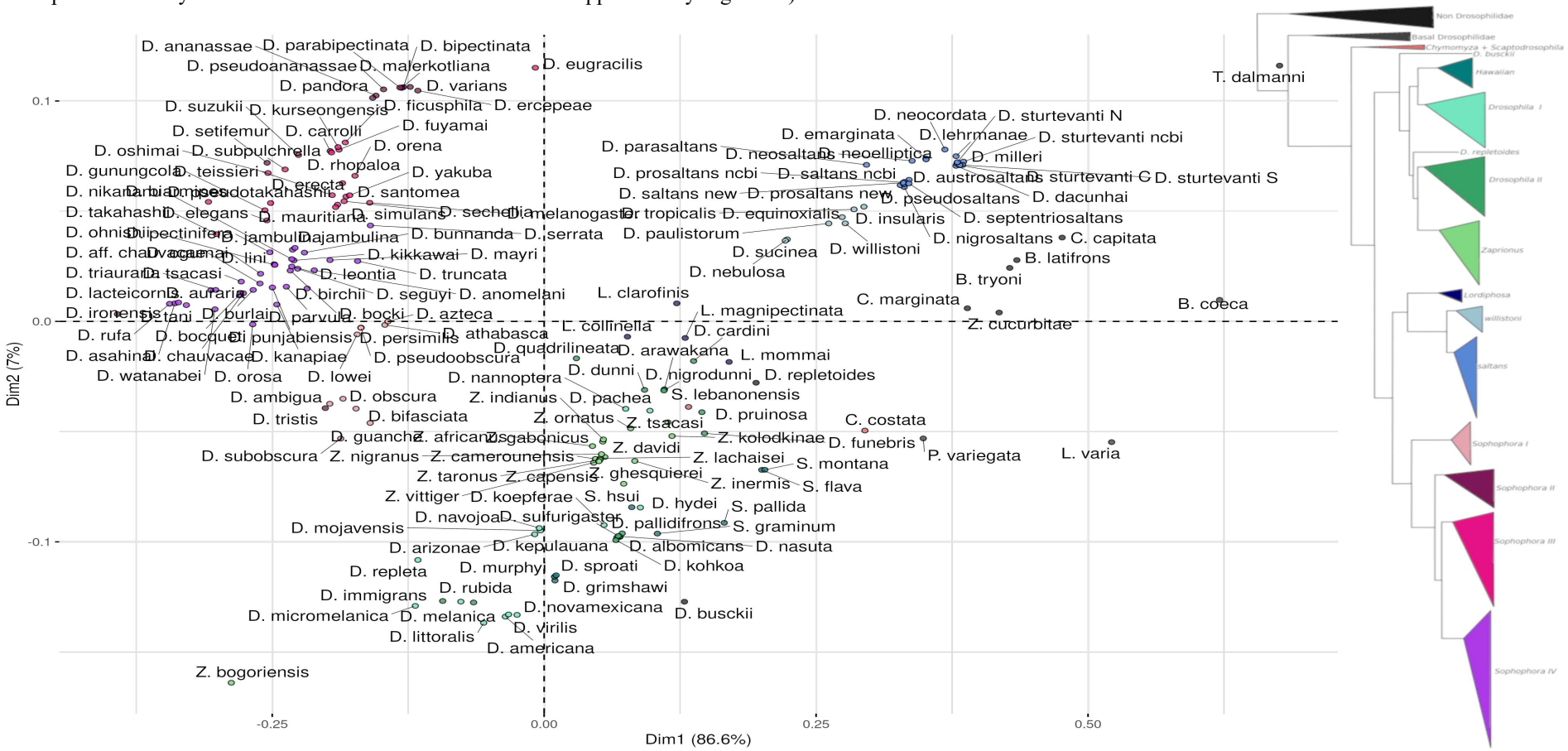
C


0.5 1 1.5 2 2.5 3
Drosophila saltans ncbi
Drosophila prosaltans ncbi
Drosophila saltans NEW
Drosophila prosaltans new
Drosophila pseudosaltans
Drosophila nigrosaltans
Drosophila septentriosaltans
Drosophila austrosaltans
Drosophila neoelliptica
Drosophila emarginata
Drosophila neosaltans
Drosophila dacunhai
Drosophila milleri
Drosophila sturtevanti-s3
Drosophila lehrmanae
Drosophila sturtevanti-s2
Drosophila sturtevanti-s1
Drosophila sturtevanti ncbi
Drosophila neocordata
Drosophila willistoni
Drosophila equinoxialis
Drosophila paulistorum
Drosophila tropicalis
Drosophila insularis
Drosophila parasaltans
Drosophila sucinea
Drosophila nebulosa
Bactrocera latifrons
Bactrocera tryoni
Zeugodacus cucurbitae
Ceratitis capitata
Phortica variegata
Coremacera marginata
Leucophenga varia
Braula
Teleopsis dalmanni

These findings underscore the intricate nuances in codon preferences, even among closely related species or subgroups. They suggest potential evolutionary or functional distinctions in the utilization of the genetic code within the Drosophilids. To further investigate the patterns of codon usage across species, we conducted a correspondence analysis using the average relative synonymous codon usage (RSCU). While the resulting analysis revealed the three major groups, it differed from the heatmap clusterization. In this analysis, the *saltans-willistoni* clade exhibited greater similarity to the *Dorsiphola*, *Siphlodora*, *Drosophila* subgenera, *Hawaiian Drosophila*, and *Zaprionus* genus in terms of the first dimension, which explains 86.6% of the variation in the indexed residuals. Notably, the codons CTG (leucine), CGT (arginine), TTA (leucine), TCA (serine), ATC (isoleucine), and ATT (isoleucine) played significant roles in shaping this dimension and exhibited differential preferences across the three clusters. In the second dimension, which accounted for 7% of the variation, the codon usage in the neotropical *Sophophora* species group appeared more akin to the old-world *Sophophora* than the clade comprising the remaining species. Argenine- (CGC, AGG, CGG) and glycine- (GGC, GGA) carrying tRNAs contributing significantly for this dimension (Figures 3, Supplementary Table S2, Supplementary Figure S1).

It is notable that the correspondence analysis separates 3 major clusters, one contains the *saltans-willistoni* groups, other with the remained *Sophophora* and the last clusted with the *Lordiphosa*, *Zaprionus* genera, the *Drosophila*, *Dorsilopha*, *Siphlodora* subgenera. This analyses also shows that the neotropical *Sophophora* codon usage pattern resemble the outgroups and basal drosophilids. Also the codon usage difference in the *Zaprionus bogoriensis* and other *Zaprionus* species is noticed. Other species that presented a notable deviation from their close relatives are, namely, *D. ironensis* (ananassae group)*, D. setifemur* (*setifemur* group)*, D. tristis* (obscura group*). D. eugracilis* (*melanogaster* group), *D. immigrans* (*immigrans* group)*, D. melanica,* (*melanica* group) *D. nannoptera* and *D. pachea* (*nannoptera* group), although the deviation was not as strong as the found in *Zaprionus* genus.

Figure 3. Correspondence analyses of the average relative synonymous codon usage between species recover 3 major clusters, the I - neotropical *Sophophora,* II - *Sophophora – old world* and III *Lordiphosa* and *Zaprionus genera* and *Drosophila*, *Siphlodora*, *Dorsilopha* subgenera. The plot shows each of the 174 genomes examined in this study along the first two dimensions (the X and Yaxes) of a correspondence analysis. Each axis is labeled with the percent variance explained by the corresponding dimension. The the codons correspondence analysis and each codon contribution is seen in Supplementary Figure S1).



Genetic drift and mutational bias

Aiming to evaluate the impact of genetic drift on the shaping of codon usage patterns in drosophilids, the phylogenetic signal of Relative Synonymous Codon Usage (RSCU) was assessed using Pagel's λ and Blomberg's K tests (Table 1). In one hand, Pagel's λ values range from 0 to 1, representing different degrees of phylogenetic signal. A value of 0 indicates the absence of any phylogenetic signal, suggesting that the trait is not influenced by evolutionary relatedness. On the other hand, a value of 1 indicates that the trait follows a Brownian model of random genetic drift, meaning that its variation can be attributed to the evolutionary relationships among species. On the other hand, Blomberg's K is a measure that compares the variation of a trait among species to the variance of trait differences among species contrasts. When a trait follows a Brownian model of random genetic drift, Blomberg's K will equal 1, indicating that trait variation is evenly distributed among species. However, Blomberg's K can be greater than 1, suggesting that trait variance is primarily observed between distinct clades rather than within them. We observed very high Pagel's λ for all codons, in concordance with high Blomberg's K, in fact the only codon the presented Blomberg's K smaller than 1 was CGA, the arginine-carrier tRNA, this indicate that other evolutionary forces may be shaping the choices of use-not use this tRNA. The high phylogenetic signal observed suggests a significant correlation between codon usage patterns and the evolutionary relationships among species. However, in agreement with this results the RSCU analyses, presented previously, the heatmap and the correspondence analysis (Figures 2 and 3), trend to recover similarities patterns among closely related species, but by the evaluation of the codon usage (RSCU) showed that some species can present patterns quite different than their close related.

Table 1. Testing of phylogenetic concordance of the RSCU for each codon across all 174 genomes. The Blomberg's K, Pagel's λ, and corresponding P-value are reported for each codon.

codon	Blomberg's K,	p (K)	Pagel's λ	p (λ)
AAA	14,2210	1,00E-03	0,99993	1,50E-147
AAC	6,3350	1,00E-03	0,99993	2,10E-138
AAG	14,2210	1,00E-03	0,99993	1,50E-147
AAT	6,3350	1,00E-03	0,99993	2,10E-138
ACA	9,5410	1,00E-03	0,99993	1,20E-141
ACC	12,8970	1,00E-03	0,99993	6,10E-149
ACG	3,1050	1,00E-03	0,99993	6,60E-113
ACT	4,4550	1,00E-03	0,99993	1,40E-114
AGA	3,5770	1,00E-03	0,99993	2,20E-110
AGC	6,5550	1,00E-03	0,99993	1,50E-129
AGG	8,3620	1,00E-03	0,99993	1,50E-149
AGT	5,5350	1,00E-03	0,99993	2,00E-127

codon	Blomberg's K,	p (K)	Pagel's λ	p (λ)
ATA	13,0910	1,00E-03	0,99993	4,90E-148
ATC	8,3630	1,00E-03	0,99993	4,70E-139
ATT	2,9960	1,00E-03	0,99993	1,50E-114
CAA	13,2060	1,00E-03	0,99993	1,80E-144
CAC	5,8060	1,00E-03	0,99993	4,10E-136
CAG	13,2060	1,00E-03	0,99993	1,80E-144
CAT	5,8060	1,00E-03	0,99993	4,10E-136
CCA	7,6570	1,00E-03	0,99993	1,50E-128
CCC	9,1420	1,00E-03	0,99993	7,00E-122
CCG	2,5660	1,00E-03	0,99993	3,80E-106
CCT	3,2640	1,00E-03	0,99993	4,20E-93
CGA	0,4580	1,00E-03	0,99993	6,40E-71
CGC	2,7400	1,00E-03	0,99993	4,70E-108
CGG	6,8320	1,00E-03	0,99993	4,00E-136
CGT	9,6890	1,00E-03	0,99993	7,00E-142
CTA	1,9320	1,00E-03	0,99993	4,40E-99
CTC	4,5250	1,00E-03	0,99993	3,50E-111
CTG	7,7690	1,00E-03	0,99993	3,80E-130
CTT	3,39	1,00E-03	0,99993	1,40E-95
GAA	12,52	1,00E-03	0,99993	2,10E-136
GAC	4,46	1,00E-03	0,99993	1,90E-129
GAG	12,52	1,00E-03	0,99993	2,10E-136
GAT	4,46	1,00E-03	0,99993	1,90E-129
GCA	10,22	1,00E-03	0,99993	4,90E-133
GCC	13,23	1,00E-03	0,99993	1,60E-138
GCG	1,74	1,00E-03	0,99993	1,30E-112
GCT	5,45	1,00E-03	0,99993	8,00E-115
GGA	1,07	1,00E-03	0,99993	8,40E-100
GGC	2,77	1,00E-03	0,99993	4,40E-105
GGG	2,12	1,00E-03	0,99993	2,40E-99
GGT	8,16	1,00E-03	0,99993	2,30E-128
GTA	7,87	1,00E-03	0,99993	2,60E-119
GTC	4,37	1,00E-03	0,99993	1,60E-96
GTG	8,00	1,00E-03	0,99993	1,60E-124
GTT	7,49	1,00E-03	0,99993	2,30E-120
TAC	5,64	1,00E-03	0,99993	1,10E-138
TAT	5,64	1,00E-03	0,99993	1,10E-138
TCA	12,03	1,00E-03	0,99993	2,50E-137
TCC	12,23	1,00E-03	0,99993	6,30E-145
TCG	2,65	1,00E-03	0,99993	3,40E-93
TCT	4,57	1,00E-03	0,99993	1,20E-102
TGC	6,67	1,00E-03	0,99993	5,70E-127
TGT	6,67	1,00E-03	0,99993	5,70E-127
TTA	17,76	1,00E-03	0,99993	1,80E-143
TTC	4,43	1,00E-03	0,99993	1,00E-114
TTG	4,26	1,00E-03	0,99993	1,10E-132
TTT	4,43	1,00E-03	0,99993	1,00E-114

Genetic drift, which refers to random changes in isomorphism frequencies over time, may play a significant role in driving codon usage patterns within closely related species. The genetic drift effect is more pronounced at shorter evolutionary timescales, leading to greater similarity in codon usage patterns among closely related species. However, at deeper branches, where the influence of genetic drift might be attenuated or obscured by other evolutionary processes, the recovery of similar codon usage patterns becomes more challenging.

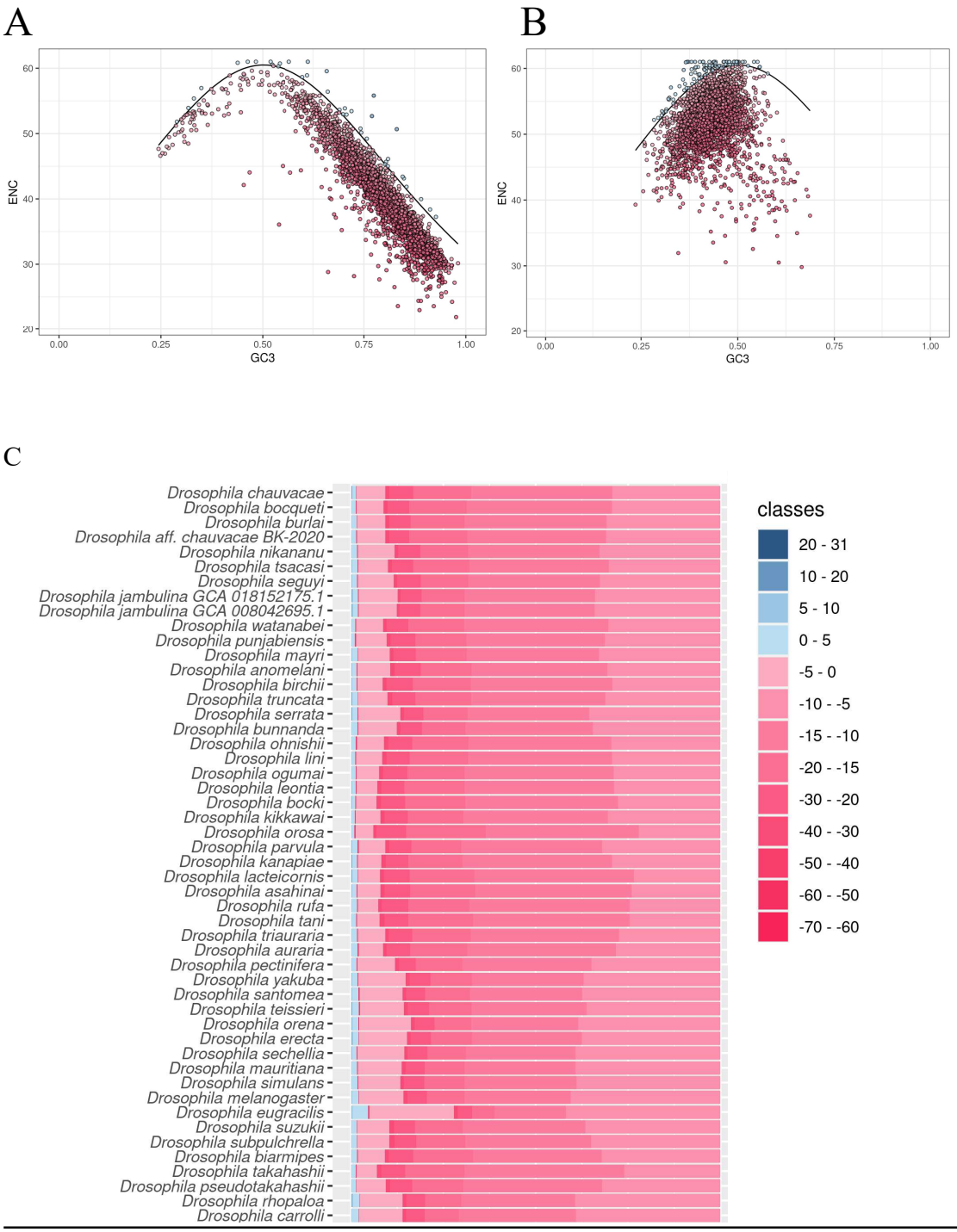
Therefore, while the phylogenetic signal analysis indicates a strong association between codon usage patterns and the overall evolutionary relationships among drosophilids, it is important to consider the limitations of the RSCU analyses in capturing similarities at deeper branches. The interplay between genetic drift, other evolutionary forces, and the complex dynamics of codon usage bias requires further investigation to fully understand the evolutionary mechanisms underlying the observed patterns.

We investigated the influence of mutational bias on codon usage patterns in various species. Mutational bias refers to the tendency of certain nucleotides to be more likely to mutate than others, which can impact the composition of codons. To assess this effect, we compared the effective number of codons (ENC) with the GC3 content (the proportion of G and C nucleotides at the third codon position) for each gene identified by BUSCO. By comparing the distribution of ENC as a function of GC3 and comparing the observed values with the expected ENC under the assumption that the only factor influencing codon usage bias is CG composition, we were able to evaluate the specific contribution of mutational bias to the observed codon usage patterns, and we estimated the percentage of deviation from expectation for each gene. Even the species that presented the lowest ENC values (*D. ironensis,* Figure 1), appears to be significantly affected by the mutational Bias (Figure 4A), althoug the selection force is pushing the obseved values of ENC down. The pattern between The neotropical sophophora and basal species resemble much more with the one demostrated with the values of D. sturtevanti (Figure 4B), and although a clear correlation is not seen, the deviation from the expectation does not appear to favore drastically increase of the observed ENC (Figure 4C), indicationg that the fact that the *saltans-willistoni* clade has GC3 composition close to 50% may be the major force acting on the lack of bias observed .

In addition to investigating codon usage bias and its evolutionary mechanisms in the Drosophilidae family, an important perspective of this work is to quantify the number of genes that deviate from the expectations of mutational bias and to evaluate the impact of translational selection. This approach aims to provide a clearer and more comprehensive

understanding of the codon usage patterns in the Drosophilidae family. By identifying the genes that fall outside the range of mutational bias, it becomes possible to uncover the selective pressures acting on codon usage. Furthermore, assessing the influence of translational selection helps elucidate the role of natural selection in shaping the observed codon preferences. This perspective enhances our understanding of the complex interplay between mutational biases, translational selection, and codon usage patterns, ultimately contributing to a more nuanced characterization of the evolutionary dynamics within the Drosophilidae family. Moreover, analyzing alterations in mutation rates can offer novel insights into the mechanisms underlying shifts in codon usage among Drosophilidae.

Figure 4. Influence of Mutational Bias on Codon Usage in *Drosophila*. The relationship between GC3 and ENC values reveals the impact of mutational bias, even in species with low Effective Number of Codons (ENC), as exemplified by *D. ironensis* (A). In contrast, species with higher ENC values, like *D. sturtevanti* (B), exhibit a less clear pattern. Notably, the percentage of genes deviating from the expected only by mutational bias (C) suggests that, despite a influence of selection, the neotropical *Sophophora* species tend to have higher ENC values due to their GC3 content approaching 50%. The reddish points indicate a higher likelihood of selection influence, while the bluish points signify a lower impact of selection on codon usage.

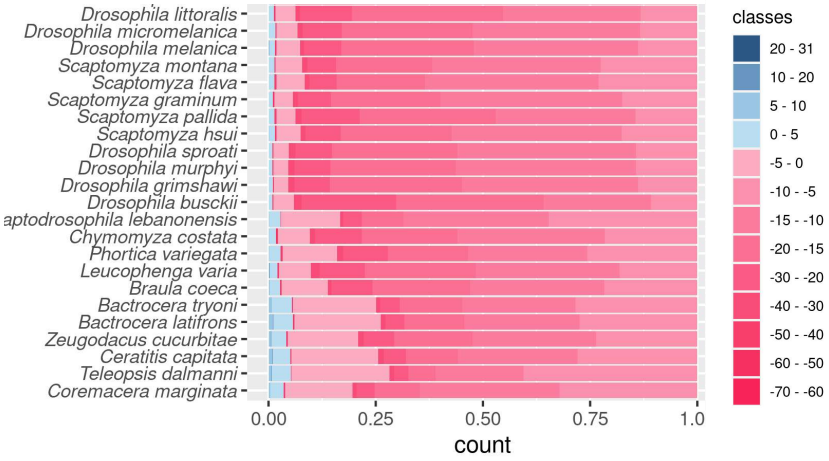


Continue


species
Drosophila carrolli
Drosophila kurseongensis
Drosophila fuyamai
Drosophila gunungcola
Drosophila elegans
Drosophila ficusphila
Drosophila oshimai
Drosophila parabipectinata
Drosophila bipectinata
Drosophila malerkotliana
Drosophila pseudoananassae
Drosophila pandora
Drosophila ananassae
Drosophila varians
Drosophila ercepeae
Drosophila ironensis
Drosophila setifemur
Drosophila tristis
Drosophila ambigua
Drosophila obscura
Drosophila bifasciata
Drosophila subobscura
Drosophila guanche
Drosophila pseudoobscura
Drosophila persimilis
Drosophila lowei
Drosophila azteca
Drosophila athabasca
Drosophila saltans NEW
Drosophila prosaltans ncbi
Drosophila saltans ncbi
Drosophila prosaltans new
Drosophila septentriosaltans
Drosophila pseudosaltans
Drosophila nigrosaltans
Drosophila austrosaltans
Drosophila neoelliptica
Drosophila emarginata
Drosophila neosaltans
Drosophila neocordata
Drosophila sturtevanti-s2
Drosophila sturtevanti-s1
Drosophila sturtevanti-s3
Drosophila sturtevanti ncbi
Drosophila milleri
Drosophila dacunhai
Drosophila lehrmanae
Drosophila parasaltans
Drosophila paulistorum
Drosophila equinoxialis
Drosophila willistoni
Drosophila tropicalis
Drosophila insularis
Drosophila nebulosa
Drosophila sucinea
Lordiphosa magnipectinata
Lordiphosa clarofinis
Lordiphosa collinella
Lordiphosa mommai
Zaprionus nigranus
Zaprionus camerounensis
Zaprionus lachaisei
Zaprionus vittiger
Zaprionus taronus
Zaprionus davidi
Zaprionus capensis
Zaprionus indianus
Zaprionus gabonicus
Zaprionus africanus
Zaprionus ornatus
Zaprionus tsacasi
Zaprionus kolodkinae
Zaprionus inermis
Zaprionus ghesquierei
Zaprionus bogoriensis
Drosophila pallidifrons
Drosophila kohkoa
Drosophila sulfurigaster
Drosophila nasuta
Drosophila albomicans
Drosophila kepulauana
Drosophila rubida
Drosophila immigrans
Drosophila pruinosa
Drosophila nigrodunni
Drosophila arawakana
Drosophila dunni
Drosophila cardini
Drosophila funebris
Drosophila quadrilineata
Drosophila repletoides
Drosophila mojavensis
Drosophila arizonae
Drosophila navojoa
Drosophila koepferae
Drosophila repleta
Drosophila hydei
Drosophila pachea
Drosophila nannoptera
Drosophila novamexicana
Drosophila americana
Drosophila virilis
Drosophila littoralis
classes
20 - 31
10 - 20
5 - 10
0 - 5
-5 - 0
-10 - -5
-15 - -10
-20 - -15
-30 - -20
-40 - -30
-50 - -40
-60 - -50
-70 - -60

continue

continue



Codon Usage impact in the phylogeny

To assess the impact of codon usage bias and saturation on the phylogenetic relationships of drosophilids, multiple maximum likelihood trees were reconstructed, using amino acids, first, second, and third codon positions, as well as the first and second codon positions. Notably, our analyses revealed several noteworthy topological changes within the Drosophilidae phylogeny (Supplentary Figures S2-S6), particularly in the context of the *saltans* species group (Figure 5).The most significant changes occurred in the *saltans* subgroup, but changes within the *sturtevanti* subgroup were also noted, although only in the third possition of the codon.

