Evolutionary landscape of dipteran insects

A Thesis Submitted in Partial Fulfilment of the Requirement for the Degree of

Doctor of Philosophy

by

Debajyoti Kabiraj

Department of Biosciences and Bioengineering

Indian Institute of Technology Guwahati

Guwahati, Assam-781039, India

May 2022

TH-2821_146106003

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Department of Biosciences and Bioengineering

DECLARATION

This is to declare that the content embodied in this thesis entitled **"Evolutionary landscape of dipteran insects"** is the result of investigations carried out by me under the supervission of **Prof. Utpal Bora**, and is submitted to Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam-781039, India for the award of degree of **Doctor of Philosophy in Biosciences and Bioengineering**. This work has not been submitted elsewhere for any degree or diploma of any institute or university to the best of knowledge and belief.

In keeping with the general practice of reporting scientific investigations, due acknowledgements have been made wherever the work of other investigators are referred.

Debajyoti Kabiraj

Debajyoti Kabiraj
Roll No- 146106003
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati,
Guwahati, Assam-781039, India

Guwahati

May, 2022

TH-2821_146106003

INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI

Department of Biosciences and Bioengineering

CERTIFICATE

This is to certify that the work embodied in this thesis entitled **"Evolutionary landscape of dipteran insects"** is the result of the investigations carried out by **Debajyoti Kabiraj (Roll No- 146106003)** under my supervission in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam-781039, India and is submitted for the award of degree of **Doctor of Philosophy in Biosciences and Bioengineering**. This work has not been submitted elsewhere for a degree.

Prof. Utpal Bora
Thesis Supervisor,
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati,
Guwahati, Assam-781039, India

Guwahati

May, 2022

TH-2821_146106003

Dedication

This Thesis is dedicated to my mother,

***Late Manika Kabiraj**,*

May her memory forever be a

comfort and a blessing.

TH-2821_146106003

ACKNOWLEDGEMENTS

As I am on the verge of this phase in my life, I would like to express my heartfelt gratitude to the people who have trusted and supported me in developing it into a successful one. I am extremely grateful and indebted to my thesis supervisor ***Prof. Utpal Bora*** *for introducing me to the exciting world genomics and evolutionary science. I thank him for providing me the opportunity to study on the issues related to one of the most valuable natural and economic resources of Assam and contribute towards its knowledgebase. I also thank him for encouraging me to perform my research independently and for tutoring me to develop my scientific communication and interpersonal skills. I hope I've been able to imbibe his enthusiasm and boldness before I plunge into the ocean of scientific explorations.*

I would like to thank the members of my doctoral committee, ***Prof. Kannan Pakshirajan*** *(Chairperson),* ***Prof. Ranjan Tamuli***, ***Prof. Karuna Kalita***, *and* ***Prof. Bulu Pradhan***, *for their valuable suggestions, encouragement, and scientific guidance, which constantly enabled me to improve my work.*

I would take this opportunity to acknowledge the present and past ***Heads of the Department of Biosciences and Bioengineering*** *for providing all essential facilities and a conductive academic environment. I would also like to convey my gratitude to the* ***Department of Biosciences and Bioengineering***, *Institutional Biotech Hub at the* ***Centre for the Environment*** *and* ***Param-Ishan***, ***IIT Guwahati's high-performance computing cluster*** *for providing me all the necessary facilities to pursue my research.*

I sincerely acknowledge the financial support from ***Ministry of Human Resource Development (MHRD)***, *Government of India for providing me fellowship as well as* ***Department of Biotechnology (DBT)***, ***Government of India*** *and* ***Central Silk Board*** *for* ***funding our laboratory***.

TH-2821_146106003

I would like to express my gratitude to ***Dr. Kartik Neog*** *and* ***Dr. Palash Dutta*** *from* ***Central Muga Eri Research and Training Institute (CMER&TI) Lahdoigarh*** *for their cooperation in sample collection for my Ph.D research.*

I would like to thank all the staff members of ***Department of Biosciences and Bioengineering, Centre for the Environment, Student Affairs Section, Research & Development Section, and Academic Affairs Section*** *for their constant co-operation.*

My lab mates have been a wonderful group of individuals with a variety of personalities and interests, and I've learned a lot from them. A note of gratitude to my inspiring seniors ***Dr. Arghya Sett***, ***Dr. Deepika Singh***, ***Dr. Suradip Das***, ***Dr. Sambhabi***, ***Dr. Sunita Ojha***, ***Dr. Papori Borgohain*** *and* ***Ms. Swagata Sharma*** *for their guidance and companionship; to my outstanding peers* ***Dr. Hasnahana Chetia*** *and* ***Mr. Vimal Moshahari*** *for their friendship, guidance and all-round support; to my aspiring juniors* ***Mr. Jon Jyoti Kalita***, ***Mr. Adhiraj Nath***, ***Ms. Dharitri Saikia***, ***Ms. Biju Bharali***, ***Mr. Pulakeswar Basumatari***, ***Ms. Tinka Singh*** *for their enthusiastic assistance and cooperation and to* ***Pragya Ma'am*** *for her care and culinary prowess. I also like to extend my love and gratitude towards my trainees,* ***Mr. Vijay Daharia (MTP)***, ***Ms. Kavya***, ***Ms. Parishmita***, ***Ms. Diksha***, ***Ms. Ayesha***, ***Mr. Priyanuj*** *for their desire and dedication.*

I would like to thank my friends and seniors in the Institute especially Dr. S. Mukherjee, Dr. P. Sarkar, Dr. N. Mandal, Dr. G. Saha, Dr. P. Das, Dr. I. Das, Mr. S. Roy, Ms. S. Dutta Choudhury, Mr. S. Jana, Mr. A. Sinha, Dr. D. Kumar for overall support during my Ph.D.