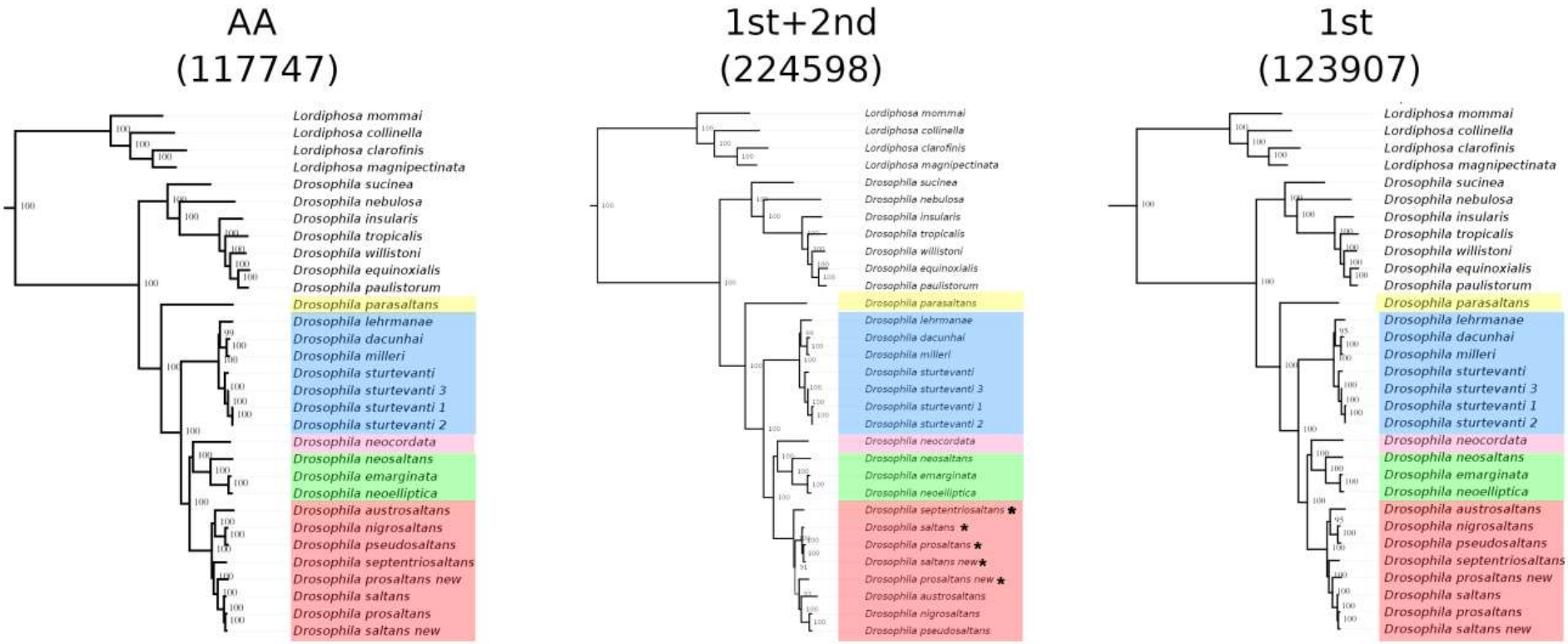
Furthermore, for the *saltans* subgroup the analysis of the first and second codon positions presented a markedly different topology compared to the amino acid-based tree. Surprisingly, these differences appeared to be primarily driven by mutations in the second codon position. The distinct influence of the second codon position on phylogenetic relationships suggests that mutations in this position may carry more phylogenetic signal than those in the first codon position. These results emphasize the need for a more nuanced consideration of codon position and its context in phylogenetic studies.

It's essential to note that these findings underscore the critical influence of codon usage bias and saturation on the reconstruction of evolutionary relationships within the *saltans* species group. The saturation effects, as observed in the third codon position, can further complicate phylogenetic reconstructions. In summary, our study highlights the intricate interplay between codon usage bias, saturation, and codon position in shaping the phylogenetic relationships of drosophilids. These findings have implications for the

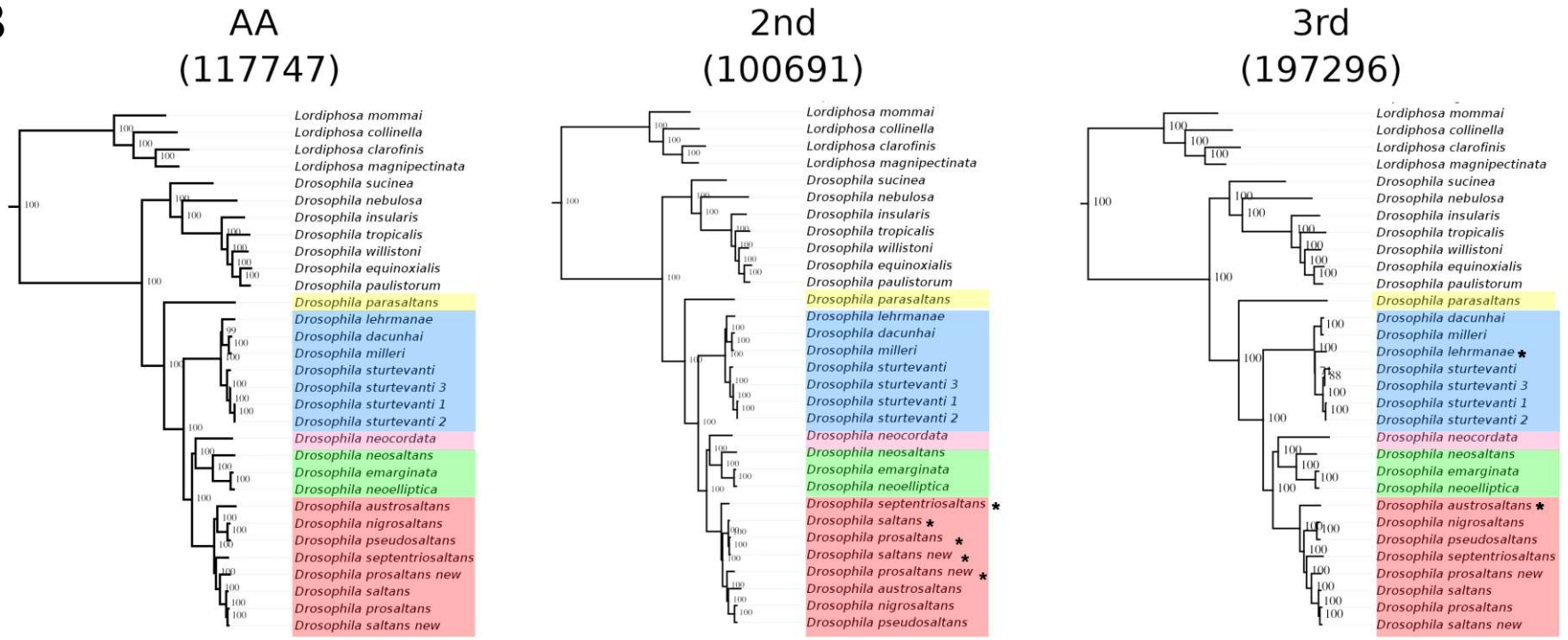
interpretation of molecular data in evolutionary studies and suggest the need for further research to elucidate the specific mechanisms behind these codon-specific effects. This, in turn, could contribute to the development of more robust and accurate phylogenetic methodologies.

Figure 5. Comparison of topologies generated with amino-acids, first+second (A), first (A), second (B) and third (B) codon bases focusing in the neotropical *Sophophora* clade. See Supplentary Figures S2-S6 for the all evaluated Drosophilidae topologies. Species deviating from the topology generated with amino acid data are highlighted with *.

A



B


AA
(117747)
2nd
(100691)
3rd
(197296)
Lordiphosa mommai
Lordiphosa collinella
Lordiphosa clarofinis
Lordiphosa magnipectinata
Drosophila sucinea
Drosophila nebulosa
Drosophila insularis
Drosophila tropicalis
Drosophila willistoni
Drosophila equinoxialis
Drosophila paulistorum
Drosophila parasaltans
Drosophila lehrmanae
Drosophila dacunhai
Drosophila milleri
Drosophila sturtevanti
Drosophila sturtevanti 3
Drosophila sturtevanti 1
Drosophila sturtevanti 2
Drosophila neocordata
Drosophila neosaltans
Drosophila emarginata
Drosophila neoelliptica
Drosophila austrosaltans
Drosophila nigrosaltans
Drosophila pseudosaltans
Drosophila septentriosaltans
Drosophila prosaltans new
Drosophila saltans
Drosophila prosaltans
Drosophila saltans new
Drosophila septentriosaltans *
Drosophila saltans *
Drosophila prosaltans *
Drosophila saltans new *
Drosophila prosaltans new *
Drosophila lehrmanae *
Drosophila austrosaltans *

Material and Methods

Data acquisition and ortholog identification

To acquire genome assemblies of Acalyptratae, specifically focusing on the family Drosophilidae, a search was conducted on the NCBI database until March 2023, using the key searches txid7214[Organism:exp] and txid43741[Organism:exp] to target Drosophilidae taxa and Acalyptratae taxa, respectively. The obtained results underwent filtering, to ensure that multiple genomes of the same species were removed, selecting the one that presented better assembly report (higher N50). The primary objective of the study was to investigate the neotropical Sophophoran clade (*saltans* and *willistoni* groups). As such, genome sequencing and draft assembly were performed for 15 species belonging to the *Drosophila saltans* species group, as described in Prediger et al. (Chapter 3).

In order to compare codon usage bias across approximately of the retrived 197 species, a critical step involved utilizing the single-copy genes identified by BUSCO v.5, a computational tool that searches for highly conserved single-copy genes across genomes (WATERHOUSE et al., 2018). By focusing exclusively on single-copy genes, the analysis avoids the inclusion of paralogous genes, ensuring a more accurate and unbiased comparison of codon usage bias among the species. This approach allows for a robust evaluation of codon usage patterns, as it provides a standardized set of genes for comparison across the diverse range of species included in the study. The downstream analysis was performed using the genomes that exhibited at least 90% completeness for the identified single-copy genes, which corresponded to a minimum of 2,956 SCG present in the genome.

Codon usage bias measure

In order to obtain the general view of codon usage between species, the ENC were calculated to 33,408 single copy genes from 174 species using the R package Vhica (WALLAU et al., 2016).The effective number of codons (ENC or Nc, WRIGHT, 1990), is a mesure of the preference in a gene codon usage and it quantify the departure of a gene from the equal use of synonymous codon. It was defined bt Wright (1990) seen in equation 1:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \quad (1)$$

Where “F_2”, “F_3”, “F_4”,and “F_6” are the average homozygosity for the amino acids having a degeneracy of two, three, four and six respectively. And are calculated as shown in equation 2, where “n” represents the overall occurrence of the amino acid within the gene, and

"p_i" denotes the frequency of the "i^{th}" synonymous codon corresponding to that specific amino acid.

$$F = \frac{\left(n \sum_{i-1}^{k} p_i^2\right) - 1}{n - 1} \quad (2)$$

The ENC value ranges from 20 to 61 due to the presence of 61 tRNA molecules carrying the 20 amino acids. A value of 20 indicates complete bias, where each amino acid is carried by a unique tRNA molecule. Conversely, a value of 61 indicates complete unbiased selection of codons, where all synonymous codons are used equally. Due to the ENC be independent of gene length and amino acid composition, does not rely on organism-specifc data and can be easily applied to study new organisms.

Other while spread that we used to evaluate the codon usage bias is the relative synonymous codon usage (RSCU) (SHARP; TUOHY; MOSURSKI, 1986) is the observed frequency of a codon divided by the expected frequency if all the synonymous codons were used equally, as demostrated in equation 2. Where where "i" is the i^{th} aminoacid carried by the codon "c". X_{ic} is the frequency of the codon "c" among the gene and "n_i" represents the total number of synonymous codons for the i^{th}-aminoacid. RSCU values can be 2, 3, 4 and 6 when a single codon is used to encode amino acids having 2,3,4 and 6 synonymous codons respectively (Supplementary Table S3)

$$RSCU_{ic} = \frac{X_{ic}}{1/n_i} \quad (2)$$

To analyze the codon usage patterns, we calculated the RSCU values for the 33,408 single-copy genes across 174 species using the R package SeqinR (CHARIF; LOBRY, 2007). To gain insights into general trends, we employed hierarchical clustering with the gplots package (WARNES et al., 2005) to visualize the average RSCU values for each species. Additionally, we performed correspondence analysis based on the average RSCU values using the FactoMineR package (LÊ; JOSSE; HUSSON, 2008) to identify specific codons that contribute to variations in codon usage between species.

Genetic drift and mutational bias

To reconstruct the evolutionary relationships within Drosophilidae, partitioned Maximum Likelihood analysis were performed in IQ-TREE (NGUYEN et al., 2015), using

the 192 SCGs shared by the 174 genomes that had completness higher the 90%. The aligment of SCGs coding sequences were carried out using Clustal-Codons methods inplemented in Mega X (megacc, KUMAR et al., 2018), whereas the alignmentfor SCG amino acids sequences they were carried out in mafft (KATOH; STANDLEY, 2013).The CDS sequences were evaluated using five different approaches: considering only the first and second bases of each codon, considering only the first base, considering only the second base, considering only the third base, and considering all bases. Additionally, partitionated Maximum Likelihood analysis was also applied for the protein sequences. To examine the influence of phylogeny on the observed variation in codon bias, we computed two measures of phylogenetic signal in R, Pagel's λ and Blomberg's K implemented in the R package phytools (REVELL, 2012), using the phylogenetic tree reconstructed with amino acids. Aiming to infer the role of mutational bias in every genome, the ENC effective number of codons (ENC) of each gene to the synonymous GC3 proportion of that gene using the R package vhcub (ANWAR; SOUDY; MOHAMED, 2019)

REFERENCES

ANWAR, A. M.; SOUDY, M.; MOHAMED, R. ***vhcub*: Virus-host codon usage co-adaptation analysis**, F1000Research, 2019. Disponível em: <https://f1000research.com/articles/8-2137>. Acesso em: 22 jun. 2023.

ARELLA, D.; DILUCCA, M.; GIANSANTI, A. Codon usage bias and environmental adaptation in microbial organisms. Molecular genetics and genomics: MGG, 6, n. 3, p. 751–762, 2021.

BALLARD, A.; BIENIEK, S.; CARLINI, D. B. The fitness consequences of synonymous mutations in Escherichia coli: Experimental evidence for a pleiotropic effect of translational selection. Gene, 4, p. 111–120, 2019.

CHARIF, D.; LOBRY, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: BASTOLLA, U.; PORTO, M.; ROMAN, H. E.; VENDRUSCOLO, M. (Eds.). **Structural Approaches to Sequence Evolution: Molecules, Networks, Populations**. Biological and Medical Physics, Biomedical EngineeringBerlin, Heidelberg: Springer, 2007. p. 207–232.

FIERS, W.; CONTRERAS, R.; DUERINCK, F.; HAEGEMAN, G.; ISERENTANT, D.; MERREGAERT, J.; MIN JOU, W.; MOLEMANS, F.; RAEYMAEKERS, A.; VAN DEN BERGHE, A.; VOLCKAERT, G.; YSEBAERT, M. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature, 0, n. 5551, p. 500–507, 1976.

GRANTHAM, R.; GAUTIER, C.; GOUY, M.; JACOBZONE, M.; MERCIER, R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Research, n. 1, p. 213, 1981.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, , n. 4, p. 772–780, 2013.

KOKATE, P. P.; TECHTMANN, S. M.; WERNER, T. Codon usage bias and dinucleotide preference in 29 Drosophila species. G3 Genes|Genomes|Genetics, , n. 8, p. jkab191, 2021.

KUMAR, S.; STECHER, G.; LI, M.; KNYAZ, C.; TAMURA, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, , n. 6, p. 1547–1549, 2018.

LABELLA, A. L.; OPULENTE, D. A.; STEENWYK, J. L.; HITTINGER, C. T.; ROKAS, A. Variation and selection on codon usage bias across an entire subphylum. PLoS Genetics, , n. 7, p. e1008304, 2019.

LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, , n. 1 SE-Articles, p. 1–18, 2008.

NGUYEN, L. T.; SCHMIDT, H. A.; VON HAESELER, A.; MINH, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution, , n. 1, p. 268–274, 2015.

POWELL, J. R.; SEZZI, E.; MORIYAMA, E. N.; GLEASON, J. M.; CACCONE, A. Analysis of a Shift in Codon Usage in Drosophila. Journal of Molecular Evolution, , n. SUPPL. 1, p. 214–225, 2003.

RETCHLESS, A. C.; LAWRENCE, J. G. Ecological Adaptation in Bacteria: Speciation Driven by Codon Selection. Molecular Biology and Evolution, , n. 12, p. 3669–3683, 2012.

REVELL, L. J. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, n. 2, p. 217–223, 2012.

RODRÍGUEZ-TRELLES, F.; TARRÍO, R.; AYALA, F. J. Switch in codon bias and increased rates of amino acid substitution in the Drosophila saltans species group. Genetics, 3, n. 1, p. 339–350, 1999. a.

RODRÍGUEZ-TRELLES, F.; TARRÍO, R.; AYALA, F. J. Molecular evolution and phylogeny of the Drosophila saltans species group Inferred from the Xdh Gene. Molecular Phylogenetics and Evolution, , n. 1, p. 110–121, 1999. b.

ROTA-STABELLI, O.; LARTILLOT, N.; PHILIPPE, H.; PISANI, D. Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. Systematic Biology, , n. 1, p. 121–133, 2013.

SHARP, P. M.; AVEROF, M.; LLOYD, A. T.; MATASSI, G.; PEDEN, J. F.; GERHART, J.; HUNT, R. T.; KIRSCHNER, M. W.; WOLPERT, L. DNA sequence evolution: the sounds of silence. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 9, n. 1329, p. 241–247, 1997.

SHARP, P. M.; STENICO, M.; PEDEN, J. F.; LLOYD, A. T. Codon usage: mutational bias, translational selection, or both? Biochemical Society Transactions, , n. 4, p. 835–841, 1993.

SHARP, P. M.; TUOHY, T. M. F.; MOSURSKI, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Research, , n. 13, p. 5125–5143, 1986.

SUN, Y.; TAMARIT, D.; ANDERSSON, S. G. E. Switches in Genomic GC Content Drive Shifts of Optimal Codons under Sustained Selection on Synonymous Sites. Genome Biology and Evolution, n. 10, p. 2560–2579, 2017.

TARRÍO, R.; RODRÍGUEZ-TRELLES, F.; AYALA, F. J. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The Drosophila saltans and willistoni groups, a case study. Molecular Phylogenetics and Evolution, , n. 3, p. 344–349, 2000.

VICARIO, S.; MORIYAMA, E. N.; POWELL, J. R. Codon usage in twelve species of Drosophila. BMC Evolutionary Biology, n. 1, 2007. . Acesso em: 17 abr. 2021.

WALLAU, G. L.; CAPY, P.; LORETO, E.; LE ROUZIC, A.; HUA-VAN, A. VHICA, a New Method to Discriminate between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within Drosophila. Molecular Biology and Evolution, , n. 4, p. 1094–1109, 2016.

WAN, X. F.; XU, D.; KLEINHOFS, A.; ZHOU, J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evolutionary Biology, p. 1–11, 2004.

WARNES, G.; BOLKER, B.; BONEBAKKER, L.; GENTLEMAN, R.; HUBER, W.; LIAW, A.; LUMLEY, T.; MÄCHLER, M.; MAGNUSSON, A.; MÖLLER, S. **gplots: Various R programming tools for plotting data**. v. 2journalAbbreviation: R package version.

WATERHOUSE, R. M.; SEPPEY, M.; SIMAO, F. A.; MANNI, M.; IOANNIDIS, P.; KLIOUTCHNIKOV, G.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution, , n. 3, p. 543–548, 2018.

WRIGHT, F. The “effective number of codons” used in a gene. Gene, , n. 1, p. 23–29, 1990.

YASSIN, A. Phylogenetic relationships among species subgroups in the Drosophila saltans group (Diptera: Drosophilidae): Can morphology solve a molecular conflict. Zoological Research, , n. 3, p. 225–232, 2009.

5 Final considerations

My thesis leveraged the advantages of new genomic resources for 16 species of the *Drosophila saltans* species. I investigated two aspects: the phylogenomic relationships (Chapter 3) and the evolution of a set of highly conserved genes (Chapter 4). A specific discussion on each of these questions was given at the end of the corresponding chapters. In this General discussion and prospectives chapter, I would like to discuss three major implications of my thesis in the evolution of the *Drosophila saltans* species group as well as of Neotropical insects in general.

5.1 Phylogenetic systematics of the *Drosophila saltans* species group

The *parasaltans* species subgroup consists of two Amazonian species, *D. parasaltans* and *D. subsaltans*, but no study has ever included both species. Indeed, only two nuclear sequence from the *Xanthine dehydrogenase (Xdh)* and a fragment of alcohol dehydrogenase (*Adh*) genes are available for *D. subsaltans* (Tarrío et al. 1998). By extracting this sequence from all the genomes, we sequenced and reconducting a phylogenetic analysis, we confirmed the monophyly of the *parasaltans* subgroup based on *Xdh* but not with the very short fragment of *Adh* that is available (Appendix A, Supplementary Figure S4). The two species are however strongly reproductively and geographically isolated (Bicudo and Prioli 1978).

The *sturtevanti* species subgroup includes seven species, of which four were included in our study. Of the species that were not included, *D. pulchella*, which was described from a male specimen from the Island of Saint Vincent, is likely a synonym of the widespread *D. sturtevanti* (Vilela and Bächli 1990). Whereas our biogeographical analysis indicated that the ancestral range of all four sequenced species was likely the northern coasts of South America, the unsampled species *D. magalhaesi* was originally described from southern Brazil (Mourão and Bicudo 1967). The Southern strain of this species is strongly reproductively isolated from both *D. sturtevanti* and *D. milleri* (Bicudo 1979). Recent studies reported its presence in northern Brazil, but we were not able to collect it despite our extensive sampling of drosophilids in Brazil, including in the type locality of this species. The third unsampled species, *D. rectangularis*, is endemic to Mexico and its males have genitalia that are completely distinct from those of the remaining species of the subgroup (Magalhães 1962). Indeed, the morphology of the male genitalia approaches *D. rectangularis* to species of the *cordata* subgroup.

The *cordata* subgroup consists of two species, *D. cordata* which is endemic to Guatemala and the widespread *D. neocordata*, which was originally described from Brazil. We have used two South American strains of *D. neocordata*. However, larger collections in Central America would be required to gain better insights on the taxonomic delimitation of this subgroup, especially given that *D. cordata* has never been included in any molecular analysis and that the Mexican *D. rectangularis* is likely member of this clade. The taxonomic increase could also help better identifying the exact placement of the *cordata* subgroup relative to both the *elliptica* and *saltans* subgroups that differed between nuclear and mitochondrial trees.

The *elliptica* subgroup includes four species, of which three were sequenced here. Remarkably, the two species only found in Brazil, *i.e. D. neosaltans* and *D. neoelliptica*, were not sisters, with *D. neoelliptica* being closely related to the widespread *D. emarginata*. Previous studies demonstrated morphological differences between widespread populations of *D. emarginata* and a population from Peru (Magalhães 1962) and there is evidence for partial reproductive isolation between populations in Central America (Bicudo and Prioli 1978). No molecular study has ever included the Mexican endemic *D. elliptica* and further geographical sampling is needed.

The *saltans* subgroup is by far the more speciose in the group, including eight species. Six out of these species were sequenced here. The two unsampled species included *D. lusaltans*, a species endemic to Haiti (Magalhães 1962), and *D. neoprosaltans*, which was recently described from Ecuador (Guillín and Rafael 2017). Earlier molecular studies included a short fragment of the *Alcohol dehydrogenase (Adh)* nuclear gene and the complete sequence of the mitochondrial gene *Cytochrome oxidase subunit 1 and 2* (*COI* and *COII*) (O'Grady et al. 1998; Roman et al. 2022). We were not able to conclusively place *D. lusaltans* using the short *Adh* sequence, but the mitochondrial analysis indicated that this species belongs to the mitotype P clade (Appendix A, Supplementary Figure S4). Indeed, Magalhães (1962) noted that the male genitalia of *D. lusaltans* closely resemble those of *D. prosaltans*. We were recently collecting and molecularly characterizing drosophilids in Ecuador, and found specimens of *D. prosaltans* and *D. austrosaltans* but we were not able to collect *D. neoprosaltans*.

5.2 Genome evolution of the *Drosophila saltans* species group

My analysis of the relaxation of codon usage bias in the *saltans* subgroup treated a single aspect of molecular evolution in this clade. Given the wide phylogenetic breadth of the

study, only highly conserved single-copy genes were considered. How much inference from this subset of genes reflect whole-genome patterns requires further investigations. In particular, the evolution of codon usage bias may reflect a general pattern of relaxation of selection and increase of genetic load in this clade. Such relaxation could occur due to decrease of effective population size, which in its turn could result from demographic effects or due to reduction in recombination rate reducing the efficiency of purging slightly deleterious synonymous mutations. Indeed, species of the neotropical *Sophophora* have long been notorious for their highly frequent polymorphic large chromosomal inversions. Large inversions often reduce recombination and increase gene load (Jay et al. 2019). Testing this hypothesis would require assembling the genomes of multiple lines of species of the *saltans* group using long-read sequencing approaches and inferring the boundaries and extent of inversions (cf. Ferreira et al. 2023). The genomes would then need to be annotated in order to quantify the extent of codon usage bias in genes present on different chromosomes or in different positions in respect to the inversions. To account for the effects of translational selection, the expression levels of the different genes need to be quantified using recent transcriptomic technologies and compared to their codon usage. Furthermore, studies on population genomics of species of the *Drosophila saltans* group can also shed a light on the recombination landscape in these species, since in most species local levels of nucleotide diversity correlate with recombination rate (Kern and Hahn 2018; Ferreira et al. 2023). Despite these gaps of our knowledge, the genomic analyses provided in this thesis represent a major step towards the application of next-generation-sequencing tools to understand the evolution of this longly neglected clade.

5.3 Phenotypic evolution of the *Drosophila saltans* species group

Compared to other species of the subgenus *Sophophora*, the genetic basis of phenotypic evolution has rarely been investigated in the *saltans* species group. A very interesting aspect of the *saltans* group is related to the male external genitalia morphology. The *saltans* group species are cryptic and show wide variation in male genital morphology, with phallic structures evolving rapidly and significantly differing among the five subgroups, but not within them (MAGALHÃES; BJÖRNBERG, 1957; SOUZA et al., 2014; SEGALA, 2019; ROMAN; MADI-RAVAZZI, 2021). In addition to the rapid evolution of male genitalia, a peculiar characteristic of this group is the evolution of male genital size. The *elliptica* subgroup is an exception to the pattern of negative allometry for the male reproductive organ in arthropods, exhibiting a large size of aedeagi. For example, the greatest aedeagus length (measured between

the phalapodeme and the apix of aedeagus) is observed in *D. emarginata* (1.07 mm), followed by *D. neosaltans* (0.85 mm), *D. elliptica* (0.73 mm), and *D. neoelliptica* (0.61 mm) (after the scaled illustrations given in MAGALHÃES; BJÖRNBERG, 1957). These numbers are significantly higher when compared, for example, with the species *D. melanogaster*, which has an approximate length of 0.2 mm (TSACAS et al., 1971), although the difference in body length of this species is not that remarkable. The enlargement of this structure represents a nice opportunity to test this species as model to understand the genetic basis of genitalia development, and the mechanisms that constrain aedeagus enlargements. I investigated from the taxonomic literature the distribution of the lengths of the aedeagus relative to body size in multiple *Drosophila* species and found that indeed the enlargement in the *elliptica* species group is unique. I have also started to investigate the pupal development of the male and female genital discs. In the future, I would like to better understand the changes in the developmental networks underlying the evolution of this spectacular phenotype.

REFERENCES

AKASHI, H. Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. Genetics, 6, n. 3, p. 927–935, 1994.

ANDERSON, E.; STEBBINS, G. L. Hybridization as an Evolutionary Stimulus. Evolution, n. 4, p. 378, 1954.

ANWAR, A. M.; SOUDY, M.; MOHAMED, R. ***vhcub*: Virus-host codon usage co-adaptation analysis**, F1000Research, 2019. Disponível em: <https://f1000research.com/articles/8-2137>. Acesso em: 22 jun. 2023.

ARELLA, D.; DILUCCA, M.; GIANSANTI, A. Codon usage bias and environmental adaptation in microbial organisms. Molecular genetics and genomics: MGG, 6, n. 3, p. 751–762, 2021.

AVISE, J. C.; ROBINSON, T. J. Hemiplasy: A New Term in the Lexicon of Phylogenetics. Systematic Biology, , n. 3, p. 503–507, 2008.

BAIÃO, G. C.; SCHNEIDER, D. I.; MILLER, W. J.; KLASSON, L. Multiple introgressions shape mitochondrial evolutionary history in Drosophila paulistorum and the Drosophila willistoni group. Molecular Phylogenetics and Evolution, 0, 2023. . Acesso em: 22 fev. 2023.

BALLARD, A.; BIENIEK, S.; CARLINI, D. B. The fitness consequences of synonymous mutations in Escherichia coli: Experimental evidence for a pleiotropic effect of translational selection. Gene, 4, p. 111–120, 2019.

BALLARD, J. W. O. Comparative genomics of mitochondrial DNA in members of the Drosophila melanogaster subgroup. Journal of Molecular Evolution, , n. 1, p. 48–63, 2000.

BICUDO, H. E. M. C. Reproductive isolation in the saltans group of Drosophila. I. The saltans subgroup. Genetica, , n. 3, p. 313–329, 1973. a.

BICUDO, H. E. M. C. Chromosomal polymorphism in the saltans group of Drosophila I. The saltans subgroup. Genetica, , n. 4, p. 520–552, 1973. b.

BICUDO, H. E. M. C. Reproductive isolation of the saltans group of Drosophila. IV. the sturtevanti subgroup. Revista Brasileira de Genética, , n. 4, p. 247–258, 1979.

BICUDO, H. E. M. C.; HOSAKI, M. K.; MACHADO, J.; MARQUES, M. C. N. Chromosomal polymorphism in the saltans group of Drosophila II. Further study on D. prosaltans. Genetica, , n. 1, p. 5–15, 1978.

BICUDO, H. E. M. C.; PRIOLI, A. J. Reproductive isolation in the saltans group of Drosophila II. The parasaltans subgroup. Genetica, , n. 1, p. 17–22, 1978.

BLISCHAK, P. D.; CHIFMAN, J.; WOLFE, A. D.; KUBATKO, L. S. HyDe: A Python Package for Genome-Scale Hybridization Detection. Systematic Biology, , n. 5, p. 821–829, 2018.

BOUCKAERT, R.; VAUGHAN, T. G.; BARIDO-SOTTANI, J.; DUCHÊNE, S.; FOURMENT, M.; GAVRYUSHKINA, A.; HELED, J.; JONES, G.; KÜHNERT, D.; DE MAIO, N.; MATSCHINER, M.; MENDES, F. K.; MÜLLER, N. F.; OGILVIE, H. A.; DU PLESSIS, L.; POPINGA, A.; RAMBAUT, A.; RASMUSSEN, D.; SIVERONI, I.; SUCHARD, M. A.; WU, C. H.; XIE, D.; ZHANG, C.; STADLER, T.; DRUMMOND, A. J. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Computational Biology, , n. 4, p. 1–28, 2019.

BULMER, M. The selection-mutation-drift theory of synonymous codon usage. Genetics, 9, n. 3, p. 897–907, 1991.

CAMACHO, C.; COULOURIS, G.; AVAGYAN, V.; MA, N.; PAPADOPOULOS, J.; BEALER, K.; MADDEN, T. L. BLAST+: Architecture and applications. BMC Bioinformatics, , p. 1–9, 2009.

CAVALCANTI, A. G. L. Geographic variation of chromosome structure in Drosophila prosaltans. Genetics, , n. 6, p. 529–536, 1948.

CHANG, E. S.; NEUHOF, M.; RUBINSTEIN, N. D.; DIAMANT, A.; PHILIPPE, H.; HUCHON, D.; CARTWRIGHT, P. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proceedings of the National Academy of Sciences, 2, n. 48, p. 14912–14917, 2015.

CHARIF, D.; LOBRY, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: BASTOLLA, U.; PORTO, M.; ROMAN, H. E.; VENDRUSCOLO, M. (Eds.). **Structural Approaches to Sequence Evolution: Molecules, Networks, Populations**. Biological and Medical Physics, Biomedical EngineeringBerlin, Heidelberg: Springer, 2007. p. 207–232.

CHARLESWORTH, B.; CAMPOS, J. L.; JACKSON, B. C. Faster-X evolution: Theory and evidence from Drosophila. Molecular Ecology, , n. 19, p. 3753–3771, 2018.

CHEN, S. L.; LEE, W.; HOTTES, A. K.; SHAPIRO, L.; MCADAMS, H. H. Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences, 1, n. 10, p. 3480–3485, 2004.

CONNER, W. R.; DELANEY, E. K.; BRONSKI, M. J.; GINSBERG, P. S.; WHEELER, T. B.; RICHARDSON, K. M.; PECKENPAUGH, B.; KIM, K. J.; WATADA, M.; HOFFMANN, A. A.; EISEN, M. B.; KOPP, A.; COOPER, B. S.; TURELLI, M. A phylogeny for the Drosophila montium species group: A model clade for comparative analyses. Molecular Phylogenetics and Evolution, 8, n. April 2020, p. 107061, 2021.

DAVID, J. R.; FERREIRA, E. A.; JABAUD, L.; OGEREAU, D.; BASTIDE, H.; YASSIN, A. Evolution of assortative mating following selective introgression of pigmentation genes between two Drosophila species. Ecology and Evolution, , n. 4, p. e8821, 2022.

DE CASTRO, J. P.; CARARETO, C. M. A. P elements in the saltans group of Drosophila: A new evaluation of their distribution and number of genomic insertion sites. Molecular Phylogenetics and Evolution, , n. 1, p. 383–387, 2004.

DE SETTA, N.; LORETO, E. L. S.; CARARETO, C. M. A. Is the evolutionary history of the O-type P element in the saltans and willistoni groups of Drosophila similar to that of the canonical P element? Journal of Molecular Evolution, , n. 6, p. 715–724, 2007.

DEGNAN, J. H.; ROSENBERG, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution, , n. 6, p. 332–340, 2009. a.

DEGNAN, J. H.; ROSENBERG, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution, , n. 6, p. 332–340, 2009. b.

DINIZ-FILHO, J. A. F.; LOYOLA, R. D.; RAIA, P.; MOOERS, A. O.; BINI, L. M. Darwinian shortfalls in biodiversity conservation. Trends in Ecology and Evolution, , n. 12, p. 689–695, 2013.

DOBZHANSKY, T. Experiments on Sexual Isolation in Drosophila: III. Geographic Strains of Drosophila Sturtevanti. Proceedings of the National Academy of Sciences, , n. 11, p. 335–339, 1944.

DOBZHANSKY, T. G.; PAVAN, C. Studies on Brazilian species of Drosophila. Boletim da Faculdade de Filosofia, Ciências e Letras da Universidade de São Paulo. Biologia Geral., , n. 4, p. 7–72, 1943.

DORONINA, L.; CHURAKOV, G.; SHI, J.; BROSIUS, J.; BAERTSCH, R.; CLAWSON, H.; SCHMITZ, J. Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria, Carnivora). Molecular Biology and Evolution, , n. 12, p. 3194–3204, 2015.

DURAND, E. Y.; PATTERSON, N.; REICH, D.; SLATKIN, M. Testing for ancient admixture between closely related populations. Molecular Biology and Evolution, , n. 8, p. 2239–2252, 2011.

ELLEGREN, H. The different levels of genetic diversity in sex chromosomes and autosomes. Trends in Genetics, , n. 6, p. 278–284, 2009.

FENG, S.; BAI, M.; RIVAS-GONZÁLEZ, I.; LI, C.; LIU, S.; TONG, Y.; YANG, H.; CHEN, G.; XIE, D.; SEARS, K. E.; FRANCO, L. M.; GAITAN-ESPITIA, J. D.; NESPOLO, R. F.; JOHNSON, W. E.; YANG, H.; BRANDIES, P. A.; HOGG, C. J.; BELOV, K.; RENFREE, M. B.; HELGEN, K. M.; BOOMSMA, J. J.; SCHIERUP, M. H.; ZHANG, G. Incomplete lineage sorting and phenotypic evolution in marsupials. Cell, 5, n. 10, p. 1646- 1660.e18, 2022.

FIERS, W.; CONTRERAS, R.; DUERINCK, F.; HAEGEMAN, G.; ISERENTANT, D.; MERREGAERT, J.; MIN JOU, W.; MOLEMANS, F.; RAEYMAEKERS, A.; VAN DEN BERGHE, A.; VOLCKAERT, G.; YSEBAERT, M. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature, 0, n. 5551, p. 500–507, 1976.

GAGNON, E.; HILGENHOF, R.; OREJUELA, A.; MCDONNELL, A.; SABLOK, G.; AUBRIOT, X.; GIACOMIN, L.; GOUVÊA, Y.; BRAGIONIS, T.; STEHMANN, J. R.; BOHS, L.; DODSWORTH, S.; MARTINE, C.; POCZAI, P.; KNAPP, S.; SÄRKINEN, T. Phylogenomic discordance suggests polytomies along the backbone of the large genus Solanum. American Journal of Botany,

GLOR, R. E. Phylogenetic Insights on Adaptive Radiation. Annual Review of Ecology, Evolution, and Systematics, , n. 1, p. 251–270, 2010.

GRANTHAM, R.; GAUTIER, C.; GOUY, M.; JACOBZONE, M.; MERCIER, R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Research, n. 1, p. 213, 1981.

GUILLÍN, E. R.; RAFAEL, V. Cinco especies nuevas del género Drosophila (Diptera, Drosophilidae) en la provincia de Napo, Ecuador. Iheringia - Serie Zoologia, 7, p. 1–12, 2017.

HALLSTRÖM, B. M.; JANKE, A. Mammalian evolution may not be strictly bifurcating. Molecular Biology and Evolution, , n. 12, p. 2804–2816, 2010.

HIME, P. M.; LEMMON, A. R.; LEMMON, E. C. M.; PRENDINI, E.; BROWN, J. M.; THOMSON, R. C.; KRATOVIL, J. D.; NOONAN, B. P.; PYRON, R. A.; PELOSO, P. L. V.; KORTYNA, M. L.; KEOGH, J. S.; DONNELLAN, S. C.; MUELLER, R. L.; RAXWORTHY, C. J.; KUNTE, K.; RON, S. R.; DAS, S.; GAITONDE, N.; GREEN, D. M.; LABISKO, J.; CHE, J.; WEISROCK, D. W. Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life. Systematic Biology, , n. 1, p. 49–66, 2021.

INAGAKI, Y.; ROGER, A. J. Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. Molecular Phylogenetics and Evolution, , n. 2, p. 428–434, 2006.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, , n. 4, p. 772–780, 2013.

KHALLAF, M. A.; CUI, R.; WEISSFLOG, J.; ERDOGMUS, M.; SVATOŠ, A.; DWECK, H. K. M.; VALENZANO, D. R.; HANSSON, B. S.; KNADEN, M. Large-scale characterization of sex pheromone communication systems in Drosophila. Nature Communications, , n. 1, p. 4165, 2021.

KIM, B. Y.; WANG, J. R.; MILLER, D. E.; BARMINA, O.; DELANEY, E.; THOMPSON, A.; COMEAULT, A. A.; PEEDE, D.; D'AGOSTINO, E. R. R.; PELAEZ, J.; AGUILAR, J. M.; HAJI, D.; MATSUNAGA, T.; ARMSTRONG, E. E.; ZYCH, M.; OGAWA, Y.; STAMENKOVIĆ-RADAK, M.; JELIĆ, M.; VESELINOVIĆ, M. S.; TANASKOVIĆ, M.; ERIĆ, P.; GAO, J. J.; KATOH, T. K.; TODA, M. J.; WATABE, H.; WATADA, M.; DAVIS, J. S.; MOYLE, L. C.; MANOLI, G.; BERTOLINI, E.; KOŠŤÁL, V.; HAWLEY, R. S.; TAKAHASHI, A.; JONES, C. D.; PRICE, D. K.; WHITEMAN, N.; KOPP, A.; MATUTE, D. R.; PETROV, D. A. Highly contiguous assemblies of 101 drosophilid genomes. eLife, , p. 1–32, 2021.

KOKATE, P. P.; TECHTMANN, S. M.; WERNER, T. Codon usage bias and dinucleotide preference in 29 Drosophila species. G3 Genes|Genomes|Genetics, , n. 8, p. jkab191, 2021.

KUBATKO, L. S.; CHIFMAN, J. An invariants-based method for efficient identification of hybrid species from

large-scale genomic data. BMC Evolutionary Biology, , n. 1, p. 112, 2019.

KUMAR, S.; STECHER, G.; LI, M.; KNYAZ, C.; TAMURA, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, , n. 6, p. 1547–1549, 2018.

KUZNETSOV, D.; TEGENFELDT, F.; MANNI, M.; SEPPEY, M.; BERKELEY, M.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Research, , n. D1, p. D445–D451, 2023.

LABELLA, A. L.; OPULENTE, D. A.; STEENWYK, J. L.; HITTINGER, C. T.; ROKAS, A. Variation and selection on codon usage bias across an entire subphylum. PLoS Genetics, , n. 7, p. e1008304, 2019.

LACHAISE, D.; CARIOU, M.-L.; DAVID, J. R.; LEMEUNIER, F.; TSACAS, L.; ASHBURNER, M. Historical Biogeography of the *Drosophila melanogaster* Species Subgroup. In: HECHT, M. K.; WALLACE, B.; PRANCE, G. T. (Eds.). **Evolutionary Biology**. Evolutionary BiologyBoston, MA: Springer US, 1988. p. 159–225.

LACHAISE, D.; SILVAIN, J.-F. How two Afrotropical endemics made two cosmopolitan human commensals: the Drosophila melanogaster-D. simulans palaeogeographic riddle. Genetica, 0, n. 1–3, p. 17–39, 2004.

LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, , n. 1 SE-Articles, p. 1–18, 2008.

LI, B.; LOPES, J. S.; FOSTER, P. G.; EMBLEY, T. M.; COX, C. J. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Molecular Biology and Evolution, , n. 7, p. 1697–1709, 2014.

LI, D.; LIU, C.-M.; LUO, R.; SADAKANE, K.; LAM, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, , n. 10, p. 1674–1676, 2015.

LI, F.; RANE, R. V.; LURIA, V.; XIONG, Z.; CHEN, J.; LI, Z.; CATULLO, R. A.; GRIFFIN, P. C.; SCHIFFER, M.; PEARCE, S.; LEE, S. F.; MCELROY, K.; STOCKER, A.; SHIRRIFFS, J.; COCKERELL, F.; COPPIN, C.; SGRÒ, C. M.; KARGER, A.; CAIN, J. W.; WEBER, J. A.; SANTPERE, G.; KIRSCHNER, M. W.; HOFFMANN, A. A.; OAKESHOTT, J. G.; ZHANG, G. Phylogenomic analyses of the genus Drosophila reveals genomic signals of climate adaptation. Molecular Ecology Resources, , n. 4, p. 1559–1581, 2022.

LI, G.; DAVIS, B. W.; EIZIRIK, E.; MURPHY, W. J. Pervasive signals of ancient hybridization in the genomes of living cats (Felidae). Genome Research, 11, 2016.