Last but not least I express my deepest sense of gratitude and love to my Parents, my younger brother and family friends for their irresistible love, constant support and patience.

May 2022 ***- Debajyoti***

TH-2821_146106003

Table of Contents

TH-2821_146106003

TH-2821_146106003

TH-2821_146106003

TH-2821_146106003

TH-2821_146106003

SYNOPSIS

"Satisfaction of one's curiosity is one of the greatest sources of happiness in life."

— Linus Pauling

TH-2821_146106003

SYNOPSIS

Insects are the most successful animals present on earth since 480 Mya, and holometabola insects account for about ~87% of the entire insect species. Their diversification time is related to the radiation of flowering plants at about 140-150 Mya. However, the first true fly evolved on Earth around 260-290 Mya, and since then, flies have gone through three episodes of rapid radiation following three major extinctions. The flies are noted for their diverse dietary behaviours across a wide range of climates and morphological stages. Tachinidae are the most common Diptera fly, known for their koinobiont, endoparasitism, and unusual respiration inside a wide variety of arthropod hosts, and are responsible for the deaths of approximately 80% of other insects. The Uzi fly is an assemblage of flies belonging to the Tachinidae family that causes serious economic harm to sericulture industries. The Muga silkworm (*Antheraea assamensis*) is infested with *Blepharipa* sp., a typical Uzi fly prevalent in the Brahmaputra valley's adjacent areas, mainly in Assam and Meghalaya. Exploration of the biological basis of such threatening organisms will definitely reveal new information about their molecular functions.

A typical metazoan mitogenome is small, maternally inherited, mutation prone, with low or no homologous recombination, conserved gene content, and high genetic polymorphism, making it a potential sequence for barcoding, phylogeography, phylogenetic, and molecular dating research. Variation in codon usage represents an evolutionary strategy to modulate gene expression, and it is correlated with protein level, precisely, translational bias, which has been

frequently observed due to differential codon usage in diverse prokaryotes as well as in eukaryotes. Every species requires energy to live a successful life, and mitochondria, as the powerhouse of the cell, are affected by a variety of factors, including climate, temperature, and nutrition. Therefore, as selection acts on phenotypes of translation and energy metabolism, studying nucleotide substitution and codon adaptation are quite genuine ways to infer the patterns of evolution and environmental adaptation in dipteran mitochondria.

Phylogenetics is the reconstruction of the ancestral history of a set of taxa, involves identifying homologous characters across a given set of data, and determining the evolutionary history of species by comparing those characters using tree reconstruction techniques. Rapid expansion of large genomics datasets for phylogenetic analysis benefits biologists, yet such data expansion poses a slew of analytical challenges. These instances frequently feature situations in which various methods infer evidence for distinct phylogenetic resolutions. Many earlier investigations have found mitochondrial protein coding genes to be troublesome. As a result, genes selected for phylogenetic reconstruction should include the necessary information to deliver the best phylogenetic tree resolution for a particular taxon. Generally, single-gene phylogenies are not well suited for proper species tree resolution due to a lower number of informative positions and the presence of stochastic noise. And the use of large data sets might reduce stochasticity, increasing systematic errors like base compositional heterogeneity, among-site rate variation, heterotachy may lead to inconsistencies in different sites and can result in inaccurate phylogenetic relationships regardless of the inference method used. Increasing taxon sampling leads to variation in evolutionary rates among lineages as a result of multiple substitutions at the same site (homoplasy), causing a tree to suffer from a problem known as the Long-branch attraction (LBA). The utilization of larger taxa or characters is still debatable in the design of phylogenetic studies as larger character sets contribute both to phylogenetic signal and noise. On the other side, the informativeness of increasing taxonomic

sampling is crucially dependent on the chronology of the ancestral lineage of the taxa added to the data set.

Complete phylogenetic trees are the result of speciation and extinction events, while diversification balances both the events, and their relative roles can be represented by different diversification models. The discordance between paleontological and molecular age estimates, or between ages estimated from different molecular datasets, is fairly common across different studies across the tree of life. Multiple factors may have been invoked to explain such conflicting age estimates, including paleontological calibration strategy, nucleotide substitution rate heterogeneity, and the appropriateness of the molecular clock used.

Given all issues related to molecular phylogenetics and molecular dating following objectives were formulated to gain a deeper understanding of parasitism of Tachinidae flies, phylogenetic relationship of Diptera evolution, adaptability and diversification events.

Objective 1: Mitochondrial genome sequencing of *Blepharipa sp.*, an endoparasite of Muga silkworm, and comparative codon usage analysis with other Oestroidea flies.

Objective 2: Reconstruction of Diptera phylogeny with larger taxa.

Objective 3: Reconstruction of Dipteran phylogeny and molecular dating using larger dataset (Orthologous genes) and comparative analysis.

Objective 4: Classification of Diptera on the basis of different dietary adaptation and the tempo and mode of their diversification.

Objective 5: Deciphering major diversification time through molecular dating and dynamics of speciation-extinction events of Diptera.

THESIS TITLE: Evolutionary landscape of dipteran insects

CHAPTER 1 of the thesis provides some general facts about Diptera and their evolutionary perception as described in earlier studies. Later in this Chapter we also mentioned different challenges related with phylogenetics and evolutionary analysis with molecular data.