LI, H. Protein-to-genome alignment with miniprot. Bioinformatics, , n. 1, p. btad014, 2023.

LLOPART, A.; HERRIG, D.; BRUD, E.; STECKLEIN, Z. Sequential adaptive introgression of the mitochondrial genome in Drosophila yakuba and Drosophila santomea. Molecular Ecology, , n. 5, p. 1124–1136, 2014.

MADDISON, W. Reconstructing character evolution on polytomous cladograms. Cladistics, n. 4, p. 365–377, 1989.

MADDISON, W. P. Gene trees in species trees. Systematic Biology, , n. 3, p. 523–536, 1997.

MADI-RAVAZZI, L.; ROMAN, B. E.; CESAR, K.; ALEVI, C.; PREDIGER, C.; YASSIN, A.; WOLFGANG, J. Integrative taxonomy and a new species description in the sturtevanti subgroup of the Drosophila saltans group (Diptera: Drosophilidae). 80, n. 2, p. 269–292, 2021.

MAGALHÃES, L. E. De. Notes on the taxonomy, morphology, and distribution of the saltans group of Drosophila, with description of four new species. The University of Texas Publication, 5–154, 1962.

MAGALHÃES, L. E. De; BJÖRNBERG, A. J. S. Estudo da genitália masculina de Drosophila do grupo saltans

(Díptera). Revista Brasileira de Biologia, , n. 4, p. 435–450, 1957.

MAI, D.; NALLEY, M. J.; BACHTROG, D.; WRIGHT, S. Patterns of genomic differentiation in the Drosophila nasuta species complex. Molecular Biology and Evolution, , n. 1, p. 208–220, 2020.

MALINSKY, M.; MATSCHINER, M.; SVARDAL, H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. Molecular Ecology Resources, , n. 2, p. 584–595, 2021.

MALLET, J.; BESANSKY, N.; HAHN, M. W. How reticulated are species? BioEssays, , n. 2, p. 140–149, 2016.

MANSOURIAN, S.; ENJIN, A.; JIRLE, E. V.; RAMESH, V.; REHERMANN, G.; BECHER, P. G.; POOL, J. E.; STENSMYR, M. C. Wild African Drosophila melanogaster Are Seasonal Specialists on Marula Fruit. Current biology: CB, , n. 24, p. 3960- 3968.e3, 2018.

MARTIN, S. H.; DAVEY, J. W.; JIGGINS, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Molecular Biology and Evolution, , n. 1, p. 244–257, 2015.

MATUTE, D. R.; COMEAULT, A. A.; EARLEY, E.; SERRATO-CAPUCHINA, A.; PEEDE, D.; MONROY-EKLUND, A.; HUANG, W.; JONES, C. D.; MACKAY, T. F. C.; COYNE, J. A. Rapid and Predictable Evolution of Admixed Populations Between Two Drosophila Species Pairs. Genetics, 4, n. 1, p. 211–230, 2020.

MAYR, E.; DOBZHANSKY, Th. Experiments on Sexual Isolation in Drosophila. Proceedings of the National Academy of Sciences, , n. 2, p. 75–82, 1945.

MEADE, A.; PAGEL, M. Ancestral State Reconstruction Using BayesTraits. Methods in Molecular Biology (Clifton, N.J.), 69, p. 255–266, 2022.

MENG, G.; LI, Y.; YANG, C.; LIU, S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Research, , n. 11, p. e63, 2019.

MILLER, J. B.; HIPPEN, A. A.; BELYEU, J. R.; WHITING, M. F.; RIDGE, P. G. Missing something? Codon aversion as a new character system in phylogenetics. Cladistics, , n. 5, p. 545–556, 2017.

MORAN, B. M.; PAYNE, C.; LANGDON, Q.; POWELL, D. L.; BRANDVAIN, Y.; SCHUMER, M. The genomic consequences of hybridization. eLife, , p. e69016, 2021.

MOREYRA, N. N.; ALMEIDA, F. C.; ALLAN, C.; FRANKEL, N.; MATZKIN, L. M.; HASSON, E. Phylogenomics provides insights into the evolution of cactophily and host plant shifts in Drosophila. Molecular Phylogenetics and Evolution, 8, p. 107653, 2023.

MORGAN, C. C.; FOSTER, P. G.; WEBB, A. E.; PISANI, D.; MCINERNEY, J. O.; O'CONNELL, M. J. Heterogeneous models place the root of the placental mammal phylogeny. Molecular Biology and Evolution, , n. 9, p. 2145–2156, 2013.

NASCIMENTO, A. P.; BICUDO, H. E. M. C. Esterase patterns and phylogenetic relationships of Drosophila species in the saltans subgroup (saltans group). Genetica, 4, n. 1, p. 41–51, 2002.

NGUYEN, L. T.; SCHMIDT, H. A.; VON HAESELER, A.; MINH, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution, , n. 1, p. 268–274, 2015.

O'DEA, A.; LESSIOS, H. A.; COATES, A. G.; EYTAN, R. I.; RESTREPO-MORENO, S. A.; CIONE, A. L.; COLLINS, L. S.; DE QUEIROZ, A.; FARRIS, D. W.; NORRIS, R. D.; STALLARD, R. F.; WOODBURNE, M. O.; AGUILERA, O.; AUBRY, M.-P.; BERGGREN, W. A.; BUDD, A. F.; COZZUOL, M. A.; COPPARD, S. E.; DUQUE-CARO, H.; FINNEGAN, S.; GASPARINI, G. M.; GROSSMAN, E. L.; JOHNSON, K. G.; KEIGWIN, L. D.; KNOWLTON, N.; LEIGH, E. G.; LEONARD-PINGEL, J. S.; MARKO, P. B.; PYENSON, N. D.; RACHELLO-DOLMEN, P. G.; SOIBELZON, E.; SOIBELZON, L.; TODD, J. A.; VERMEIJ, G. J.; JACKSON, J. B. C. Formation of the Isthmus of Panama. Science Advances, n. 8, p. e1600883, 2016.

O'GRADY, P. M.; CLARK, J. B.; KIDWELL, M. G. Phylogeny of the Drosophila saltans species group based on combined analysis of nuclear and mitochondrial DNA sequences. Molecular Biology and Evolution, , n. 6, p. 656–664, 1998.

O'GRADY, P. M.; DESALLE, R. Phylogeny of the genus Drosophila. Genetics, 9, n. 1, p. 1–25, 2018.

ØRSTED, I. V.; ØRSTED, M. Species distribution models of the Spotted Wing Drosophila (Drosophila suzukii, Diptera: Drosophilidae) in its native and invasive range reveal an ecological niche shift. Journal of Applied Ecology, , n. 2, p. 423–435, 2019.

OWEN, C. L.; MILLER, G. L. Phylogenomics of the Aphididae: Deep relationships between subfamilies clouded by gene tree discordance, introgression and the gene tree anomaly zone. Systematic Entomology, , n. 3, p. 470–486, 2022. a.

OWEN, C. L.; MILLER, G. L. Phylogenomics of the Aphididae: Deep relationships between subfamilies clouded by gene tree discordance, introgression and the gene tree anomaly zone. Systematic Entomology, , n. 3, p. 470–486, 2022. b.

PARMLEY, J. L.; HURST, L. D. Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals. Molecular Biology and Evolution, , n. 8, p. 1600–1603, 2007.

PATTERSON, N.; MOORJANI, P.; LUO, Y.; MALLICK, S.; ROHLAND, N.; ZHAN, Y.; GENSCHORECK, T.; WEBSTER, T.; REICH, D. Ancient Admixture in Human History. Genetics, 2, n. 3, p. 1065–1093, 2012.

PEASE, J. B.; HAHN, M. W. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Systematic Biology, , n. 4, p. 651–662, 2015.

PÉLANDAKIS, M.; SOLIGNAC, M. Molecular phylogeny of Drosophila based on ribosomal RNA sequences. Journal of Molecular Evolution, , n. 5, p. 525–543, 1993.

PHILIPPE, H.; BRINKMANN, H.; LAVROV, D. V.; LITTLEWOOD, D. T. J.; MANUEL, M.; WÖRHEIDE, G.; BAURAIN, D. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biology, n. 3, 2011.

PHILIPPE, H.; DERELLE, R.; LOPEZ, P.; PICK, K.; BORCHIELLINI, C.; BOURY-ESNAULT, N.; VACELET, J.; RENARD, E.; HOULISTON, E.; QUÉINNEC, E.; DA SILVA, C.; WINCKER, P.; LE GUYADER, H.; LEYS, S.; JACKSON, D. J.; SCHREIBER, F.; ERPENBECK, D.; MORGENSTERN, B.; WÖRHEIDE, G.; MANUEL, M. Phylogenomics revives traditional views on deep animal relationships. Current Biology, , n. 8, p. 706–712, 2009.

PICK, K. S.; PHILIPPE, H.; SCHREIBER, F.; ERPENBECK, D.; JACKSON, D. J.; WREDE, P.; WIENS, M.; ALIÉ, A.; MORGENSTERN, B.; MANUEL, M.; WÖRHEIDE, G. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Molecular Biology and Evolution, , n. 9, p. 1983–1987, 2010.

POWELL, J. R.; SEZZI, E.; MORIYAMA, E. N.; GLEASON, J. M.; CACCONE, A. Analysis of a Shift in Codon Usage in Drosophila. Journal of Molecular Evolution, , n. SUPPL. 1, p. 214–225, 2003.

PRESNYAK, V.; ALHUSAINI, N.; CHEN, Y.-H.; MARTIN, S.; MORRIS, N.; KLINE, N.; OLSON, S.; WEINBERG, D.; BAKER, K. E.; GRAVELEY, B. R.; COLLER, J. Codon Optimality Is a Major Determinant of mRNA Stability. Cell, 0, n. 6, p. 1111–1124, 2015.

QVARNSTRÖM, A.; BAILEY, R. I. Speciation through evolution of sex-linked genes. Heredity, 2, n. 1, p. 4–15, 2009.

RANALLO-BENAVIDEZ, T. R.; JARON, K. S.; SCHATZ, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications, , p. 1432, 2020.

REILLY, P. F.; TJAHJADI, A.; MILLER, S. L.; AKEY, J. M.; TUCCI, S. The contribution of Neanderthal introgression to modern human traits. Current Biology, , n. 18, p. R970–R983, 2022.

RETCHLESS, A. C.; LAWRENCE, J. G. Ecological Adaptation in Bacteria: Speciation Driven by Codon Selection. Molecular Biology and Evolution, , n. 12, p. 3669–3683, 2012.

REVELL, L. J. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, n. 2, p. 217–223, 2012.

RICK, J. A.; BROCK, C. D.; LEWANSKI, A. L.; GOLCHER-BENAVIDES, J.; WAGNER, C. E. Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. Systematic Biology, ad065, 2023.

RODRÍGUEZ-TRELLES, F.; TARRÍO, R.; AYALA, F. J. Molecular evolution and phylogeny of the Drosophila saltans species group Inferred from the Xdh Gene. Molecular Phylogenetics and Evolution, , n. 1, p. 110–121, 1999. a.

RODRÍGUEZ-TRELLES, F.; TARRÍO, R.; AYALA, F. J. Switch in codon bias and increased rates of amino acid substitution in the Drosophila saltans species group. Genetics, 3, n. 1, p. 339–350, 1999. b.

ROMAN, B. E.; MADI-RAVAZZI, L. Male terminalia morphology of sixteen species of the Drosophila saltans group Sturtevant (Diptera, Drosophilidae). Zootaxa, 61, n. 3, p. 523–544, 2021.

ROMAN, B. E.; SANTANA, D. J.; PREDIGER, C.; MADI-RAVAZZI, L. Phylogeny of Drosophila saltans group (Diptera: Drosophilidae) based on morphological and molecular evidence. Plos One, , n. 4, p. e0266710, 2022.

ROMIGUIER, J.; RANWEZ, V.; DELSUC, F.; GALTIER, N.; DOUZERY, E. J. P. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular Biology and Evolution, , n. 9, p. 2134–2144, 2013.

ROTA-STABELLI, O.; LARTILLOT, N.; PHILIPPE, H.; PISANI, D. Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. Systematic Biology, , n. 1, p. 121–133, 2013.

SANKARARAMAN, S.; MALLICK, S.; PATTERSON, N.; REICH, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current Biology, , n. 9, p. 1241–1247, 2016.

SAYYARI, E.; MIRARAB, S. Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. Genes, n. 3, p. 132, 2018.

SCHAEFFER, S. W. Muller “Elements” in Drosophila : How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. Genetics, 0, n. 1, p. 3–13, 2018.

SCHREMPF, D.; SZÖLLŐSI, G. The sources of phylogenetic conflicts. In: SCORNAVACCA, C.; DELSUC, F.; GALTIER, N. (Eds.). **Phylogenetics in the Genomic Era**. No commercial publisher | Authors open access book, 2020. p. chapter. 3.1, p. 3.1:1-3.1:23.

SCORNAVACCA, C.; DELSUC, F.; GALTIER, N. **Phylogenetics in the Genomic Era.** No commercial publisher, 2020.

SEIXAS, F. A.; BOURSOT, P.; MELO-FERREIRA, J. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. Genome Biology, , n. 1, p. 91, 2018.

SHARP, P. M.; AVEROF, M.; LLOYD, A. T.; MATASSI, G.; PEDEN, J. F.; GERHART, J.; HUNT, R. T.; KIRSCHNER, M. W.; WOLPERT, L. DNA sequence evolution: the sounds of silence. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 9, n. 1329, p. 241–247, 1997.

SHARP, P. M.; STENICO, M.; PEDEN, J. F.; LLOYD, A. T. Codon usage: mutational bias, translational selection, or both? Biochemical Society Transactions, , n. 4, p. 835–841, 1993.

SHARP, P. M.; TUOHY, T. M. F.; MOSURSKI, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Research, , n. 13, p. 5125–5143, 1986.

SIMÃO, F. A.; WATERHOUSE, R. M.; IOANNIDIS, P.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, , n. 19, p. 3210–3212, 2015.

SIMION, P.; PHILIPPE, H.; BAURAIN, D.; JAGER, M.; RICHTER, D. J.; DI FRANCO, A.; ROURE, B.; SATOH, N.; QUÉINNEC, É.; ERESKOVSKY, A.; LAPÉBIE, P.; CORRE, E.; DELSUC, F.; KING, N.; WÖRHEIDE, G.; MANUEL, M. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Current biology: CB, , n. 7, p. 958–967, 2017.

SKOV, L.; COLL MACIÀ, M.; LUCOTTE, E. A.; CAVASSIM, M. I. A.; CASTELLANO, D.; SCHIERUP, M. H.; MUNCH, K. Extraordinary selection on the human X chromosome associated with archaic admixture. Cell Genomics, n. 3, p. 100274, 2023.

SOLIGNAC, M.; MONNEROT, M.; MOUNOLOU, J. C. Mitochondrial DNA evolution in the melanogaster species subgroup of Drosophila. Journal of Molecular Evolution, , n. 1, p. 31–40, 1986.

SOUZA, T. A. J.; NOLL, F. B.; BICUDO, H. E. M. C.; MADI-RAVAZZI, L. Scanning electron microscopy of male terminalia and its application to species recognition and phylogenetic reconstruction in the Drosophila saltans group. PLoS ONE, n. 6, 2014.

SPASSKY, B. Morphological differences between sibling species of Drosophila. In: **Genetics of Drosophila**. v. 5721.

SPRENGELMEYER, Q. D.; MANSOURIAN, S.; LANGE, J. D.; MATUTE, D. R.; COOPER, B. S.; JIRLE, E. V.; STENSMYR, M. C.; POOL, J. E. Recurrent Collection of Drosophila melanogaster from Wild African Environments and Genomic Insights into Species History. Molecular Biology and Evolution, , n. 3, p. 627–638, 2020.

STURTEVANT, A. H. The classification of the genus Drosophila, with descriptions of nine new species. The University of Texas Publication, 13, p. 7–51, 1942.

STURTEVANT, A. H.; NOVITSKI, E. The homologies of the chromosome elements in the genus Drosophila. Genetics, , n. 5, p. 517–541, 1941.

SUH, A. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. Zoologica Scripta, , n. S1, p. 50–62, 2016.

SUN, Y.; TAMARIT, D.; ANDERSSON, S. G. E. Switches in Genomic GC Content Drive Shifts of Optimal Codons under Sustained Selection on Synonymous Sites. Genome Biology and Evolution, n. 10, p. 2560–2579, 2017.

SUVOROV, A.; KIM, B. Y.; WANG, J.; ARMSTRONG, E. E.; PEEDE, D.; D'AGOSTINO, E. R. R.; PRICE, D. K.; WADELL, P.; LANG, M.; COURTIER-ORGOGOZO, V.; DAVID, J. R.; PETROV, D.; MATUTE, D. R.; SCHRIDER, D. R.; COMEAULT, A. A. Widespread introgression across a phylogeny of 155 Drosophila genomes. Current Biology, , p. 111–123, 2022.

TARRÍO, R.; RODRÍGUEZ-TRELLES, F.; AYALA, F. J. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The Drosophila saltans and willistoni groups, a case study. Molecular Phylogenetics and Evolution, , n. 3, p. 344–349, 2000.

THROCKMORTON, L. H. The problem of phylogeny in the genus Drosophila. University of Texas Publications, 05, p. 207–343, 1962.

THROCKMORTON, L. H. **The phylogeny, ecology, and geography of Drosophila**, 1975.

THROCKMORTON, L. H.; MAGALHÃES, L. E. Changes with evolution of pteridine accumulations in species of the saltans group of the genus Drosophila. University of Texas Publications, 05, p. 489–505, 1962.

TIDON, R.; DE ALMEIDA, J. M. Family drosophilidae. Zootaxa, 22, n. 1, p. 719–751, 2016.

TOEWS, D. P. L.; BRELSFORD, A. The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, , n. 16, p. 3907–3930, 2012.

TOWNSEND, J. P.; SU, Z.; TEKLE, Y. I. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. Systematic Biology, , n. 5, p. 835–849, 2012.

TSACAS, L.; DAVID, J. Systenatics and biogeography of the Drosophila kikkawai-complex with descrition of new species (Diptera: Drosophilidae). Annales de la Société entomologique de France, n. 4, p. 675–693, 1977.

TWYFORD, A. D.; ENNOS, R. A. Next-generation hybridization and introgression. Heredity, 8, n. 3, p. 179–189, 2012.

VALIENTE-MULLOR, C.; BEAMUD, B.; ANSARI, I.; FRANCÉS-CUESTA, C.; GARCÍA-GONZÁLEZ, N.; MEJÍA, L.; RUIZ-HUESO, P.; GONZÁLEZ-CANDELAS, F. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLOS Computational Biology, , n. 1, p. e1008678, 2021.

VICARIO, S.; MORIYAMA, E. N.; POWELL, J. R. Codon usage in twelve species of Drosophila. BMC Evolutionary Biology, n. 1, 2007. . Acesso em: 17 abr. 2021.

WALLAU, G. L.; CAPY, P.; LORETO, E.; LE ROUZIC, A.; HUA-VAN, A. VHICA, a New Method to Discriminate between Vertical and Horizontal Transposon Transfer: Application to the Mariner Family within Drosophila. Molecular Biology and Evolution, , n. 4, p. 1094–1109, 2016.

WALSH, H. E.; KIDD, M. G.; MOUM, T.; FRIESEN, V. L. Polytomies and the power of phylogenetic inference. Evolution, , n. 3, p. 932–937, 1999.

WAN, X. F.; XU, D.; KLEINHOFS, A.; ZHOU, J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evolutionary Biology, p. 1–11, 2004.

WARNES, G.; BOLKER, B.; BONEBAKKER, L.; GENTLEMAN, R.; HUBER, W.; LIAW, A.; LUMLEY, T.; MÄCHLER, M.; MAGNUSSON, A.; MÖLLER, S. **gplots: Various R programming tools for plotting data**. v. 2journalAbbreviation: R package version.

WATERHOUSE, R. M.; SEPPEY, M.; SIMAO, F. A.; MANNI, M.; IOANNIDIS, P.; KLIOUTCHNIKOV, G.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution, , n. 3, p. 543–548, 2018.

WHELAN, N. V.; KOCOT, K. M.; MOROZ, L. L.; HALANYCH, K. M. Error, signal, and the placement of Ctenophora sister to all other animals. Proceedings of the National Academy of Sciences, 2, n. 18, p. 5773–5778, 2015.

WICKETT, N. J.; MIRARAB, S.; NGUYEN, N.; WARNOW, T.; CARPENTER, E.; MATASCI, N.; AYYAMPALAYAM, S.; BARKER, M. S.; BURLEIGH, J. G.; GITZENDANNER, M. A.; RUHFEL, B. R.; WAFULA, E.; DER, J. P.; GRAHAM, S. W.; MATHEWS, S.; MELKONIAN, M.; SOLTIS, D. E.; SOLTIS, P. S.; MILES, N. W.; ROTHFELS, C. J.; POKORNY, L.; SHAW, A. J.; DEGIRONIMO, L.; STEVENSON, D. W.; SUREK, B.; VILLARREAL, J. C.; ROURE, B.; PHILIPPE, H.; DEPAMPHILIS, C. W.; CHEN, T.; DEYHOLOS, M. K.; BAUCOM, R. S.; KUTCHAN, T. M.; AUGUSTIN, M. M.; WANG, J.; ZHANG, Y.; TIAN, Z.; YAN, Z.; WU, X.; SUN, X.; WONG, G. K.-S.; LEEBENS-MACK, J. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences, 1, n. 45, p. E4859–E4868, 2014.

WRIGHT, F. The “effective number of codons” used in a gene. Gene, , n. 1, p. 23–29, 1990.

YAMAGUCHI, M.; YOSHIDA, H. *Drosophila* as a Model Organism. In: YAMAGUCHI, M. (Ed.). ***Drosophila* models for human diseases**. Singapore: Springer, 2018. v. 1076p. 1–11.

YASSIN, A. Phylogenetic relationships among species subgroups in the Drosophila saltans group (Diptera: Drosophilidae): Can morphology solve a molecular conflict. Zoological Research, , n. 3, p. 225–232, 2009.

YUSUF, L. H.; TYUKMAEVA, V.; HOIKKALA, A.; RITCHIE, M. G. Divergence and introgression among the virilis group of Drosophila. Evolution Letters, n. 6, p. 537–551, 2022.

ZHANG, C.; RABIEE, M.; SAYYARI, E.; MIRARAB, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics, , n. 6, p. 153, 2018.

ZHOU, T.; WEEMS, M.; WILKE, C. O. Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. Molecular Biology and Evolution, , n. 7, p. 1571–1580, 2009.

ZHOU, Z.; DANG, Y.; ZHOU, M.; YUAN, H.; LIU, Y. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. eLife, p. e33569, 2018.

ZIMIN, A. V.; MARÇAIS, G.; PUIU, D.; ROBERTS, M.; SALZBERG, S. L.; YORKE, J. A. The MaSuRCA genome assembler. Bioinformatics, , n. 21, p. 2669–2677, 2013.

APPENDIX A: Supplementary material Chapter 3

Supplentmentary Document 1. Futher discussion. Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

Supplementary Data S1. 2,159 gene trees generated maximum likelihood estimation. Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

Supplementary Table S1. Summary of previous competing phylogenetic hypotheses in the *saltans* group. CO = *cordata* subgroup, EL = *elliptica* subgroup, ST = *stutevanti* subgroup, PA = *parasaltans* subgroup, SA = *saltans* subgroup, aus= *D. austrosaltans*, nig = *D. nigrosaltans*, sal = *D. saltans*, pro = *D. prosaltans*, lus = *D. lusaltans*, sep = *D. septentriosaltans*, pse = *D. pseudosaltans*, stu= *D. sturtevanti*, leh = *D. lehrmanae*, mil = *D. milleri*, dac = *D. dacunhai*, nsa = *D. neosaltans*, nel = *D. neoelliptica*, ema= *D. emarginada*, OS = overall similarities, ai = measure of isolation for each interspecific cross, MP = maximum parsimonia, ML = maximum likelihood, BI =Bayesian inference.

Markers	Topology	N. of species	Methods	Authors
	D. saltans group			
Morphology 1	(CO,(EL,(ST,(PA,SA))));	14	OS	Throckmorton & Magalhães 1962
Spermatheca	(CO,EL,ST,(SA,PA));	5	OS	Throckmorton 1962
Morphology 2	(ST,SA,(PA,(CO,EL)));	8	MP	Magalhaes 1962 in O'Grady 1998
28S	((ST,EL),(CO,SA));	4	MP	Pélandakis & Solignac 1993
Adh	*D. saltans* subgroup not monophyletic	8	MP	O'Grady et al. 1998
COI	(CO,EL,(ST,PA,SA));	8	MP	O'Grady et al. 1998
COI + COII + ITS1 + AdH + morphology2	(CO, (EL, (ST, PA, SA)));	8	MP	O'Grady et al. 1998
COII	(PA,(CO,SA,EL,ST));	8	MP	O'Grady et al. 1998
ITS1	(PA,(CO,EL), ST, SA);	8	MP	O'Grady et al. 1998
Xdh	(PA,(ST,(EL,(CO,SA))));	6	ML	Rodríguez-Trelles et al., 1998
Xdh + Adh + ITS1	(PA,(ST,((EL,CO),SA)));	6	NJ	Rodríguez-Trelles et al., 1998
Xdh + Adh + ITS1 + COI	(PA,(ST,((EL,CO),SA)));	6	NJ	Rodríguez-Trelles et al., 1998
Xdh + Adh + ITS1 + COI + COII	(PA,(ST,((EL,CO),SA)));	6	NJ	Rodríguez-Trelles et al., 1998
Xdh + Adh + ITS1 + COII	(PA,(ST,((EL,CO),SA)));	6	NJ	Rodríguez-Trelles et al., 1998
P element	((CO,EL),(PA,(SA,ST))); or (CO,EL,(PA,(SA,ST)));	10	PA	Castro & Cararetto 2004
Adh	(CO,(SA,(PA,(EL,ST))));	9	NJ	Setta et al. 2007

Morphology 3	(ST, ((CO,EL),(PA,SA)));	9	MP	Yassin, 2009
Morphology 4	(CO,((EL, ST),(PA,SA)));	10	MP	Souza, 2014
CO I+ COII + morphology 5	(PA,(SA,(ST,(EL,CO))));	16	BI	Roman et al. 2022
COI + COII	(PA,(SA,(ST,(EL,CO))));	16	BI	Roman et al. 2022
Morphology 5	((PA,SA),ST,EL,CO));	16	MP	Roman et al. 2022
D. saltans subgroup				
Reproductive isolation	((aus, nig),(sep,(lus,sal,pro)),pse);	7	Ai	Bicudo 1973
chromosome inversion	((((lus,sep,sal),(aus,nig)),pse),pro);	7	OS	Bicudo 1973
Adh + ITS1 + COI + COII + morphology2	(sal,aus,lus,pro);	4	MP	O'Grady, 1998
Esterases	(pro,(aus,(sep,sal)));	4	NJ	Nascimento & Bicudo 2002
Adh	(pro, (sal, (lus, aus)));	4	NJ	Setta et al. 2007
morphology 3	((aus,lus),pro,sal);	4	MP	Yassin 2009
Morphology 4	(sal,(pro,lus,aus));	4	MP	Souza et al. 2014
COI + COII	((lus,pro,pse,sep),(nig,(aua,sal))	7	BI	Roman et al. 2022
CO I + COII + morphology 5	(lus,pse,(pro,sep),(nig,(aus,sal)));	7	BI	Roman et al. 2022
morphology 5	(sal,(sep,aus),(pse,nig),pro,lus)	7	MP	Roman et al. 2022
D. sturtevanti subgroup				
reproductive isolation	((stu,mil),mag);	3	Ai	Bicudo 1979
Morphology 4	(stu, (dac,mil));	3	MP	Souza, 2014
COI + COII + ND4 + ND2	((stu,leh),(mil,dac));	4	BI	Madi-Ravazzi et al. 2021
morphology5	((stu,leh),(mil,dac));	4	MP	Roman et al. 2022
COI + COII	((stu,leh),(mil,dac));	4	BI	Roman et al. 2022
CO I+ COII + morphology 5	((stu,leh),(mil,dac));	4	BI	Roman et al. 2022

D. elliptica subgroup				
COI + COII	(nsa,(nel,ema));	3	BI	Roman et al. 2022
CO I+ COII + morphology 5	(nsa,(nel,ema));	3	BI	Roman et al. 2022
morphology 5	(nsa,(nel,ema));	3	MP	Roman et al. 2022

Supplementary Table S2. Assembly quality, completeness of the genome of the saltans group. The total number of single copy genes used as baits is 3,285.
Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

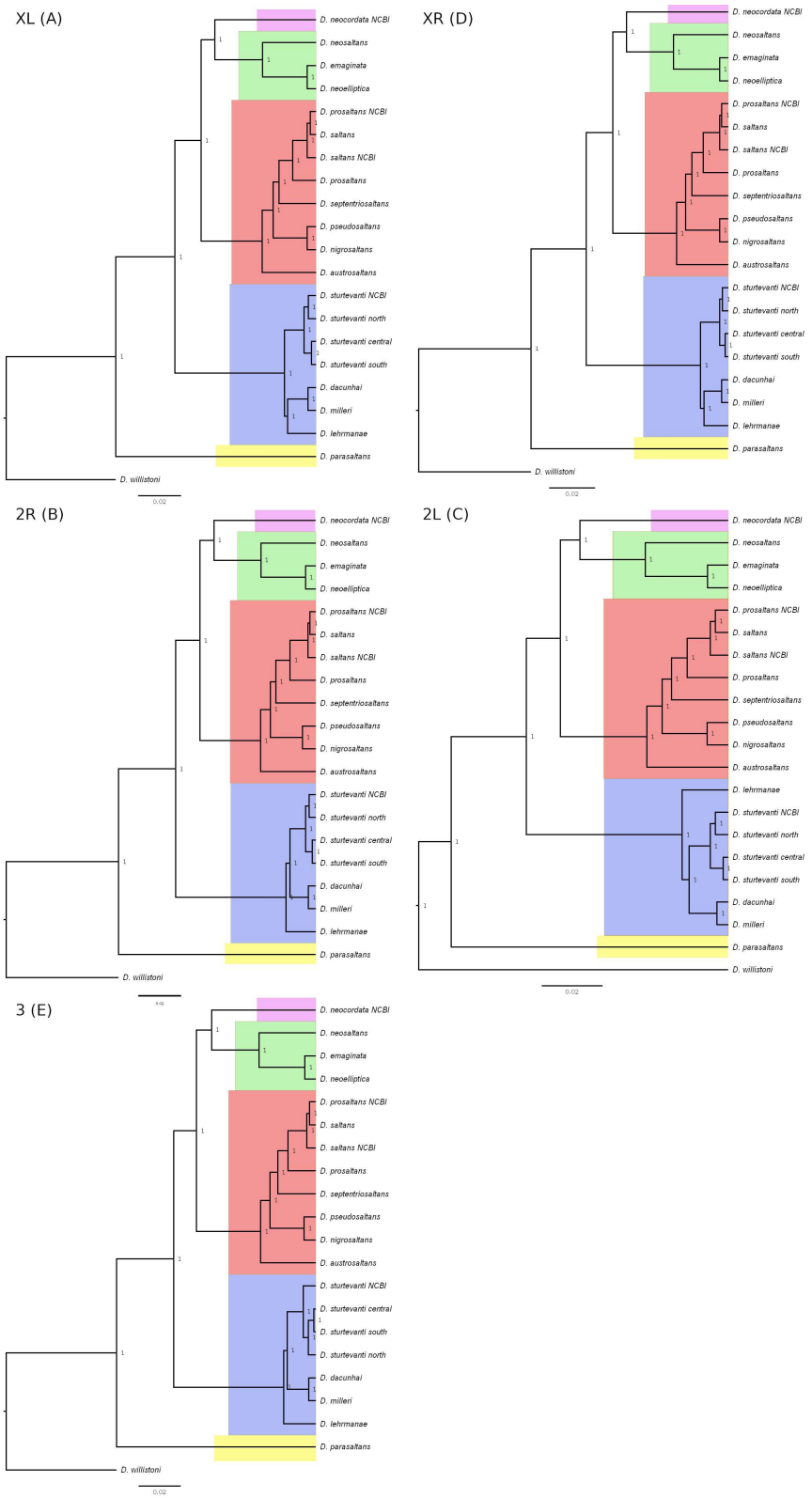
Supplementary Table S3. 2A2B Results for every quartets
Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

SupplementaryTable S4. Node ages, ancestral area for the *saltans* group
Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

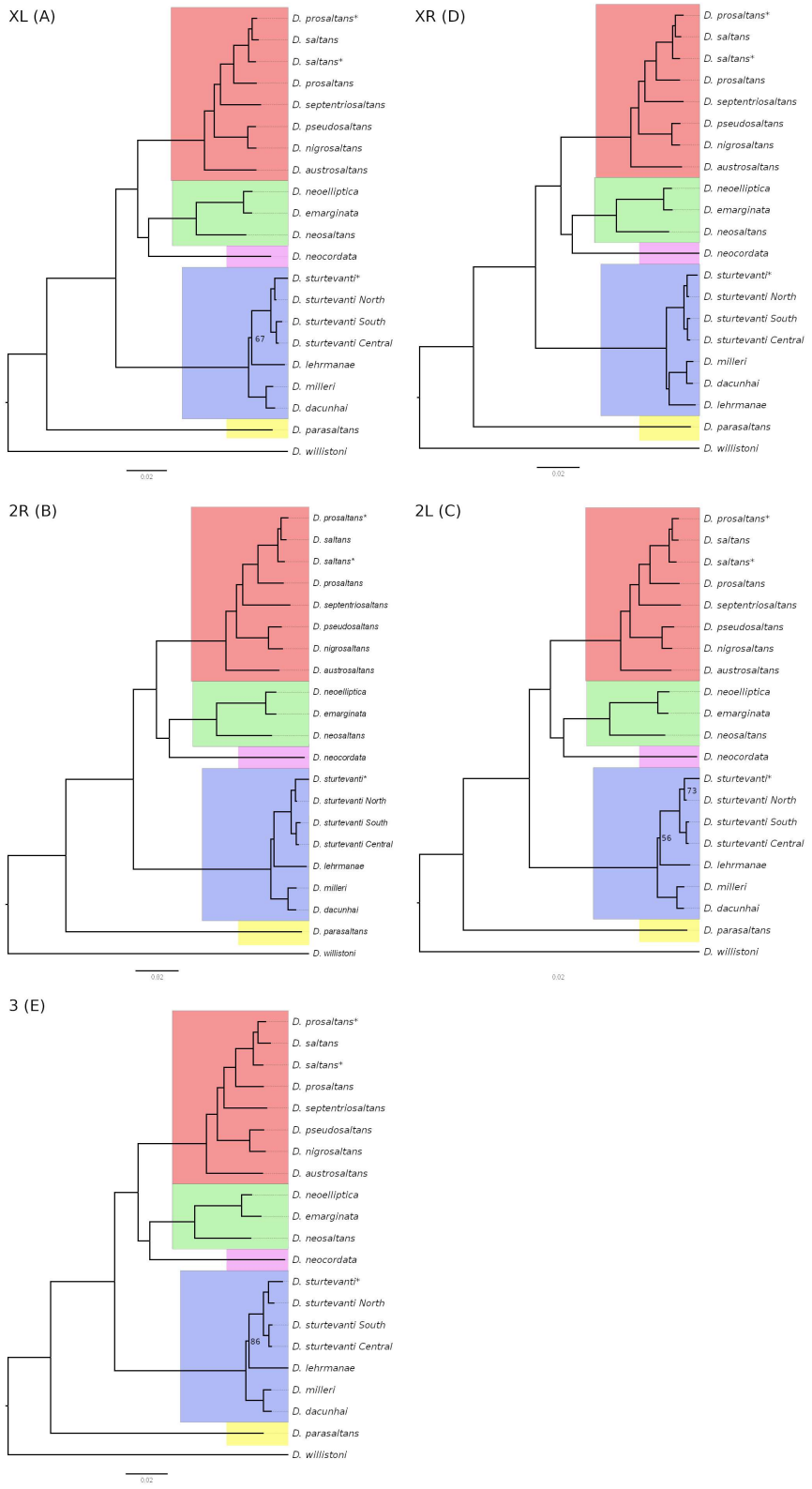
Supplementary Table S5. T2/T1 and H2/H1ratio and reticulation estimated for the *saltans* group.
Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

Supplementary Table S6. Location and number of individuals used in the illumina PoolSeq.
Available link: https://drive.google.com/drive/folders/15evhOfc3jxKn_rZoxOb1PwQFiH84iPTP?usp=drive_link

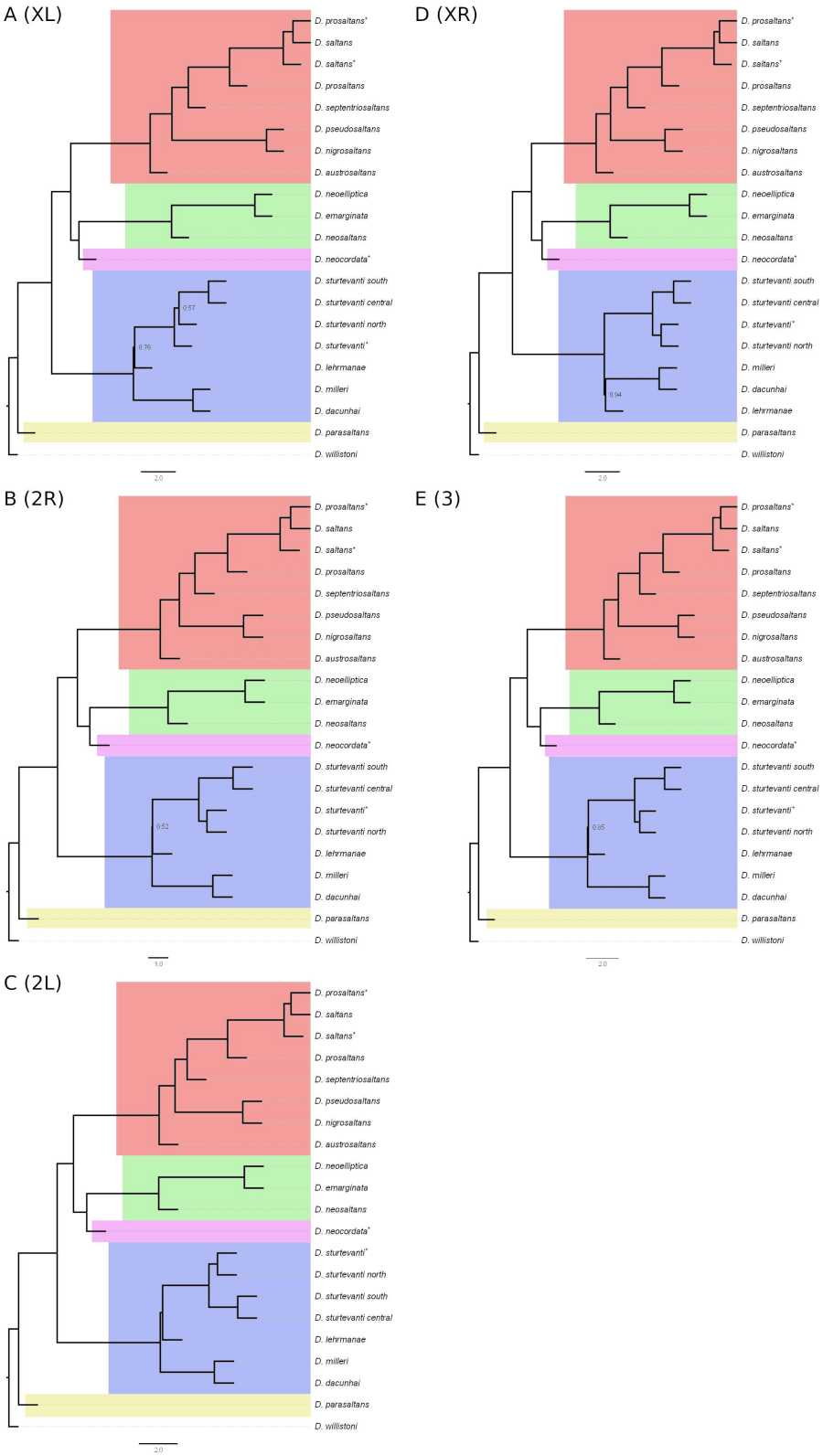
Supplementary Figure S1. Bayesian Inference trees generated with 5 independent datasets, chromosome arms and respecteve Muller elements are indicated in each tree. Branch Posterior probabilities are shown for each node. The *parasaltans, sturtevanti, saltans, elliptica* and *cordata* subgroups are highlighted in yellow, blue, red, green and pink, respectively.



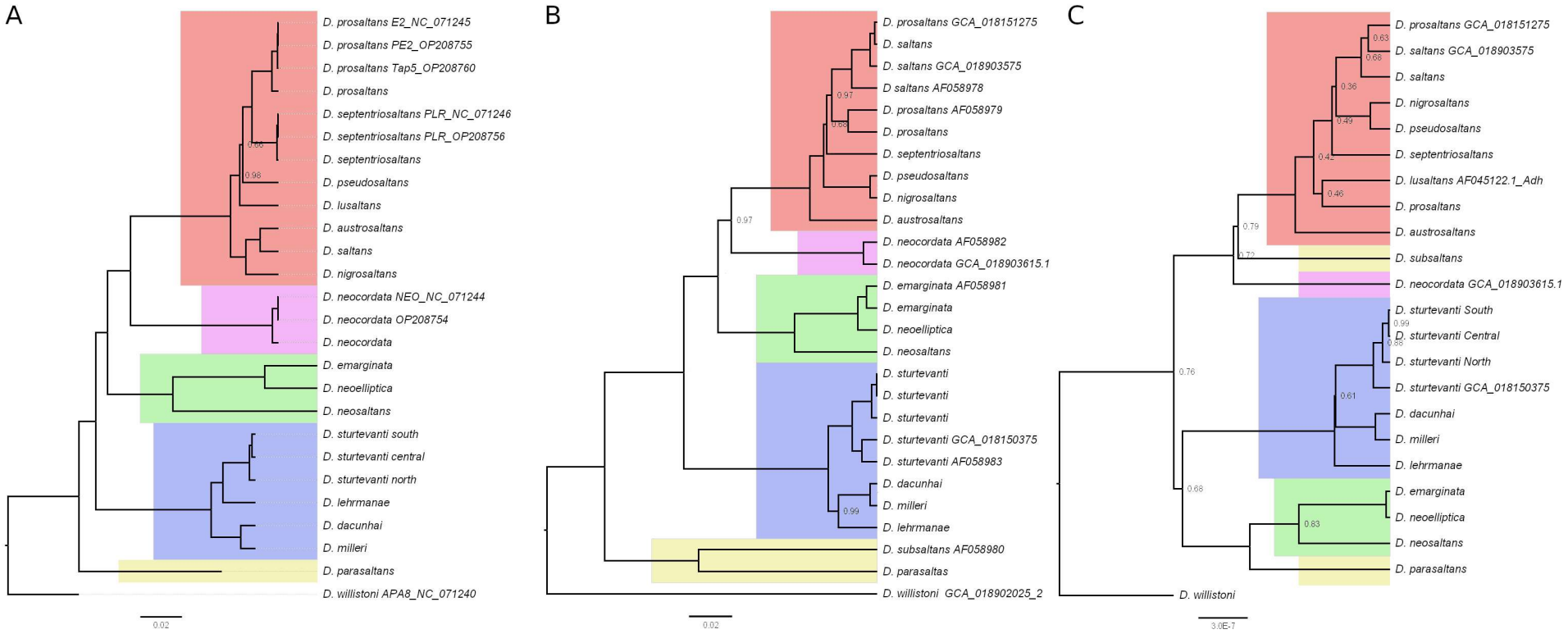
Supplementary Figure S2. Maximum likelihood trees generated with 5 independent datasets, each comprise the concatenate genes predicted to the Muller elements A-F. UltraFast Bootrap values are shown for each node. The *parasaltans, sturtevanti, saltans, elliptica* and *cordata* subgroups are highlighted in yellow, blue, red, green and pink, respectively.



Supplementary Figure S3. Species Tree generated under the multi-species coalecent model impremented in ASTRAL-III, from the 2,156 genes tree available in Supplementary Data S1 and evaluated as 5 different data sets, acording to genes predicted to the Muller elements A-F. Branch support are shown for each node. The *parasaltans, sturtevanti, saltans, elliptica* and *cordata* subgroups are highlighted in yellow, blue, red, green and pink, respectively.