CHAPTER 2 reports first full mitogenome of a dipteran parasitoid of the Muga silkworm (*Antheraea assamensis*) found in the Indian states of Assam and Meghalaya. The complete mitochondrial genome of Blepharipa sp. (Uziflies, Family: Tachinidae) (Acc: KY644698, 15080 bp, A+T = 78.41%) was sequenced by the Next-Generation Sequencing technique. The mitogenome of Blepharipa sp. comprises typical dipteran gene organisation and number (37 genes: 13 protein coding genes, 22 tRNAs, and 2 rRNAs). This mitochondrial genome was multifocally compared with other species of the same superfamily (Oestroidea) as well as some species from the order Diptera. This study has confirmed that Blepharipa sp. mitogenome gene content and arrangement is similar to other Tachinidae and Sarcophagidae flies of the Oestroidea superfamily, typical of ancestral Diptera. However, Calliphoridae and Oestridae flies have undergone tRNA translocation and insertion, forming unique intergenic spacers (IGS) and overlapping regions (OL), and a few of them (IGS, OL) have been conserved across Oestroidea flies. Tandem repeat variation and sequence coverage of the control region (CR) influence mitogenome size, and tachinid flies have a smaller mitogenome due to their small CR (*Blepharipa* sp.: 168 bp). This study shows that mitogenomes of the Tachinidae family have high AT content and AT biased codons in their protein-coding genes (PCGs) compared to their Oestroidea counterpart. The PCGs of this new species are highly skewed towards AT content, and 92.07% of all (3722) codons have A or T in their 3rd codon position. The neutrality test shows that natural selection has a stronger influence on codon usage bias than directed

mutational pressure. This study also reveals that longer PCGs (e.g., *nad5*, *cox1*) have a higher codon usage bias than shorter PCGs (e.g., *atp8*, *nad4l*). The divergence rates increase nonlinearly as AT content at the 3rd codon position increases, and a higher rate of synonymous divergence than nonsynonymous divergence causes strong purifying selection. The phylogenetic analysis explains that *Blepharipa* sp. is well suited to the family of insectivorous tachinid maggots. It's possible that biased codon usage in the Tachinidae family reduces the effective number of codons, while purifying selection preserves the fundamental activities in their mitogenome, aiding efficient metabolism in their endo-parasitic life strategy.

CHAPTER 3 lent a drawback of improper phylogenetic relationships among different Diptera flies. Here, in this chapter, we reconstructed phylogeny by increasing the taxa up to 112 Diptera species and 4 outgroup species using multiple homogenous models and methods. As a result, the homogenous models yield some substandard phylogenetic resolution. The profile of phylogenetic informativeness (PI), which captures the signal inferred by the tree building methods, indicates that inconsistency of PI persists across the depth of the tree. This is due to the presence of several kinds of noise from the dataset, namely, synonymous codon substitution, compositional heterogeneity among taxa, among site composition heterogeneity, heterotachy within-site rate variations, homoplasy (character shared by a set of species but not present in their common ancestor), substitution saturation, or reticulate evolution. The coalescent-based species tree also exposed gene-tree discordance with Diptera's phylogenetic backbone. Furthermore, examination of datasets with networks reveals obvious ambiguity in the signal observed as a result of rapid and reticulate diversifications, which may explain gene conflicts and the difficulties in resolving evolutionary relationships. Later in this section, we also observed some evidence of gene flow between distantly related species. The signs of other

historical processes, such as recombination or reticulation, and their outcomes invoked the idea of inferring unique features from the data without limiting our focus to a single tree.

CHAPTER 4 describes phylogenetic reconstruction of Diptera using a large genomic dataset. The learning from the previous chapter was that increasing the taxon set led to systematic error in phylogenetic relations. Here in this chapter, constraining the taxon set (52 Diptera, 2 Out-group), we increased the amount of data by identifying the orthologous protein coding genes from the available genome in the public databases. A general comparison and correlation study was done on the identified genes, such as the correlation of the size of the genomes between total gene and CDS size. Correlation among different codon usage indices of 52 species of identified orthologous genes. Then we identified 335 common orthologous genes that are present in all 54 taxa and used them for phylogenetic analysis. Different concatenation and coalescent based methods were applied for building phylogenetic trees. The Internode certainty analysis suggests that some nodes have incongruence in the backbone of the species tree. In addition, the divergence time estimated using orthologous genes indicates that Diptera originated during the Late Permian to Late Triassic period. The comparison of phylogenetic trees using multidimensional scaling of different distance-based methods in treespace provides multiple clusters of similar species trees and gene trees.

CHAPTER 5 investigates whether nucleotide substitutions in mitochondrial genes have influenced the evolution of successful niches in various climates, lifestyles, and diets. Because mitochondria provide more than 95% of the cell's energy, the bioenergetic efficiency of mitochondrial ATP generation is dependent on the variability of nutrient molecules and the availability of oxygen. Our investigation established that the food habits of Diptera are significantly correlated with nucleotide substitution, either nonsynonymous (dN) or

synonymous (dS) or their ratio (dN/dS). Variations in selection patterns in mitochondrial genes due to changes in the source of nutrients mean that dietary ability has a differential effect on mitochondrial energy metabolism. Further, we explored the variation in diversification imposed by different dietary habits. In addition, the role of dS, dN, and ω as continuous traits is explored, and we employ phylogenetic modelling to investigate the tempo and mode of evolutionary diversification of these organisms. This chapter also looks at whether species with different food habits evolved in a random or deterministic fashion, and if it is not random, whether they merge around a single or many optimal values of molecular traits.