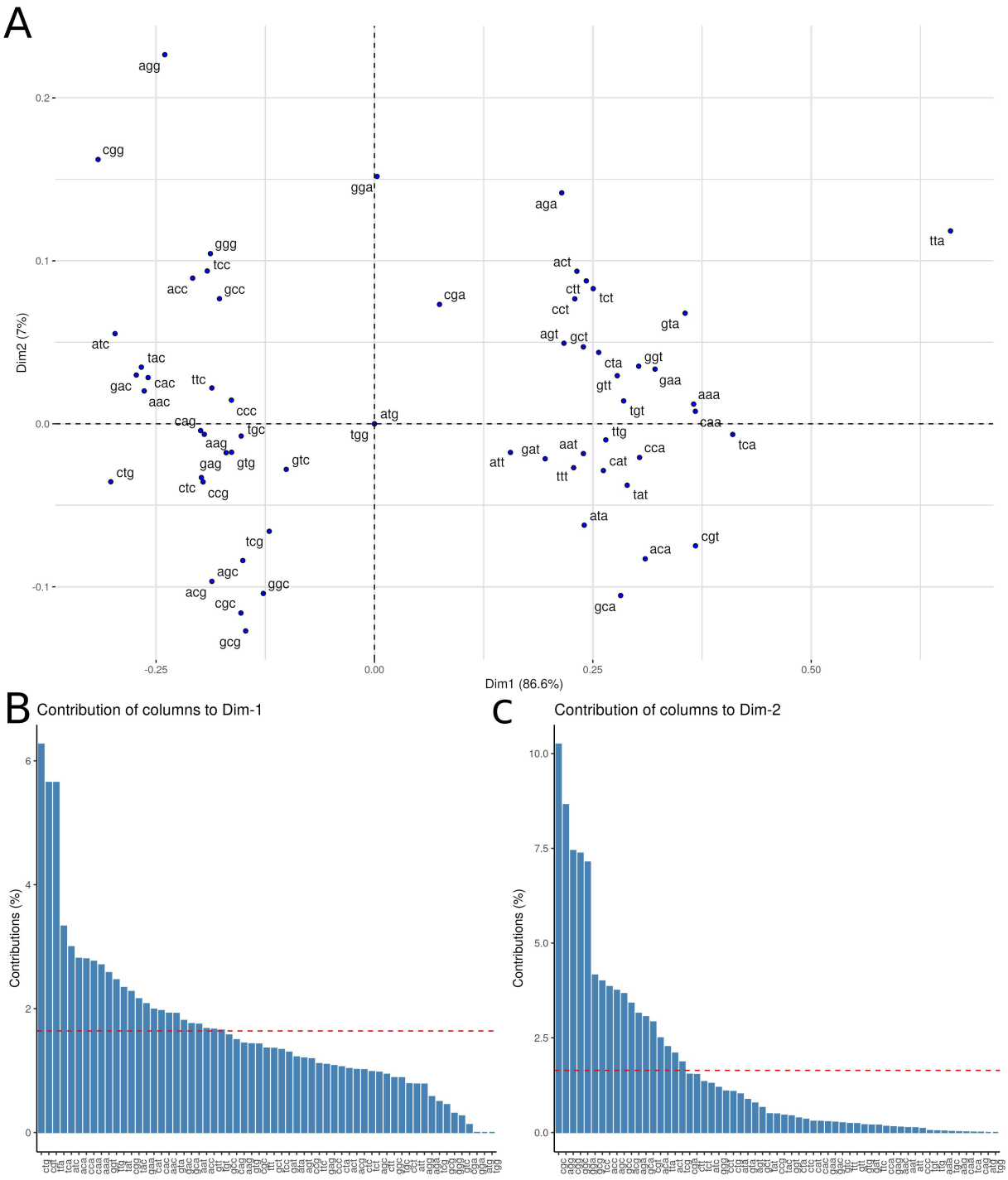


Supplementary Figure S4. Phylogenetic tree with inclusion of *D. lusaltans* and *D.* subsaltans. Mitochondrial tree reconstructed with inclusion of mitochondrial genes of *D. lusaltans* (A) and nuclear trees generated with the *Xdh* (B) and *Adh* (C) genes, which includes sequences of *D. subsaltans*. Branch supporter different than 1 are shown. The *parasaltans, sturtevanti, saltans, elliptica* and *cordata* subgroups are highlighted in yellow, blue, red, green and pink, respectively.

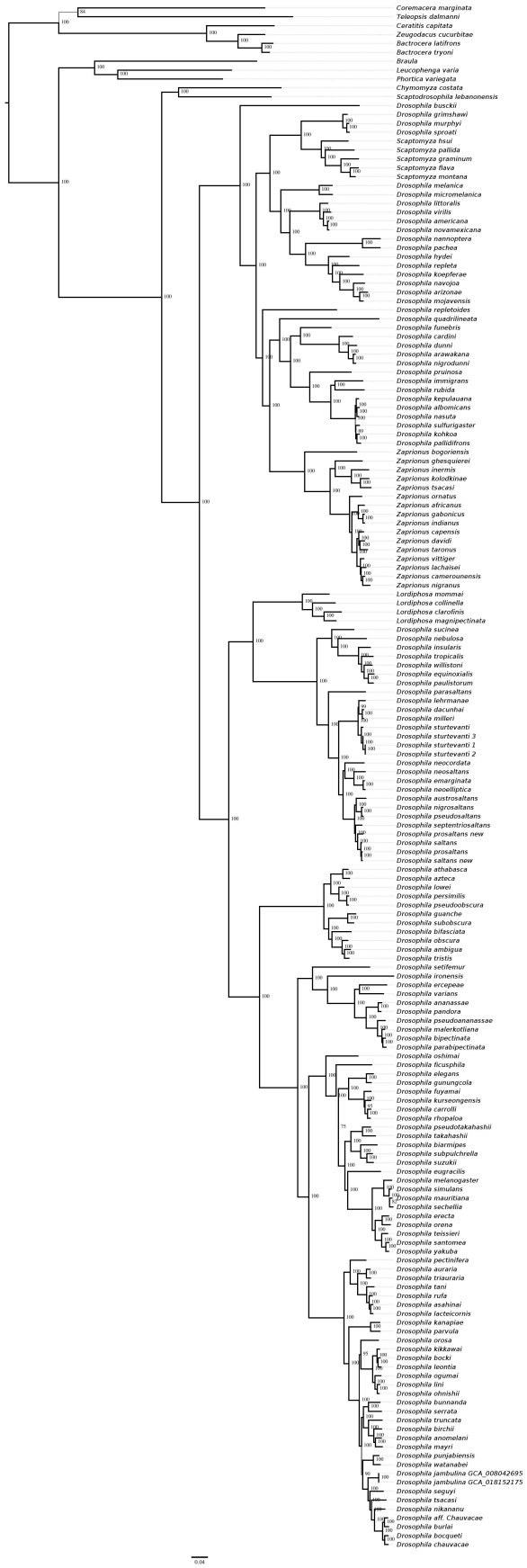


APPENDIX B: supplementary material chapter 4

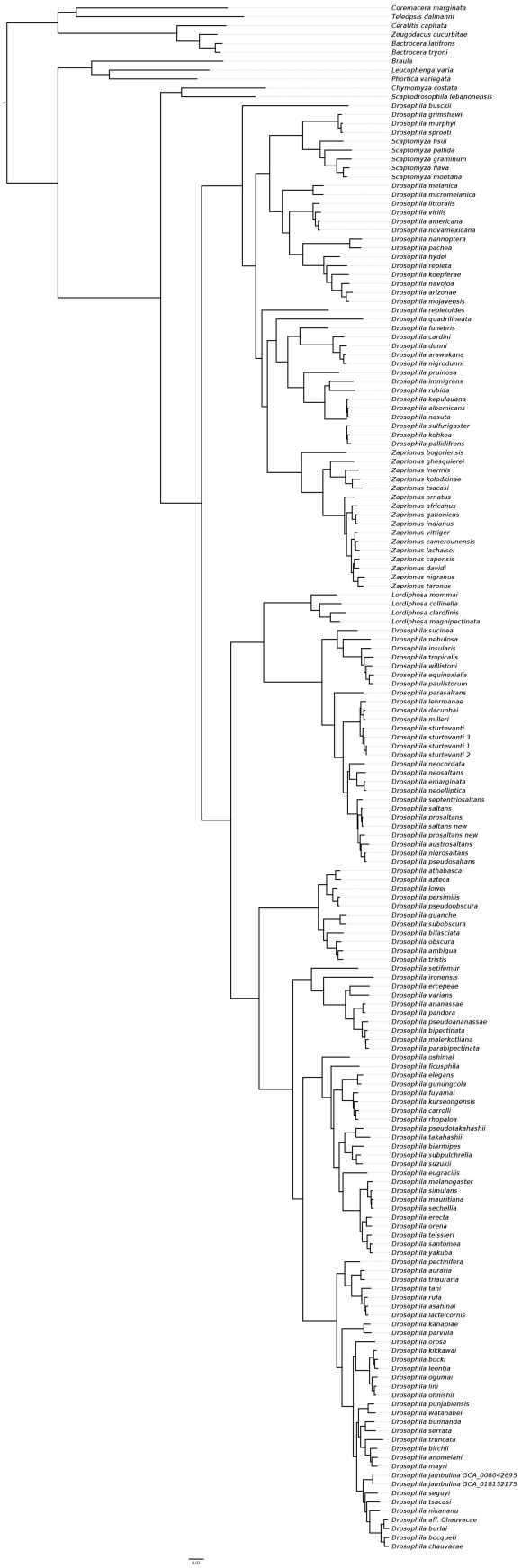
Supplementary Figure S1. Correspondence analysis of codons (A) and their contributions to the first (B) and second (C) diminutions. Red dashed lines in B and C represent the expected values for contributions assuming equal contributions from all factors.



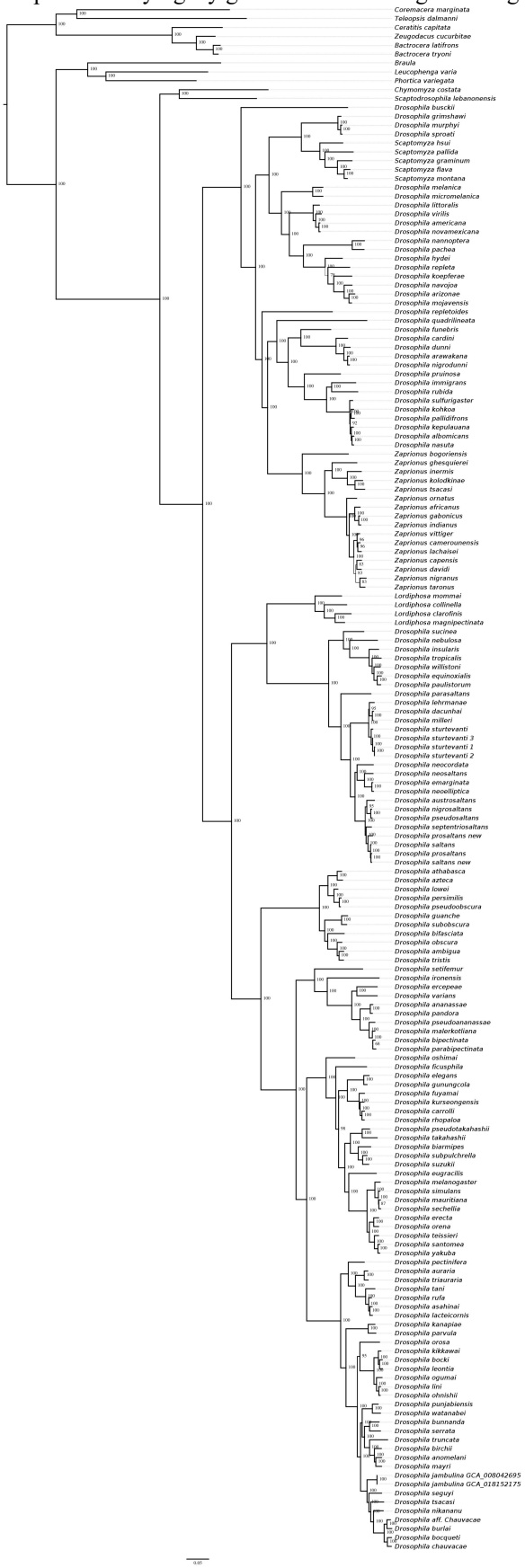
Supplementary Figure S2. Drosophilidae Phylogeny generated with 192 gene sequenced translated to amino-Acids.



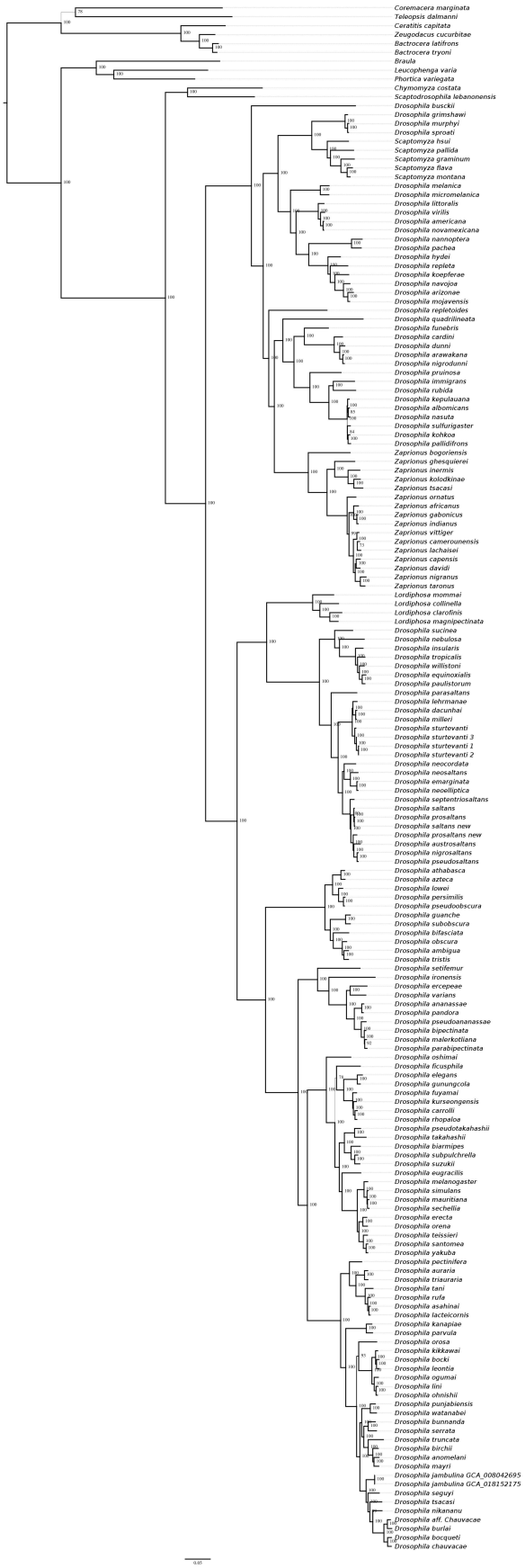
Supplementary Figure S3. Drosophilidae Phylogeny generated with 192 gene using the first and second codon bases.



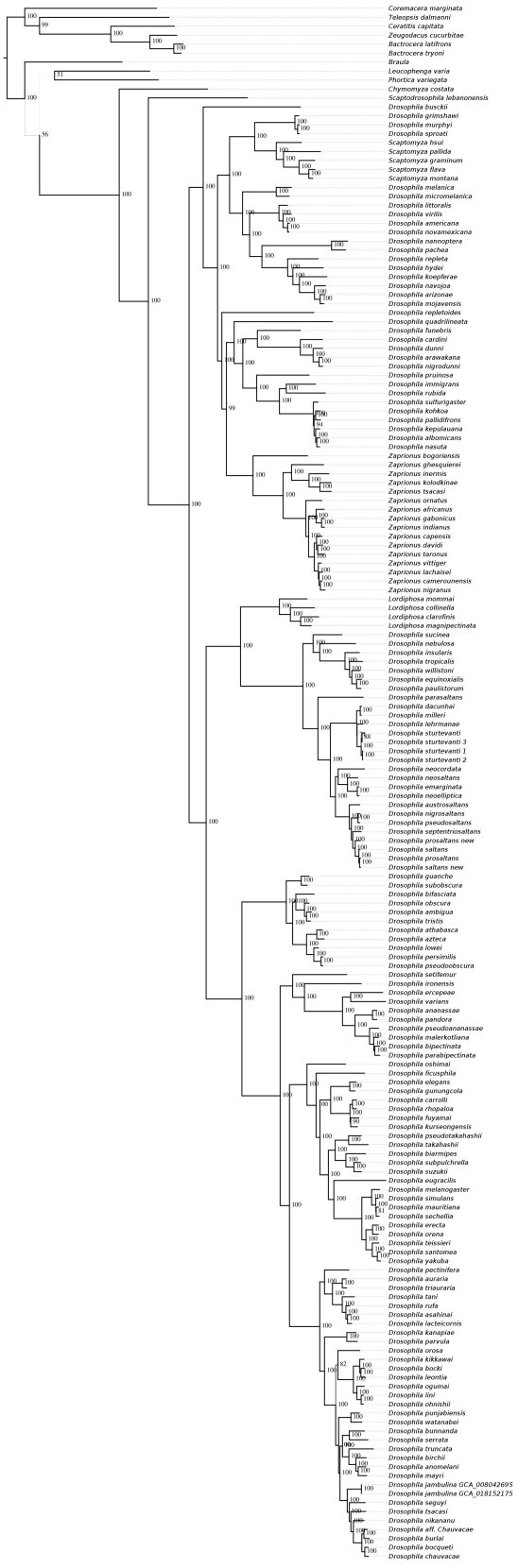
Supplementary Figure S4. Drosophilidae Phylogeny generated with 192 gene using the first codon base.



Supplementary Figure S5. Drosophilidae Phylogeny generated with 192 gene using the second codon base.



Supplementary Figure S6. Drosophilidae Phylogeny generated with 192 gene using the third codon base.



Supplementary Table S1. Total number of recover genes for each evaluated genome.

Species	Number of recover single copy genes (3285 baits genes)
Drosophila sproati	3238
Drosophila tsacasi	3028
Braula	3100
Coremacera marginata	3157
Scaptomyza flava	3088
Drosophila pandora	3125
Drosophila lacteicornis	3184
Drosophila sturtevanti-s3	2969
Drosophila sulfurigaster	3215
Drosophila elegans	3250
Drosophila gunungcola	3207
Drosophila tani	3112
Drosophila orosa	3086
Drosophila lehrmanae	2996
Scaptomyza graminum	3225
Drosophila rubida	3214
Zeugodacus cucurbitae	3235
Drosophila lini	3096
Drosophila jambulina GCA 018152175.1	3229
Drosophila ogumai	3140
Drosophila emarginata	3183
Drosophila leontia	3099
Drosophila mojavensis	3233
Drosophila pectinifera	3217
Drosophila ohnishii	2959
Lordiphosa magnipectinata	3094
Zaprionus indianus	3220
Drosophila pallidifrons	3204
Drosophila sturtevanti ncbi	3217
Drosophila nikananu	3152
Drosophila ercepeae	3248
Drosophila cardini	3211
Drosophila athabasca	3200
Drosophila koepferae	3220
Drosophila immigrans	3231
Drosophila pachea	3233
Bactrocera latifrons	3212
Drosophila ficusphila	3236
Drosophila parabipectinata	3240
Drosophila bunnanda	3224
Teleopsis dalmanni	3058
Drosophila milleri	3183
Drosophila mauritiana	3245
Drosophila teissieri	3222
Drosophila nigrodunni	3218
Drosophila quadrilineata	3237

Species	Number of recover single copy genes (3285 baits genes)
Zaprionus nigranus	3115
Drosophila oshimai	3206
Lordiphosa mommai	3131
Drosophila aff. chauvacae BK-2020	3145
Scaptomyza pallida	3169
Drosophila dunni	3238
Drosophila lowei	3102
Drosophila melanogaster	3234
Drosophila neoelliptica	3194
Phortica variegata	3076
Lordiphosa collinella	3077
Drosophila burlai	3114
Drosophila obscura	3204
Scaptomyza montana	3214
Drosophila truncata	3216
Zaprionus kolodkinae	3231
Drosophila nannoptera	3201
Drosophila subobscura	3226
Drosophila repletoides	3232
Zaprionus africanus	3233
Drosophila melanica	3178
Drosophila repleta	3244
Drosophila kepulauana	3233
Zaprionus lachaisei	3210
Drosophila willistoni	3212
Zaprionus bogoriensis	3243
Drosophila parasaltans	3166
Drosophila novamexicana	3202
Zaprionus capensis	3213
Drosophila subpulchrella	3082
Drosophila birchii	3218
Drosophila auraria	3192
Drosophila rhopaloa	3220
Drosophila malerkotliana	3236
Drosophila takahashii	3198
Drosophila varians	3235
Drosophila murphyi	3238
Drosophila sucinea	3215
Drosophila prosaltans ncbi	3229
Drosophila triauraria	2998
Zaprionus gabonicus	3237
Zaprionus davidi	3220
Drosophila albomicans	3230
Drosophila insularis	3238
Leucophenga varia	2961
Drosophila chauvacae	2983
Drosophila eugracilis	3218
Drosophila septentriosaltans	3156

Species	Number of recover single copy genes (3285 baits genes)
Drosophila carrolli	3213
Drosophila austrosaltans	3069
Drosophila tropicalis	3196
Drosophila watanabei	3008
Drosophila pseudotakahashii	3241
Drosophila santomea	3249
Drosophila seguyi	3182
Drosophila neosaltans	3208
Drosophila punjabiensis	3073
Drosophila anomelani	3104
Drosophila paulistorum	3224
Drosophila littoralis	3243
Drosophila persimilis	3182
Drosophila azteca	3195
Scaptomyza hsui	3212
Zaprionus ornatus	3213
Drosophila fuyamai	3234
Drosophila arawakana	3214
Drosophila neocordata	3232
Drosophila saltans ncbi	3231
Drosophila sturtevanti-s2	3154
Drosophila micromelanica	3204
Drosophila biarmipes	3245
Drosophila sturtevanti-s1	3035
Drosophila ironensis	3200
Drosophila kurseongensis	3239
Drosophila ananassae	3243
Drosophila dacunhai	3190
Drosophila pseudoobscura	3219
Drosophila tristis	3178
Scaptodrosophila lebanonensis	3210
Drosophila suzukii	3105
Drosophila bifasciata	3024
Drosophila nebulosa	3187
Drosophila virilis	3195
Drosophila setifemur	3234
Drosophila jambulina GCA 008042695.1	3202
Drosophila navojoa	3216
Drosophila bipectinata	3235
Drosophila americana	3238
Zaprionus taronus	3141
Zaprionus camerounensis	3226
Drosophila simulans	3246
Drosophila nigrosaltans	3153
Drosophila arizonae	3124
Drosophila kohkoa	3036
Drosophila busckii	3177
Drosophila bocki	3198

Species	Number of recover single copy genes (3285 baits genes)
Drosophila saltans NEW	3162
Drosophila equinoxialis	3158
Drosophila erecta	3245
Drosophila bocqueti	3202
Chymomyza costata	3225
Zaprionus ghesquierei	3218
Drosophila guanche	3234
Drosophila nasuta	3224
Drosophila pseudosaltans	3120
Drosophila parvula	2971
Drosophila yakuba	3247
Ceratitis capitata	3236
Drosophila sechellia	3241
Drosophila funebris	3217
Drosophila grimshawi	3235
Drosophila rufa	3213
Bactrocera tryoni	3210
Zaprionus tsacasi	3232
Drosophila asahinai	3192
Drosophila pseudoananassae	3231
Drosophila orena	3062
Drosophila hydei	3187
Lordiphosa clarofinis	3118
Drosophila kanapiae	3218
Drosophila pruinosa	3233
Zaprionus vittiger	3236
Drosophila ambigua	3211
Drosophila mayri	3220
Drosophila prosaltans new	3114
Drosophila serrata	3132
Zaprionus inermis	3236
Drosophila kikkawai	3217
TOTAL	553627

Supplementary Table S2. Average RSCU calculated from all SCGs. The most frequently used codon for each amino acid is highlighted in blue, and favored codons, denoted by RSCU greater than 1, are displayed in bold. Availlable at : https://drive.google.com/file/d/1foT434g7BhqVricHRZplhq_f3tyjIkZo/view?usp=drive_link

Supplementary Table S3 The relationship between the amino acinds and the tRNAs responsible for carrying them, their codons and the range of their RSCU.