CHAPTER 6 describes the molecular dating analysis using different approaches and comparative analysis of their results. We estimated the diversification time of dipteran major clades from mitochondrial and genomic datasets using different calibration methods. We compared the variation of time assuming different nucleotide substitution rates (similar, different), different branching prior processes (yule and birth-death), and different molecular dating methods. Results exhibit that by increasing the number of calibration points, the time of diversification events becomes older in stem lineages. Investigation of the diversification patterns and rate shift on the time calibrated trees shows mainly 2-3 diversification rates shifted in different time trees. The results indicate that different dating methods have a significant impact on estimating the chronology of major Diptera lineages and subsequent macroevolutionary analyses. We observed a variation in diversification rates associated with different Diptera traits, as well as diversification that is dependent on diversity. This chapter also explores the role of paleoenvironmental instability in Diptera speciation and extinction events over the course of their evolution. This study suggests the importance of a combination of factors rather than a single explanation in explaining lineage diversification.

CHAPTER 7 summarizes the thesis chapter by chapter, outlining its limitations and prospects for the future. The entire study was developed with the intent of learning more about the

evolution of Diptera flies. The central focus of the thesis is to reconstruct Diptera phylogeny using two major molecular datasets (mitochondrial and genomic) and to correlate various molecular phenomena with physical and environmental parameters. The mitochondrial genome of *Blepharipa* sp., an endoparasite of the Muga silkworm, is described in Chapter 2 along with the phylogenetic relationship of the Oestroidea superfamily (n = 36) based on mitochondrial genes. In Chapter 3, the taxon sampling is expanded (Diptera = 112) and the impact of several factors that pose difficulty during Diptera phylogeny reconstruction is explored. The dataset is increased in Chapter 4, and nuclear orthologous genes are used to reconstruct Diptera phylogeny and estimate divergence time. The association between food pattern and nucleotide substitution rate of mitochondrial DNA is inferred in Chapter 5. Finally, Chapter 6 examines the divergence trends of major Diptera clades, as well as the impact of various biotic and abiotic factors on Fly evolution. The main constraint of this study was the restricted and unequal taxon sampling; we expected that adequately increasing the taxon sets would yield more conclusive results.

Publications			
2022	**Research article**	**Kabiraj, D.**, Chetia, H., Nath, A., Sharma, P., Mosahari, P. V., Singh, D., ... & Bora, U. (2022). Mitogenome-wise codon usage pattern from comparative analysis of the first mitogenome of *Blepharipa sp.* (Muga uzifly) with other Oestroid flies. Scientific Reports, 12(1), 1-33.	**From Thesis work**
2019	**Book Chapter**	Chetia H, **Kabiraj D**, Bharali B, Ojha S, Barkataki MP, Saikia D, Singh T, Mosahari PV, Sharma P, Bora U. 2019. Exploring the benefits of endophytic fungi via omics, p. 51–81. In. Springer, Cham.	
2018	**Book Chapter**	Ojha S, Singh D, Sett A, Chetia H, **Kabiraj D**, Bora U. 2018. Nanotechnology in Crop Protection. Nanomater Plants, Algae, Microorg 345–391.	
2018	**Book Chapter**	Mosahari PV, Singh D, Kalita JJ, Sharma P, Chetia H, **Kabiraj D**, Mahanta C, Bora U. Nanotoxicity: Impact on Health and Environment. Environmental Toxicity of Nanomaterials. 2018 Apr 17:21.	
2017	**Research article**	Chetia H, **Kabiraj D**, Singh D, Mosahari PV, Das S, Sharma P, Neog K, Sharma S, Jayaprakash P, Bora U. 2017. De novo	

		transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). Gene 611.	Other Collabo-rative work
2017	**Research article**	Singh D, **Kabiraj D**, Sharma P, Chetia H, Mosahari PV, Neog K, Bora U. 2017. The mitochondrial genome of muga silkworm (Antheraea assamensis) and its comparative analysis with other lepidopteran insects. PLoS One 12.	
2017	**Research article**	Chetia H, **Kabiraj D**, Sharma S, Bora U. 2017. Comparative insights to the transportome of Nosema: a genus of parasitic microsporidians. bioRxiv 110809.	
2016	**Research article**	Chattopadhyay E, De Sarkar N, Singh R, Ray A, Roy R, Paul RR, Pal M, Ghose S, Ghosh S, **Kabiraj D**, Banerjee R. Genome-wide mitochondrial DNA sequence variations and lower expression of OXPHOS genes predict mitochondrial dysfunction in oral cancer tissue. Tumor Biology. 2016 Sep;37(9):11861-71.	
2016	**Review article**	Singh D, Chetia H, **Kabiraj D**, Sharma S, Kumar A, Sharma P, Deka M, Bora U. 2016. A comprehensive view of the web-resources related to sericulture. Database 2016:baw086.	
2015	**Review article**	**Kabiraj D**, Kalita J, Chetia H, Singh D, Bora U. 2015. Expanding the frontiers of rice research through omics. Assam Sc. Soc. Vo. p. 1-28.	
2015	**Research article**	Kumar A, Chetia H, Sharma S, **Kabiraj D**, Talukdar NC, Bora U. 2015. Curcumin Resource Database. Database 2015:bav070.	