Amino Acid	N. tRNA	Codons	maximum range of RSCU values
Arginine	6 tRNAs	CGT, CGC, CGA, CGG, AGA, AGG	0-6
Leucine	6 tRNAs	TTA, TTG, CTT, CTC, CTA, CTG	0-6
Serine	6 tRNAs	TCT, TCC, TCA, TCG, AGT, AGC	0-6
Alanine	4 tRNAs	GCT, GCC, GCA, GCG	0-4
Glycine	4 tRNAs	GGT, GGC, GGA, GGG	0-4
Proline	4 tRNAs	CCT, CCC, CCA, CCG	0-4
Threonine	4 tRNAs	ACT, ACC, ACA, ACG	0-4
Valine	4 tRNAs	GTT, GTC, GTA, GTG	0-4
Isoleucine	3 tRNAs	ATT, ATC, ATA	0-3
Asparagine	2 tRNAs	AAT, AAC	0-2
Aspartic Acid	2 tRNAs	GAT, GAC	0-2
Cysteine	2 tRNAs	TGT, TGC	0-2
Glutamic Acid	2 tRNAs	GAA, GAG	0-2
Glutamine	2 tRNAs	CAA, CAG	0-2
Histidine	2 tRNAs	CAT, CAC	0-2
Lysine	2 tRNAs	AAA, AAG	0-2
Phenylalanine	2 tRNAs	TTT, TTC	0-2
Tyrosine	2 tRNAs	TAT, TAC	0-2
Methionine	1 tRNA	ATG	1
Tryptophan	1 tRNA	TGG	1

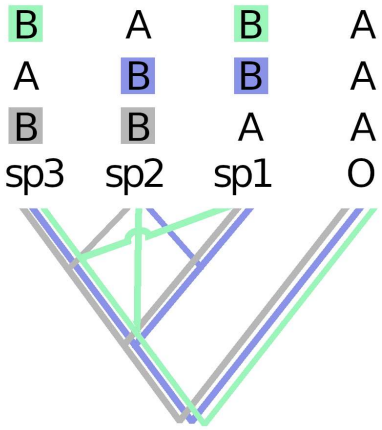
APPENDIX C : Résumé étendu de la thèse

Introduction général

Comprendre les relations entre les espèces et les processus sous-jacents à leur diversification est un objectif fondamental en biologie. Traditionnellement, les relations évolutives des espèces sont représentées dans des arbres, offrant une représentation simplifiée de leurs relations phylogénétiques. Cependant, cette représentation directe devient complexe lorsque des polytomies surviennent, ajoutant une couche de complexité au modèle d'arbre ramifié. Les polytomies sont largement classées en deux catégories : "soft" et "hard". Les polytomies "soft" surviennent lorsque les données disponibles manquent de signal phylogénétique suffisant pour résoudre un nœud, conduisant à un schéma de ramification non résolu, et la topologie changera jusqu'à ce que suffisamment de données de haute qualité soient ajoutées à l'analyse. En revanche, les polytomies "hard" sont intégrées dans la trajectoire historique des espèces, marquée par une influence plus prononcée de l'évolution réticulaire. Divers processus, tels que le transfert horizontal de gènes, le flux génique (comme la spéciation hybride et l'introgression) et la ségrégation incomplète des lignée, contribuent à la configuration complexe des ramifications. Contrairement aux polytomies "soft", les défis posés par les polytomies "hard" vont au-delà des problèmes de qualité des données, s'étendant à la complexité de l'histoire évolutive des espèces. La reconnaissance de ces deux classes de polytomies souligne la complexité inhérente à la représentation des relations évolutives. Cela met en évidence l'interaction dynamique entre les limitations méthodologiques et les processus biologiques complexes qui façonnent l'arbre de la vie. En substance, la reconnaissance des polytomies incite à une exploration plus approfondie de la nature complexe de l'évolution des espèces et des forces diverses qui contribuent aux schémas de ramification observés dans l'arbre évolutif.

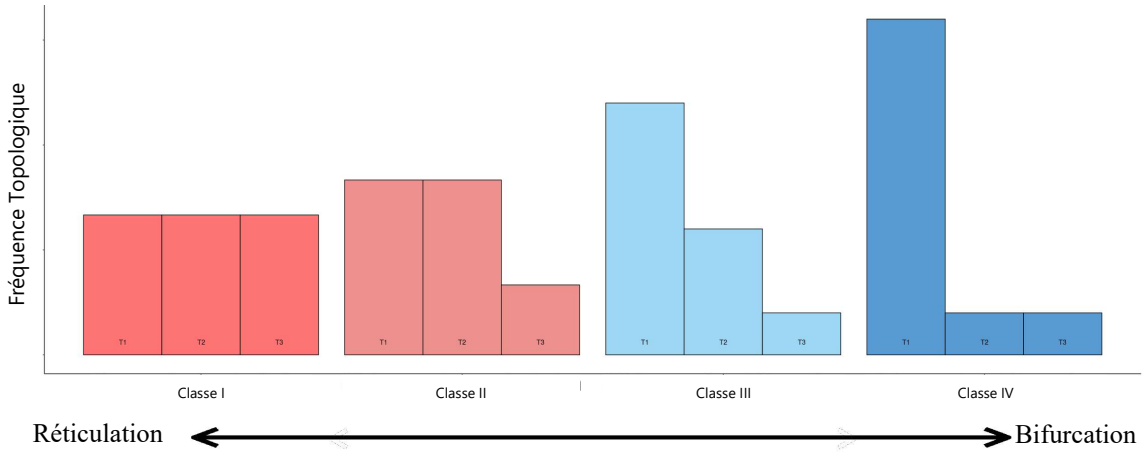
Au cours de la dernière décennie, les efforts pour distinguer tri incomplet des lignées et introgression posent des défis en raison de motifs génétiques similaires qu'ils peuvent produire. Des tests analytiques, tels que le D de Patterson et HyDe, sont utilisés pour démêler ces complexités, en évaluant le partage des mutations dans des paires d'espèces en quadruplets. Pour chaque quadruplet d'espèces, avec un groupe externe connu, trois topologies sont possibles, fréquemment appelées AABB, ABBA et BABA (Figure 1).

Figure 1. Structures topologiques potentielles dans des quatuors d'espèces avec un groupe externe, déduites de la variation génétique partagée dans des paires. Trois topologies distinctes sont possibles : AABB (gris, espèces 2 et 3 comme les plus proches parents), ABBA (bleu, espèces 1 et 2 comme les plus étroitement liées), et BABA (vert, espèces 1 et 3 comme la paire la plus proche). Sp1=espèces 1, Sp2=espèces 2,Sp3=espèces 3, O= groupe extérieur.



La fréquence de ces topologies peut être organisée en quatre classes : (i) réticulation complète, où toutes les topologies (AABB, ABBA et BABA) ont des fréquences égales, ce cas est attendu dans des cas de SIL élevée, comme dans les radiations adaptatives; (ii) réticulation incomplète, où deux topologies dépassent significativement la troisième mais ne diffèrent pas significativement l'une de l'autre, ce qui est attendu dans des cas d'échange bidirectionnel de matériel génétique entre espèces/populations, comme observé dans des cas d'hybridation complète; (iii) bifurcation incomplète, où les proportions de toutes les topologies diffèrent significativement, fréquemment interprétée comme un signal d'introgression; et (iv) bifurcation complète, où une topologie dépasse significativement les deux autres, ces dernières ayant des proportions presque égales (Figure 2). Les tests existants couvrent la classe iii, la classe iv (dans le cas du D de Patterson), et la classe ii (HyDe). Cependant, un test unifié pour une évaluation complète reste insaisissable.

Figure 2. Quatre classes possibles de distributions topologiques de fréquence à travers le génome. Ces classes offrent des perspectives sur divers scénarios évolutifs : classe I, réticulation complète, attendue dans les régions du génome affectées par l'ILS ; classe II, échange bidirectionnel de matériel génétique, interprété comme un signal d'hybridation ; classe III, échange asymétrique de matériel génétique, interprété comme une introgression ; et classe IV, bifurcation complète, attendue dans les événements de spéciation sans ILS et flux génique.



Le biais d'utilisation des codons, c'est-à-dire l'utilisation non aléatoire de codons synonymes, pose des défis pour l'inférence phylogénétique. Alors que les motifs d'utilisation des codons s'alignent généralement au sein d'espèces étroitement apparentées, les branches spécifiques aux clades peuvent présenter des similitudes en raison de la parallélisme, de la convergence ou de l'inversion. Deux explications principales de la variation non aléatoire des codons impliquent la sélection translationnelle et des processus neutres. La sélection translationnelle postule que les préférences de codons améliorent l'efficacité de la traduction, en alignant avec l'abondance des ARNt. Les processus neutres, influencés par les biais mutationnels et la dérive génétique, contribuent également au biais d'utilisation des codons. L'équilibre délicat entre les forces de sélection et neutres façonne le biais d'utilisation des codons, impactant la fitness et jouant potentiellement un rôle dans l'évolution des espèces. Le biais d'utilisation des codons a été associé au mode de vie des organismes, et son rôle potentiel dans le processus de spéciation a été proposé. Investiguer l'importance de l'utilisation des codons tout au long de l'évolution des espèces offre des perspectives sur l'interaction complexe entre la variation génétique, la sélection naturelle et le développement de la biodiversité. Le genre *Drosophila*, en particulier le sous-genre *Sophophora*, est connu pour sa préférence pour l'utilisation de codons se terminant par C et G. Cependant, un changement d'utilisation des codons a été identifié pour le clade néotropical de *Sophophora*, c'est-à-dire les groupes d'espèces *D. saltans* et *D. willistoni*. Et le manque général de biais d'utilisation des codons est rapporté pour ce *clade*. Un modèle pour étudier ces aspects de

l'évolution est le groupe *saltans* de *Drosophila*. Ce groupe d'espèces se compose de 23 espèces réparties en cinq sous-groupes, à savoir *saltans*, *cordata*, *sturtevanti*, *parasaltans* et *elliptica*. De nombreuses études utilisant divers marqueurs ont cherché à élucider les relations phylogénétiques au sein du groupe *saltans*, mais des incohérences persistent et des réserves quant aux relations évolutives subsistent. Les défis découlent des divergences entre les études utilisant différentes sources de données, et les incohérences dans les marqueurs moléculaires peuvent résulter des caractéristiques intrinsèques à l'évolution des espèces. Malgré ces défis, l'utilisation de méthodologies de séquençage de nouvelle génération offre des perspectives prometteuses pour clarifier la phylogénie de ce groupe d'espèces, offrant une multitude de caractères phylogénétiquement informatifs et contribuant à une compréhension plus approfondie de leur histoire évolutive. La grande quantité de données génomiques offre une opportunité d'évaluer le changement d'utilisation des codons qui a été rapporté en utilisant quelques gènes et seulement un génome de *Sophophora* néotropical. L'objectif principal de cette étude a été d'explorer l'évolution au sein du groupe *saltans* de *Drosophila*. Les relations phylogénétiques et l'évolution génomique de 15 espèces réparties dans ses sous-groupes ont été examinées. L'impact du flux génique et du tri incomplet des lignées a également été évalué. Enfin, les motifs d'utilisation des codons ont été analysés, examinant les rôles de la mutation et de la dérive au sein de la famille des Drosophilidae, en se concentrant particulièrement sur le groupe *saltans*.

Article I : La réticulation évolutive du groupe *saltans* de *Drosophila*

Titre du article : Saltational episodes of reticulate evolution in the jumping *Drosophila saltans* species group

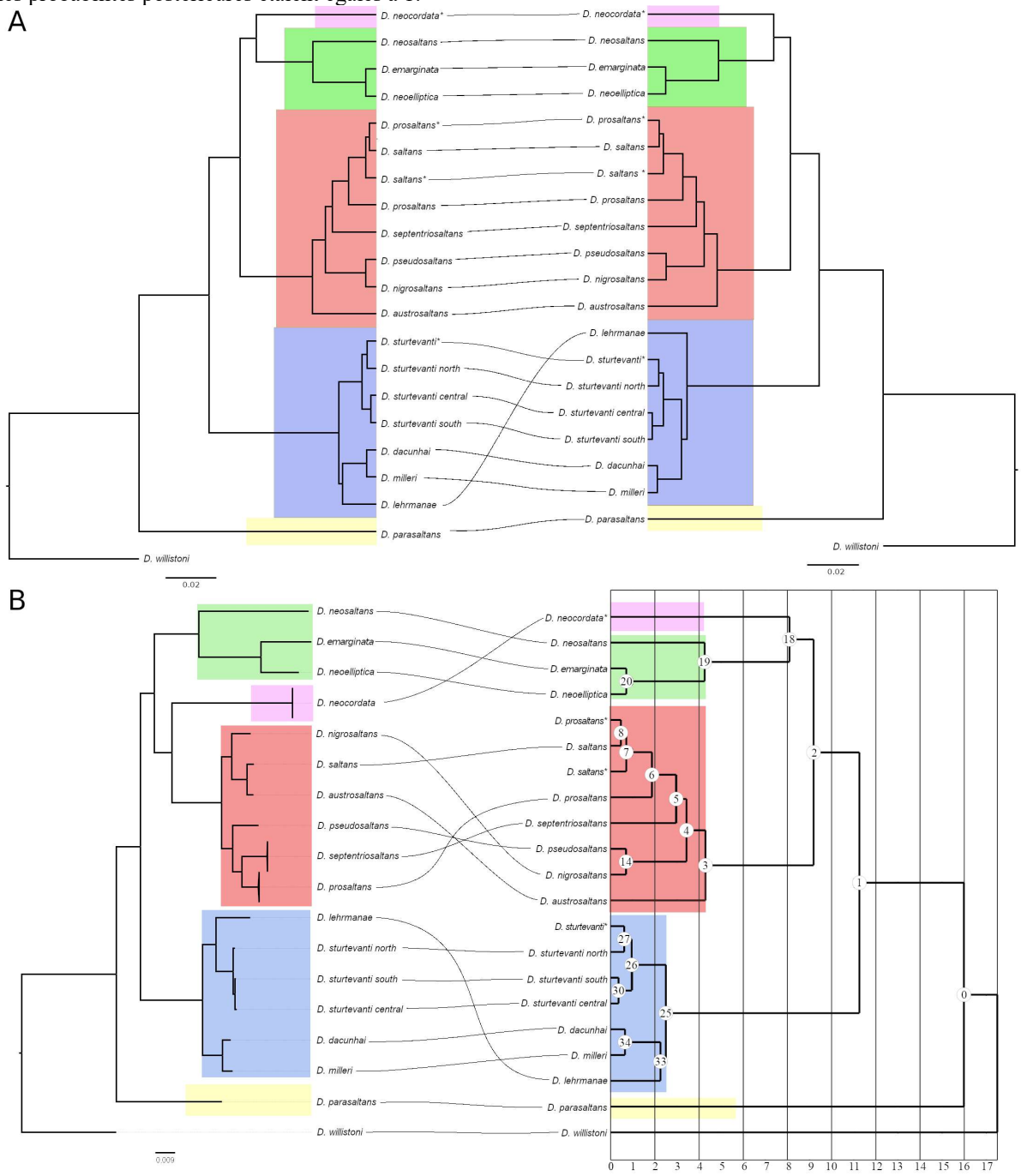
Auteurs: Carolina Prediger, Erina A. Ferreira, Samara Videira Zorzato, Aurélie Hua-Van, Lisa Klasson, Wolfgang J. Miller, Amir Yassin and Lilian Madi-Ravazzi

Dans cette étude, nous avons mené une analyse phylogénomique approfondie de 15 espèces au sein du groupe *saltans*, un clade diversifié de mouches des fruits néotropicales, en utilisant des séquences Illumina de lecture courte et des approches d'assemblage. De nombreuses approches phylogénomiques ont été réalisées, notamment l'évaluation de 5 ensembles de données de plus de 300 gènes uniques chacun et l'utilisation de gènes mitochondriaux pour la reconstruction des relations évolutives de ce groupe, ainsi que l'évaluation des topologies de conflit potentielles. De plus, nous avons généré et appliqué le

test 2A2B, un test pour l'estimation de l'évolution de la réticulation, et comparé la quantité de zone géographique ancestrale partagée et la brièveté de multiples événements de spéciation avec la fréquence du signal de réticulation.

Nos analyses phylogénétiques nucléaires ont révélé des relations cohérentes parmi les cinq sous-groupes du clade *saltans*, plaçant le sous-groupe *parasaltans* comme le premier à se diviser, suivi de *sturtevanti*. Le sous-groupe *saltans* était étroitement lié aux sœurs *cordata-elliptica* (Figure 3A). Cependant, des incongruences ont été observées au sein des sous-groupes, notamment le sturtevanti, la plupart des différences observées dans les différentes méthodes appliquées étant corrélées aux locus des gènes des chromosomes X et autosomes. L'analyse mitogénomique différait de l'analyse nucléaire dans le positionnement de certaines espèces, indicatif d'événements d'introgression cytoplasmique, notamment dans le sous-groupe *saltans*, mais cela a également été observé pour les espèces *D. sturtevanti* et *D. lehrmanae* (sous-groupe *sturtevanti*). Entre les sous-groupes, des différences mito-nucléaires apparaissent concernant le clade *saltans-elliptica-cordata* (Figure 3B).

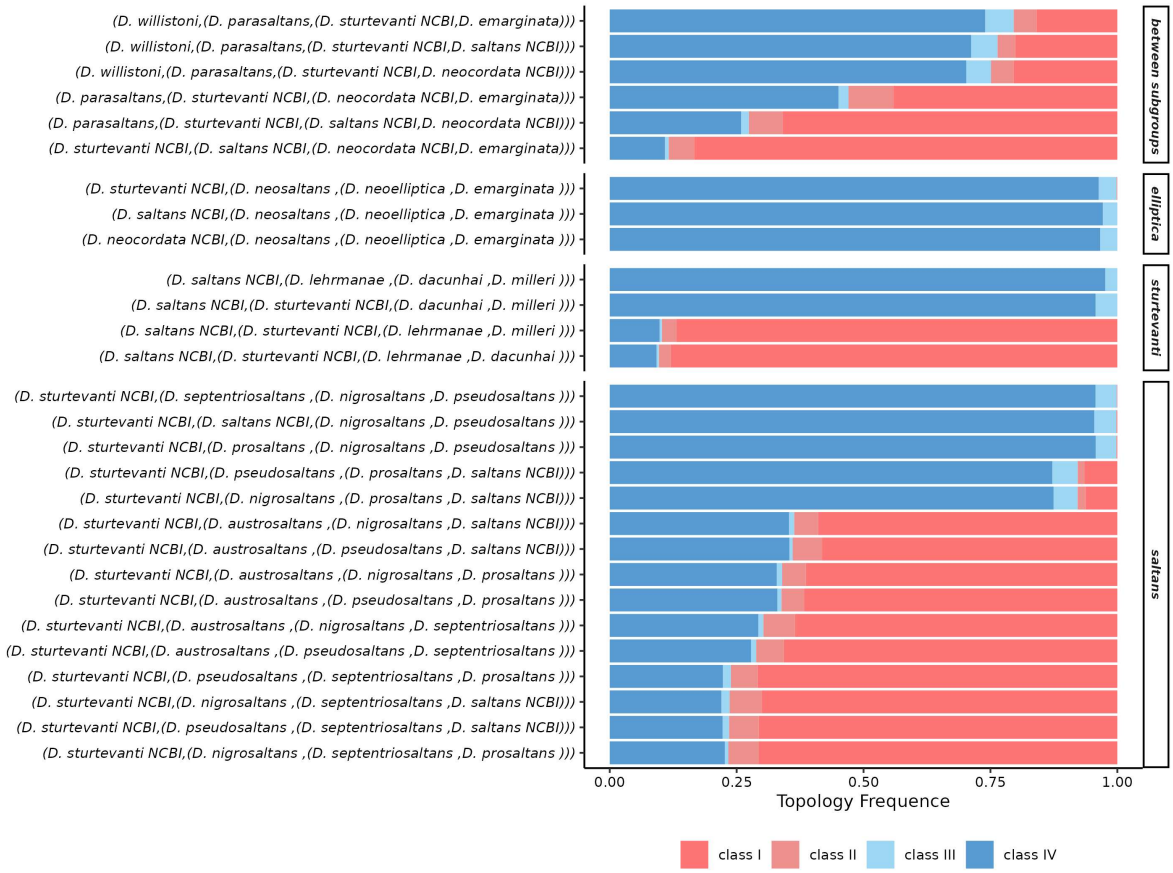
Figure 3. Conflit Phylogénomique du Chromosome X, des Autosomes et des Mitochondries. A) Analyse Comparative de la topologie autosomique (à gauche, représentée par l'élément de Muller B) et de la topologie liée au chromosome X (à droite, représentée par l'élément de Muller A) démontre un accord global avec des incongruences mineures. B) Le Désaccord Mitochondrie-Nucléaire met en évidence une incongruence plus forte entre la Topologie Mitochondriale (à gauche) et la topologie des chromosomes sexuels (à droite). L'estimation du temps de divergence (en millions d'années, Ma) pour la Topologie Chromosomique Sexuelle est fournie. Toutes les probabilités postérieures étaient égales à 1.



En ce qui concerne l'analyse de la réticulation, plusieurs points de signal élevé d'évolution de la réticulation ont été notés, notamment au niveau du clade *saltans-elliptica-cordata* et pour les espèces du sous-groupe *saltans*, ainsi que pour les espèces *D. sturtevanti*

et *D. lehrmanae* (sous-groupe *sturtevanti*) (Figure 4). Ces résultats suggèrent des processus évolutifs complexes, avec des implications pour l'isolement reproductif et la diversité génétique à ces points des relations évolutives du groupe *saltans*.

Figure 4. Le test 2A2B révèle un signal d'introgression diminué, tandis qu'un signal proéminent d'évolution réticulaire est évident au sein de sous-groupes spécifiques. La distribution des fréquences des classes i-iv, allant de la réticulation complète symétrique à la réticulation de bifurcation asymétrique, est affichée pour les espèces en quatuor. Un motif prononcé de réticulation complète est apparent dans les sous-groupes *saltans* et *sturtevanti*, tandis qu'un tel signal est absent dans le sous-groupe *elliptica*.



L'étude a également abordé les défis liés à l'identification géographique erronée et à l'isolement reproductif au sein du sous-groupe *saltans*. L'inclusion d'espèces supplémentaires, telles que *D. subsaltans* et *D. lusaltans*, grâce à des séquençages génomiques complets, promet d'apporter de nouvelles perspectives sur des questions phylogénétiques non résolues. Les disjonctions géographiques et le renforcement potentiel de la reproduction entre *D. saltans* et *D. prosaltans* soulignent la nécessité d'un échantillonnage élargi pour comprendre l'étendue de l'isolement reproductif et de la porosité génomique.