Articles under preparation/ communication from thesis	
1	Kabiraj D and Bora U, **Exploring aberrant signals in phylogeny reconstruction and testing reticulation events in Diptera flies.**
2	Kabiraj D and Bora U, **Dipteran phylogeny and molecular dating using large genomic dataset (Orthologous nuclear genes) and comparative analysis.**
3	Kabiraj D and Bora U, **Classification of Diptera based on different trophic pattern, and the tempo and mode of their diversification.**
4	Kabiraj D and Bora U, **Comparison of different molecular dating methods and perspective on speciation-extinction of Diptera from molecular and paleo-biological data.**

List of Figures

Figure 2.6: A) AT rich control region Alignment of *Blepharipa sp.* with other two Tachinidae species. **B)** Three alignments of the common overlap region between *trnW-trnC*, *atp8-atp6* and *nad4-nad4l*. **C)** Three alignment of the consensus gap region between *trnS2-nad1* (TACTAAAHHHHAWWMH), *trnE-trnF* (ACTAAHWWWAATTMHHWA), *nad5-trnH* (WGAYADATWYTTCAY) genes of all 36 Oestroidea mitogenome (Where, W= A/T, H=A/T/C, Y=T/C, D=G/T/A, M=A/C).

Figure 2.7: Usage of start and stop codons in complete Oestroidea mitogenomes. **(A)** Start codons usage of 13 PCGs in Oestroidea. **(B)** Stop codons usage of 13 PCGs in Oestroidea.

Figure 2.8: (A) Trend of AT skew across the Oestroidea superfamily and outgroups. **(B)** AT skew vs GC skew of different genetic position of 44 organisms (CR shows maximum variation and 1st codon position shows least variation).

Figure 2.9: (A) RSCU Cluster analysis of 36 species from Oestroidea Superfamily, 6 organisms from other Diptera and 2 organisms from out group (Lepidoptera). Termination codons are excluded. The heat-map was drawn with CIMminer. Bigger RSCU values, suggesting more frequent codon usage, are represented with darker shades of red. (B) Family wise Average RSCU value plot of 62 codons.

Figure 2.10: (A) RSCU value comparison between E. flavipalpis (Maximum A/U at 3rd codon position), G. intestinalis (Minimum A/U at 3rd codon position) and *Blepharipa sp.* **(B)** Average Effective codon number (ENc) of 13 PCGs of different families of Oestroidea flies and out groups.

Figure 2.11: (A) The ENc vs. GC3s plots of Oestroidean mitochondrial protein coding genes. The standard curve ENc = 2+GC3s +29/ [GC3s2+(1-GC3s)2] represents the expected ENc to GC3s. **(B)** Neutrality plots (GC12 vs. GC3) of 13 PCGs of 42 species. GC12 stands for the average value of GC content in the first and second position of the codons (GC1 and GC2).

While GC3 refers to the GC content in the third codon position (each dot signifying a gene). **(C)** Probability of selection pressure on each PCGs of Oestroidea. The regression line of all PCGs denoted by y= mx+c (Where, Mutational Pressure (M) =m*100, Natural Selection (N) =100-M).

Figure 2.12: (A) Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 44 organisms (36: Oestridea superfamily, 6: other Diptera and 2: Out group Lepidoptera) using maximum likelihood (ML) method in RaxML 8.2.x (5000 bootstrap replicates). **(B)** Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 44 organisms (36: Oestridea superfamily, 6: other Diptera and 2: Out group Lepidoptera using Bayesian inference (BI) method in MrBayes v3.2.6.

Figure 2.13: Univariate regression model fitting between response variables, divergence rate (dS, dN and ω) and predictor variables, codon usage indices (GC3, GC3s, GC12, ENc) of training datasets; **(A-D)**: average synonymous divergence (adS) rate vs codon usage indices; **(E-H)**: average nonsynonymous divergence rate (adN) vs codon usage indices; **(I-L):** omega ratio (ω) vs codon usage indices; Green: Linear Model (LM), Blue: Polynomial Model (PM), Red: Generalized Additive Model (GAM), light colour represent the 95% confidence interval; R-sq (R^2): Coefficient of determination; AIC: Akaike information criterion; deg: degree, edf: effective degrees of freedom.

Figure 2.14: Univariate regression model fitting between logarithmic response variables, divergence rate (dS, dN and ω) and predictor variables, codon usage indices (GC3, GC3s, GC12, ENc) of training datasets; **(A-D):** log of average synonymous divergence (adS) rate vs codon usage indices; **(E-H):** log of average nonsynonymous divergence rate (adN) vs codon usage indices; **(I-L):** log of omega ratio (ω) vs codon usage indices; Green: Linear Model (LM), Blue: Polynomial Model (PM), Red: Generalized Additive Model (GAM), light colour

represent the 95% confidence interval; R-sq (R^2): Coefficient of determination; AIC: Akaike information criterion; deg: degree, edf: effective degrees of freedom.

Figure 2.15: (A) RSCU value of AU ending codons and ENc of 13 concatenated PCGs. Contour map phylogeny shows the estimated evolutionary history of codon usage, and corresponding variation of ENc produced via contMap function in the R package Phytools. Note that the Tachinidae clade has evolved a AT content that is higher than the rest of the ingroup that is reflected in ENc. **(B)** and **(C)** Principal components analysis of RSCU across the Oestroidea. The Tachinidae groups are distinguishable from rest of the Oestroidea insects.

Figure 3.1: Basic principles behind the D and f-branch statistics. **(A)** Example genealogies illustrating the sharing of derived alleles, labeled as 'B' across populations P2 and P3 (the ABBA pattern) and P1 and P3 (the BABA pattern) as a result of incomplete lineage sorting. Both patterns are thought to be equally possible in the absence of gene flow **(B)** Gene flow between P2 and P3 provides additional loci with ABBA patterns, which would resulting in a positive D statistic. **(C)** A basic phylogeny for explaining f4-ratio estimation. **(D)** Interdependences between distinct f4-ratio scores are illustrated in this example, which can provide information on the time of introgression. Different choices for the P1 population in this case offer constraints on when gene flow may have occurred. **(E)** The f-branch, or fb statistic, distinguishes between admixture at different time periods by allocating signals to distinct (potentially internal) branches in the population/species tree, based on relations between the f4-ratio results from different four taxon tests. (This figure is followed from Malinsky, M. et. al 2021).

Figure 3.2: The Likelihood mapping of sequence alignment of mitochondrial protein coding genes, A) Raw Alignment, B) Gblock Alignment. The concentration of points at the ends of the triangles on the top, as well as the sum of the percentages (>90%) at the ends of the triangles

on the bottom-right, indicate the good quality of the phylogenetic signal of the investigated genes.

Figure 3.3: Complete phylogeny of Diptera flies using Maximum Likelihood method implemented in RaxML v. 8. Green branches are showing maximum bootstrap support (100). Red branches are minimum bootstrap support. Representative images of flies are taken from internet using labeled for reuse with modification.

Figure 3.4 i: (A) PI of MrBayes unconstrained tree for different codon positions. **(B)** PI of MrBayes constrained tree for different codon positions. We restrict us to continue further MrBayes analysis with GB dataset as the MrBayes analysis is computationally very expensive and unable to provide desirable result. (200 million generations took almost 19 days).

Figure 3.4 ii: PI profile of RaxML constraint tree **(A)** WG dataset **(B)** GB dataset.

Figure 3.4 iii: PI profile of RaxML unconstraint tree **(A)** WG dataset **(B)** GB dataset.

Figure 3.4 iv: PI profile of IQedge proportional tree **(A)** WG dataset **(B)** GB dataset.

Figure 3.4 v: PI profile of IQedge unlinked tree **(A)** WG dataset **(B)** GB dataset.

Figure 3.5: Nucleotide contents percentage of each species for the entire set of genes, and first, second, or third codon positions only.

Figure 3.6: Heat map of RSCU values in Diptera mitogenome. The heat-map was drawn with CIMminer, using the quantile binning method. Bigger RSCU values, suggesting more frequent codon usage, are represented with brighter shades of red.

Figure 3.7: Substitution saturation plot of Transition and Transversion vs F84 and GTR distance of two datasets (WG and GB). Substitution saturation define by slope of the regression line, lower the slope higher the substitution saturation.

Number in red colour associated with node denote < 100% bs; **(B)** The concatenation-based species tree inferred by PhyloBayes (Bayesian inference). Colour of the branches indicate the posterior probability (pp); green: maximum pp; **(C)** The coalescent-based species tree was inferred by ASTRAL. Colour of the branches indicate the posterior probability (pp); green: maximum pp, red: minimum pp. Number in red colour associated with node denote < 1 pp.

Figure 4.6: Summary of conflicting and concordant homologs. The species trees are generated by RaxML**(A)**, PhyloBayes **(B)**, and ASTRAL **(C)**. For each branch, the top number indicates the number of homologs concordant with the species tree at that node, and the bottom number indicates the number of homologs in conflict with that clade in the species tree. The pie charts at each node present the proportion of homologs that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion that inform (conflict or support) this clade that have less than 50% bootstrap support (grey).

Figure 4.7: Representation of Internode Certainty All (ICA) score of the nodes in the species trees. The species trees are generated by RaxML**(A)**, PhyloBayes **(B)**, and ASTRAL **(C)**. The numbers associated with each node is the ICA score and colour of numbers denotes the strength of internode certainty; green: high ICA (strong certainty) and red: low ICA (high conflict).

Figure 4.8: The substitution rate ω (dN/dS) (Top) and corresponding kappa (Ts/Tv) (Bottom) variation on different Phylogenetic trees using M0 model (one rate for a tree).

Figure 4.9: Scatter plot of ω and kappa values of 335 genes based on seven phylogenetic trees

Figure 4.10: Chronogram Depicting the Evolutionary Timescale of 52 Diptera and Two Outgroups (Lepidoptera and Coleoptera) based on RaxML species tree. **(A)** Divergence times were estimated using Bayesian inference of 335 genes with 1 calibration point in MCMCTree.