L'analyse des blocs synténiques a fourni une mesure quantitative de l'évolution réticulée, le sous-groupe *saltans* présentant la plus forte incidence de réticulation. Le degré de réticulation était corrélé au temps entre les événements successifs de spéciation et au degré de

conservation des aires géographiques, soulignant l'influence des échelles de temps évolutives et des facteurs géographiques sur les motifs de réticulation.

De plus, l'étude a introduit un nouveau test 2A2B, révélant des informations sur l'étendue de l'introgression et des événements de radiation rapide au sein du groupe *saltans*. La faible incidence d'introgression interspécifique, comparée à d'autres clades de *Drosophila*, suggère que les analyses traditionnelles d'introgression basées sur des arbres de spéciation bifurqués peuvent être trompées en présence de polytomies difficiles.

En conclusion, cette analyse génomique approfondie éclaire les dynamiques évolutives complexes, les conflits et l'évolution réticulée au sein du groupe d'espèces *Drosophila saltans*. Les résultats soulignent l'importance d'étendre les efforts d'échantillonnage, de traiter les erreurs d'identification géographique et de prendre en compte de grands blocs synténiques pour déchiffrer avec précision l'histoire évolutive de ce clade néotropical.

Article II : Changement dans l'utilisation des codons chez *Drosophila*

Titre du article : Ancestral state relaxation and contrasting trends in codon usage across 174 *Drosophila* species

Auteurs: Carolina Prediger, Amir Yassin and Lilian Madi-Ravazzi

Ce chapitre présente les résultats partiels de nos recherches sur l'utilisation des codons dans la famille des Drosophilidae, en mettant l'accent sur le clade néotropical *Sophophora*, composé des groupes *saltans* et *willistoni*. Cela s'explique par l'observation d'un manque de préférence pour l'utilisation de codons dans ce clade, que ce soit dans quelques gènes spécifiques étudiés précédemment ou dans un seul génome. Le travail a utilisé 197 génomes de drosophiles, après une filtration de la complétude génomique, 174 génomes ont été analysés, dont 27 représentants du clade *Sophophora* néotropical. En raison du grand nombre de génomes analysés, 3285 gènes à copie unique ont été sélectionnés pour évaluer la préférence d'utilisation des codons.

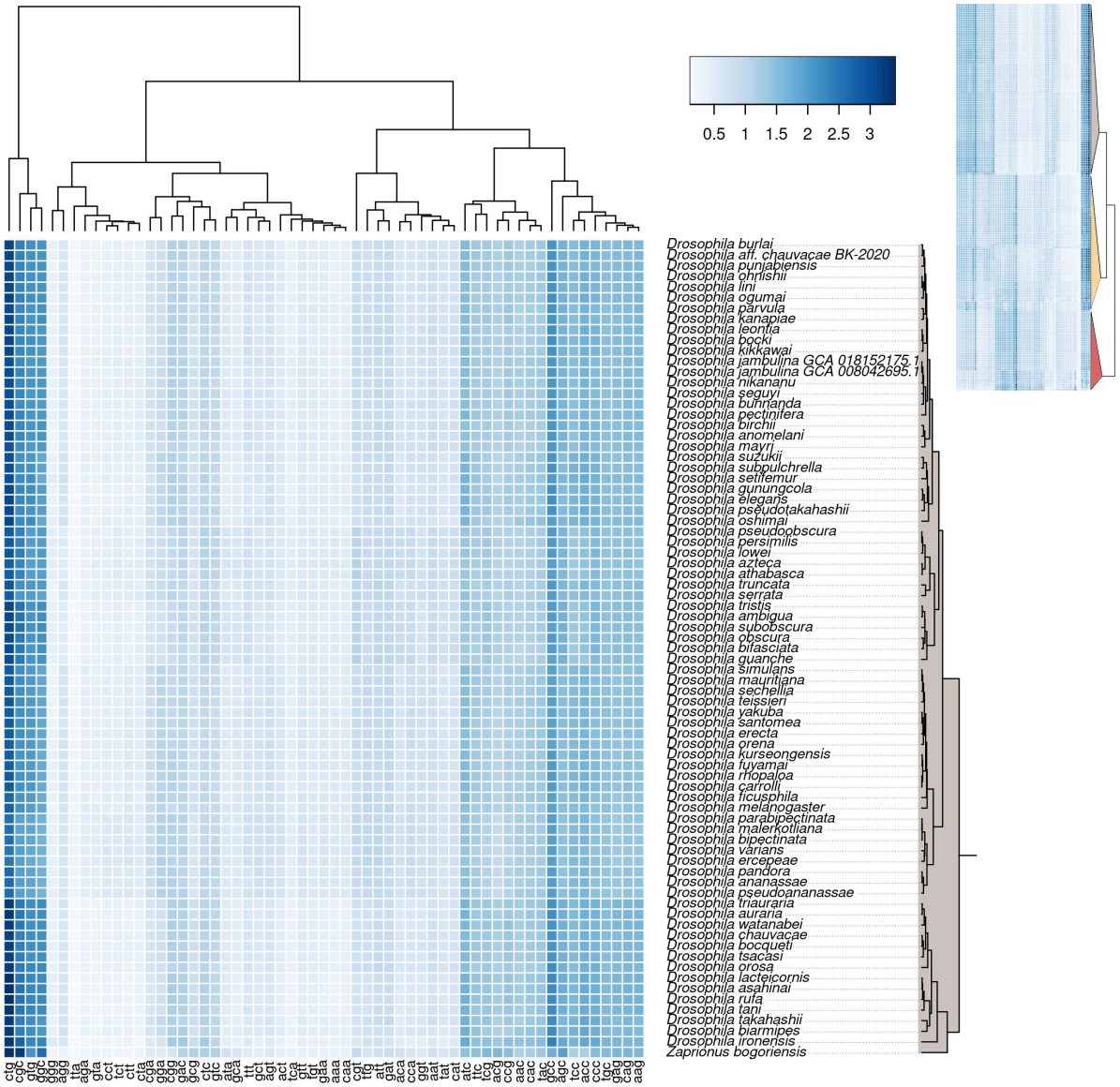
Les résultats ont révélé différents degrés de préférence entre les espèces, le clade *saltans-willistoni* présentant un manque notable de préférence pour les codons, similaire à des taxons externes. Le Calcul du Nombre Effectif de Codons (ENC) a été réalisé pour évaluer la préférence d'utilisation des codons. Les valeurs de l'ENC varient de 20 à 61, indiquant un biais complet à une sélection complètement sans biais. Les moyennes de l'ENC ont varié de 40,61 à 53,92, révélant différents degrés de préférence d'utilisation des codons entre les

espèces. Des biais forts ont été observés chez des espèces telles que *D. ironensis* et *Z. bogoriensis*, tandis que le clade saltans-willistoni présentait un manque notable de préférence pour les codons.

L'analyse de l'Utilisation Relative des Codons (RSCU) et la construction d'un regroupement hiérarchique ont mis en évidence trois clusters principaux. Le premier englobait la *Sophophora* néotropicale, les espèces externes et les drosophiles basales, le deuxième incluait les sous-genres *Dorsiphola*, *Siphlodora* et *Drosophila*, ainsi que *Drosophila* hawaïenne et *Zaprionus*. Le troisième cluster comprenait *Sophophora* - ancien monde et *Zaprionus bogoriensis*. De manière intrigante, dans le groupe *saltans*, *D. parasaltans* présentait un modèle similaire au groupe *willistoni*, indiquant des variations dans les préférences génétiques après la divergence de *D. parasaltans*. Z. bogoriensis a montré un modèle d'utilisation de codons divergent, suggérant une distinction évolutive ou fonctionnelle unique (Figure 5).

Figure 5. L'analyse de l'utilisation relative des codons synonymes montre que le motif d'utilisation des codons dans le clade néotropical de *Sophophora* est plus similaire au motif observé pour l'ancêtre des drosophiles. Les colonnes représentent les codons et les lignes correspondent aux espèces qui ont été évaluées, les cellules de couleur plus foncée correspondent aux codons des ARNt qui sont favorisés.

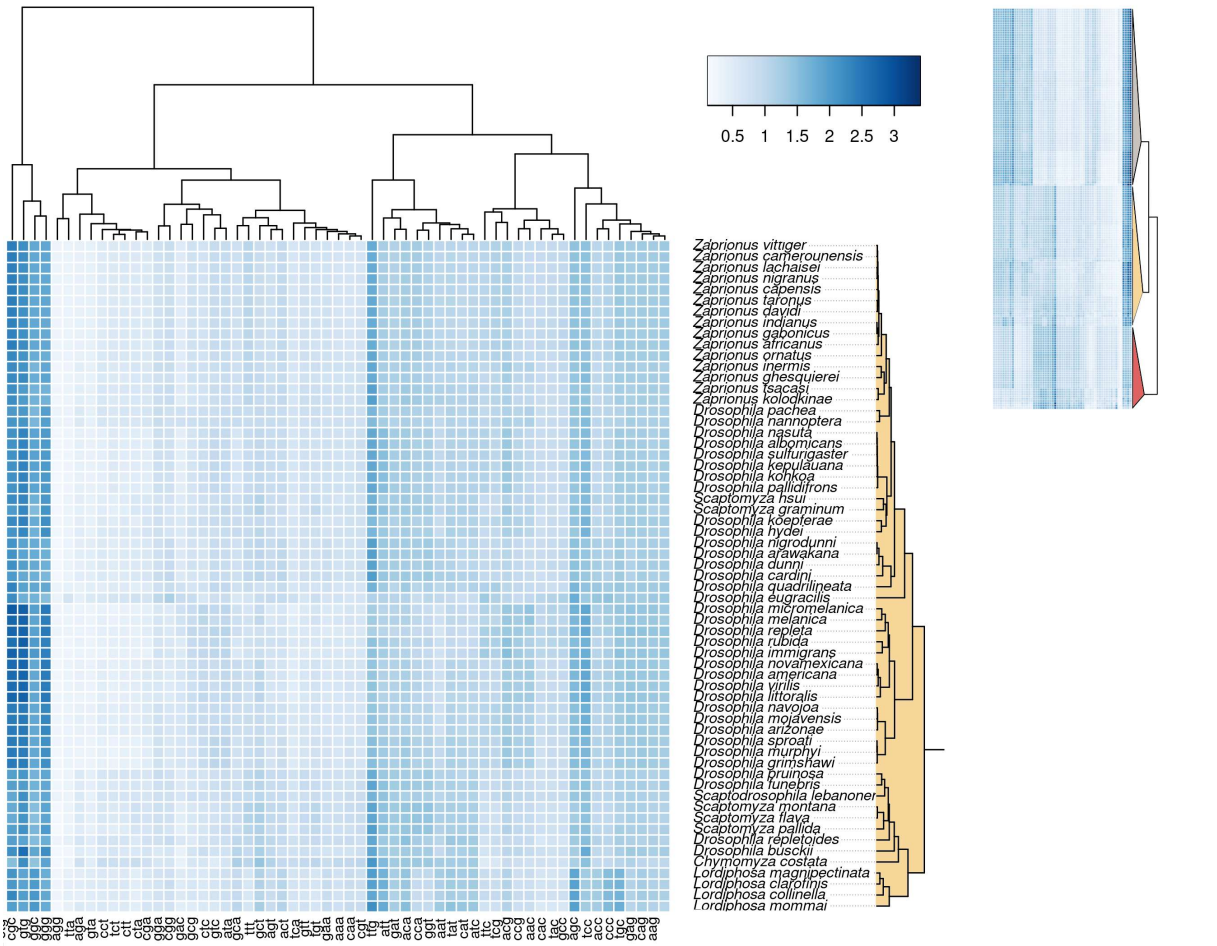
A



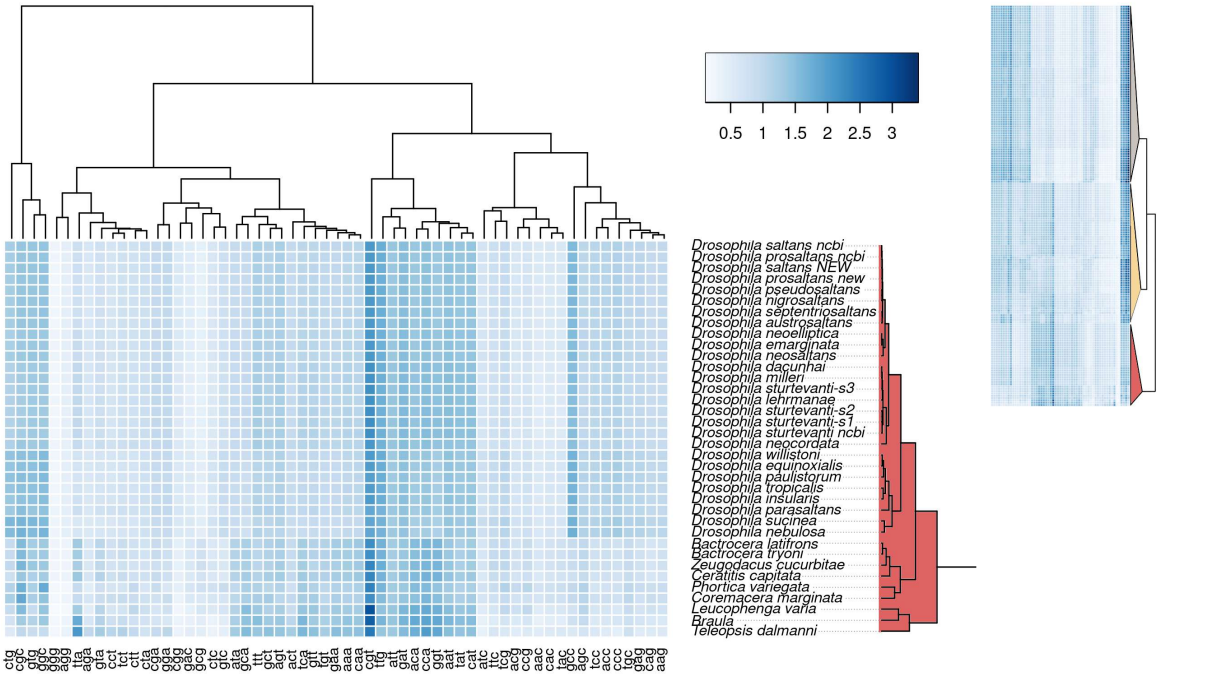
continuer

continue

B


0.5 1 1.5 2 2.5 3
Zaprionus vittiger
Zaprionus carherounensis
Zaprionus lachaisei
Zaprionus nigranus
Zaprionus capensis
Zaprionus taronus
Zaprionus davidi
Zaprionus indianus
Zaprionus gabonicus
Zaprionus africanus
Zaprionus ornatus
Zaprionus inermis
Zaprionus ghesquierei
Zaprionus tsacasi
Zaprionus kolodkinae
Drosophila pachea
Drosophila hannoptera
Drosophila nasuta
Drosophila albomicans
Drosophila sulfurigaster
Drosophila kepulauana
Drosophila kohkoa
Drosophila pallidifrons
Scaptomyza hsui
Scaptomyza graminum
Drosophila koepferae
Drosophila hydei
Drosophila nigrodunni
Drosophila arawakana
Drosophila dunni
Drosophila cardini
Drosophila quadrilineata
Drosophila eugracilis
Drosophila micromelanica
Drosophila melanica
Drosophila repleta
Drosophila rubida
Drosophila immigrans
Drosophila novamexicana
Drosophila americana
Drosophila virilis
Drosophila littoralis
Drosophila navojoa
Drosophila mojavensis
Drosophila arizonae
Drosophila sproati
Drosophila murphyi
Drosophila grimshawi
Drosophila bruneosa
Drosophila funebris
Scaptodrosophila lebanonen
Scaptomyza montana
Scaptomyza flava
Scaptomyza pallida
Drosophila repletoides
Drosophila busckii
Chymomyza costata
Lordiphosa magnipectinata
Lordiphosa clarofinis
Lordiphosa collinella
Lordiphosa mommai

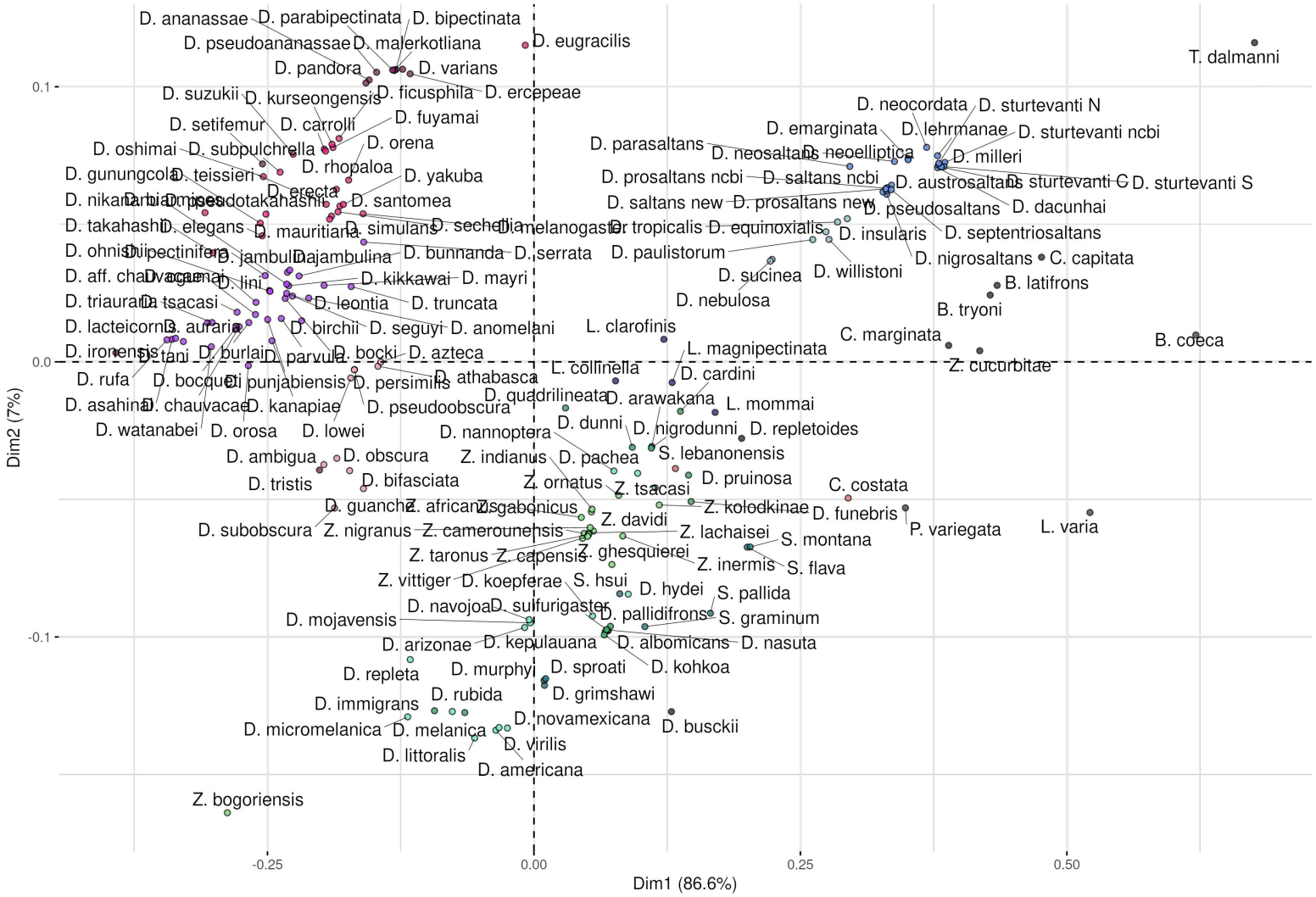
C


0.5 1 1.5 2 2.5 3
Drosophila saltans ncbi
Drosophila prosaltans ncbi
Drosophila saltans NEW
Drosophila prosaltans new
Drosophila pseudosaltans
Drosophila nigrosaltans
Drosophila septentriosaltans
Drosophila austrosaltans
Drosophila neoelliptica
Drosophila emarginata
Drosophila neosaltans
Drosophila dacunhai
Drosophila milleri
Drosophila sturtevanti-s3
Drosophila lehrmanae
Drosophila sturtevanti-s2
Drosophila sturtevanti-s1
Drosophila sturtevanti ncbi
Drosophila neocordata
Drosophila willistoni
Drosophila equinoxialis
Drosophila paulistorum
Drosophila tropicalis
Drosophila insularis
Drosophila parasaltans
Drosophila sucinea
Drosophila nebulosa
Bactrocera latifrons
Bactrocera tryoni
Zeugodacus cucurbitae
Ceratitis capitata
Phortica variegata
Coremacera marginata
Leucophenga varia
Braula
Teleopsis dalmanni

L'analyse de correspondance utilisant RSCU a révélé des similitudes entre le clade *saltans-willistoni* et les sous-genres *Dorsiphola*, *Siphlodora*, *Drosophila* et *Zaprionus*, mettant en évidence l'influence de codons spécifiques dans les différences observées. L'analyse a séparé trois clusters principaux : *saltans-willistoni*, *Sophophora* - ancien monde, et *Lordiphosa* et *Zaprionus*. Des différences notables dans le modèle d'utilisation des codons ont été observées chez *Z. bogoriensis* et d'autres espèces de *Zaprionus*, ainsi que chez certaines espèces divergeant de leurs proches parents, telles que *D. ironensis*, *D. setifemur*, *D. tristis*, *D. eugracilis*, *D. immigrans*, *D. melanica*, *D. nannoptera* et *D. pachea*. (Figure 6).

Figure 6. Les analyses de correspondance de l'utilisation relative moyenne des codons synonymes entre les espèces permettent de récupérer 3 clusters majeurs : I - néotropical *Sophophora*, II - *Sophophora* - vieux monde et III - genres *Lordiphosa* et *Zaprionus*, ainsi que les sous-genres *Drosophila*, *Siphlodora* et *Dorsilopha*. Le graphique montre chacun des 174 génomes examinés le long des deux premières dimensions d'une analyse de correspondance.

.



Ces découvertes mettent en lumière des nuances complexes dans les préférences de codons, indiquant des distinctions évolutives ou fonctionnelles même entre des espèces étroitement apparentées. Les résultats suggèrent également un possible changement évolutif dans le modèle d'utilisation des codons dans le clade saltans-willistoni. L'analyse de correspondance offre une perspective supplémentaire, montrant des similitudes entre saltans-willistoni et d'autres sous-genres en termes de RSCU, soulignant l'importance de codons spécifiques dans ces différences. Ces découvertes contribuent à une compréhension plus approfondie de la dynamique évolutive et fonctionnelle sous-jacente aux modèles d'utilisation des codons chez les Drosophilidae. L'influence du biais mutagène a été examinée dans plusieurs espèces en comparant l'ENC avec le contenu GC3. Même chez des espèces avec un faible ENC, comme D. ironensis, un impact significatif du biais mutagène a été observé, indiquant que la force de sélection pousse les valeurs d'ENC observées vers le bas.