List of Tables

List of Abbreviations

OXPHOS: Oxidative phosphorylation

ETS: Electron transport system

PCG: Protein-coding genes

rRNA: Ribosomal Ribonucleic acid

tRNA: Transfer Ribonucleic acid

mtDNA, mt-genome: Mitochondrial Deoxyribonucleic acid., mitogenome

NCBI: National Center for Biotechnology Information

WMG: Whole Mitochondrial Genome

AT: Adenosine and Tyrosine combine

GC: Guanine and Cytosine combine

GC3: GC at 3rd codon position

ATP: Adenosine triphosphate

ADP: Adenosine diphosphate

atp6: ATP synthase membrane subunit 6

atp8: ATP synthase membrane subunit 8

cox1: Cytochrome c oxidase subunit I

cox2: Cytochrome c oxidase subunit II

cox3: Cytochrome c oxidase subunit III

cytb: Cytochrome b

NADH: Nicotinamide adenine dinucleotide

nad1: NADH-ubiquinone oxidoreductase chain 1

nad2: NADH-ubiquinone oxidoreductase chain 2

nad3: NADH-ubiquinone oxidoreductase chain 3

nad4: NADH-ubiquinone oxidoreductase chain 4

nad4l: NADH-ubiquinone oxidoreductase chain 4L

nad5: NADH-ubiquinone oxidoreductase chain 5

nad6: NADH-ubiquinone oxidoreductase chain 6

S-shape curve: Sigmoid curve

CR: Control Region

CTAB: Cetyl Trimethyl Ammonium Bromide

CMER&TI: Central Muga Eri Research and Training Institute

SSPACE: SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension

BLAST: Basic Local Alignment Search Tool

CAP3: Contig Assembly Program phase 3

PCR: Polymerase Chain Reaction

nr database: non-redundant database

NGS: Next-generation sequencing

ORF: Open Reading Frame

MEGA: Molecular Evolutionary Genetics Analysis

SRA: Sequence Read Archive

Mfold: Multiple fold

MCMC: Markov chain Monte Carlo

LRT: likelihood ratio test

MAFFT: Multiple Alignment using Fast Fourier Transform

PP: Posterior Probabilities

ESS: Effective sample size

PSRF: Potential Scale Reduction Factor

RaxML: Randomized Axelerated Maximum Likelihood

iTOL: Interactive Tree Of Life

dS: Synonymous substitutions rate

dN: Non-synonymous substitutions rate

ω: dN/dS, ratio of nonsynonymous to synonymous substitutions rate

Ts: Transition, interchanges of two-ring purines (A G), or of one-ring pyrimidines (C T)

Tv: Transversion, interchanges of purine for pyrimidine bases or vice versa

κ: Transition/ Transversion

PAML: Phylogenetic Analysis by Maximum Likelihood

OL: Over-Lapping region

IGS: Inter-Genic Spacer

RSCU: Relative Synonymous Codon Usage

DAMBE: Data Analysis in Molecular Biology and Evolution

ENc: Effective Number of codons

CUB: Codon Usage Bias

CAI: Codon adaptation index

tAI: translation adaptation index

fop: frequency of optimal codons

LM: Linear Model

PM: Polynomial Model

GAM: Generalized Additive Model

AIC: Akaike information criterion

BIC: Bayesian information criterion

pBIC: phylogenetic Bayesian information criterion

LBA: Long-branch attraction

MPL: Maximum Pseudo-Likelihood

ML: Maximum Likelihood

MP: Maximum Parsimony

GTR: general-time-reversible model

F84: Felsenstein 84 model

HKY: Hasegawa, Kishino, and Yano model

GHOST: General Heterogeneous evolution On a Single Topology

IC: Internode Certainty

TC: Tree Certainty

ICA: Internode Certainty All

HGT: Horizontal Gene Transfer

ILS: Incomplete Lineage Sorting

GB dataset: GBlock trimmed dataset

WG dataset: Without GBlock trimmed dataset

PI: Phylogenetic informativeness

Iss: Index of substitution saturation

MDC: Minimizing Deep Coalescence

I_D: Disparity index

QS: Quartet Sampling

AU-test: Approximately Unbiased test

EPE: End-Permian Extinction

BUSCO: Benchmarking Universal Single-Copy Orthologs

OG: Orthologus gene

ANI: Average nucleotide identity

RF distance: Robinson-Foulds distance

MDS: Multidimensional Scaling

ARD: All rates different

ER: Equal rates

BiSSE: Binary State Speciation and Extinction

MuSSE: Multi State Speciation and Extinction

HiSSE: Hidden State Speciation and Extinction

DI: Disparity index

MRCA: Most Common Recent Ancestor

LTT: Lineage-Through-Time

MCC: Maximum clade credibility

MCCR: Monte Carlo constant-rates

RTT: Rate-Through-Time

QuaSSE: Quantitative State Speciation and Extinction

CID: Character independent

DTT: Disparity-through-time

BM: Brownian motion

OU: Ornstein-Uhlenbeck

EB: Early-Burst

LASSO: least absolute shrinkage and selection operator

RQ: Red Queen model, based on biotic factors

CJ: Court Jester model, based on abiotic factors

BEAST: Bayesian Evolutionary Analysis Sampling Trees

BAMM: Bayesian Analysis of Macroevolutionary Mixtures

CoMET: Compound Poisson process on Mass-Extinction Times

DDD: Diversity-Dependent Diversification

CHAPTER 1

Introduction and Review of Literature

> "So important are insects and other land-dwelling arthropods that if all were to disappear, humanity probably could not last more than a few months."
>
> — Edward Osborne Wilson

TH-2821_146106003

CHAPTER 1

Introduction and Review of Literature

1.1 Successful life of insects:

Biologists often refer that the "success" of a group of organisms, which typically means one of two things: **evolutionary success**—measured in terms of species diversity, geological duration, and/or geographic spread—and **ecological success**, as measured in terms of the impacts of a species or group of species upon an ecosystem[1]. If this the definition of successful life of any species, then insects are truly living their successful life since Early Ordovician (480 Mya (Million years ago)) period having most speciose class present on earth[2]. Insects comprise ~84% of the entire animal life having more than 1 million documented species. Insects are classified into three groups based on their metamorphosis style: Ametabola (no metamorphosis), Hemimetabola (partial metamorphosis), and Holometabola (Complete metamorphosis). Holometabola has three unique life stages: a) larvae, caterpillar, grub, or maggot (feeding stage), b) pupa or chrysalis (resting stage), and c) flying adult insect (reproductive stage)[3]. A major tissue restructuring occurs in the 'resting' stage between the feeding (larvae) and reproductive stages (adult insect), resulting in larvae that vary markedly from adults in terms of morphology, and function[3,4]. Very high diversities (100,000 or more species) are achieved in four orders of holometabolans, which includes Coleoptera (Beetle; 400K), Lepidoptera (Butterfly and Moth; 180K), Hymenoptera (Sawfly, Wasp, Bee, and Ant; 150K), and Diptera (True fly; 150K), account for ~87% of all insect species[3]. The 'reset button' effect of metamorphosis is regarded to be primarily responsible for holometabolous insects' success, since it permits larvae and adults to specialize more

independently for various activities, like as growth vs reproduction, achieving better efficiency in each[5,6].

1.2 Diptera, the true flies:

True flies belong to the Diptera order of insects, which gets its name from the Greek di- "two" and pteron "wing" because they only have one pair of functional wings, with the hind wings reduced to a set of stalked knobs termed halters that assist them balance during flying[7]. The most of flies have sponging mouthparts for sucking liquids, different from the mandible-like chewing mouthparts in other insects [7,8]. Flies are divided into three groups, each with its own distinct body form and antennal structure namely, A) The Nematocera, the earliest group, has a thin body and long slender antennae with 16 apparent segments; certain males have long dense hairs on these segments; e.g. mosquitoes (Culicidae), midges (Chironomidae), crane flies (Tipulidae), and blackflies (Simuliidae); B) The Lower Brachycera have antennas that are shorter and stouter, with three to ten visible segments, e.g. horse and deer flies (Tabanidae); C) The higher Brachycera have a short, wide body and three broad segments on their antennae, e.g. blow flies (Calliphoridae), fruit flies (Drosophilidae) and muscid flies (Muscidae)[7].

Diptera's ability to live on a variety of food sources has allowed them to flourish in a range of food chains, and climatic niches[7,9,10]. Several studies have described the various lifestyles and behaviours of dipterans as larvae and adults illustrating many roles that flies play in the planet's ecological interactions—large numbers of flies graze on plants, manage pest arthropods, decompose decaying organic and faecal matter, pollinate blossoms, provide nourishment for other species, and even transmit disease[7,10]. Despite their diverse dietary behaviours (scavengers, predators, parasites, parasitoids, and herbivores), larvae usually found in moist or wet habitats (e.g., within tissues or living plants, amid decaying organic materials, inside other animals, or in association with truly "aquatic" habitats). Invasion of these and other habitats is related to larval adaptations[10].

1.3 Diptera diversity:

Diptera is one of the most speciose the insect order (>152,000 species), anatomically diverse, and environmentally adaptable groups of species, accounting for more than 10% of all known animal kingdom species [9,11,12]. The extant dipteran species have been categorized into 2 suborders, 13 infraorders, 22–32 superfamilies, 157 families, and 10,000 genera [9,12–14].

With 40 families and about 55,000 species, the Nematocera or lower Diptera are an ecologically and morphologically diverse assemblage of true flies, accounting for around one-third of the order's extant variety[9,15,16]. The Nematocera is comprised of eight infraorders namely Culicomorpha (mosquitoes, midges etc.), Tipulomorpha (crane flies), Psychodomorpha (sand flies), Bibionomorpha (march flies and gall midges), Deuterophlebiomorpha, Nymphomyiomorpha, Ptychopteromorpha, and Perissommatomorpha[9,16].

The higher Diptera or Brachycera consist of more than 100,000 species and categorized within 117 families, and four major groups namely Eremoneura, Cyclorrhapha, Schizophora, and Calyptratae nested within Brachycera[9]. The Cyclorrhapha, a Brachyceran clade, is home to more than half of all true flies; its key innovations include extreme shrinking of the larval head capsule and pupation of the third instar in the final larval skin (puparium)[8,14]. The Cyclorrhaphan group, Schizophora has the most family-level variety, with 85 described families (>50,000 species), and flies of this group emerge from the puparium by inflating a membrane head sac (ptilinum)[9,13,14]. Acalyptratae and Calyptratae are the two major groups of Schizophora. Acalyptratae contains around 20% of the fly species yet accounts for nearly 40% of the order's family-level diversity (62 families). Surprisingly, six common acalyptrate families (Drosophilidae, Tephritidae, Agromyzidae, Chloropidae, Lauxaniidae, and Ephydridae) account for more than half of the Acalyptratae species diversity[8]. Calyptrate flies are classified into 13 families, some of which are relevant forensically (e.g., Calliphoridae and

Sarcophagidae), medically (e.g., Glossinidae, Muscidae, and Oestridae), or as biological control agents (e.g., Tachinidae)[8].

1.4 Phylogeny and Evolution of Diptera:

Historically, establishing relationships between distinct Dipteran lineages was dependent on an understanding of morphological characteristics. In recent decades, sophisticated and consistent methodologies, such as the incorporation of larger amounts of molecular sequence data and the introduction of a considerable number of new and well-preserved fossils, have been utilized in Diptera higher phylogenetic studies, yielding more definitive results[9,15,17–19].

Nematocera – the earliest fly lineage: Diptera's monophyly is widely established, as evidenced by a variety of complex morphological innovations known as synapomorphies, such as the transformation of hindwings into halteres and the modification of mouthpart for sponging liquids[14,20]. A number of shared-derived characters support the Brachycera's monophyly, yet the Nematocera is largely recognized as a paraphyletic assemblage of infraorders from which the Brachycera evolved[13,15,21]. Morphology-based theories contradict the composition and interrelationships of the nematoceran infraorders, making the Nematocera's evolutionary relationships extremely difficult to discern[13,21]. Several phylogenetic studies based on molecular and morphological characters contradict each other in certain ways about the composition and interrelations of the nematoceran infraorders, which we attempted to cover here.

Many morphological traits in the Tipulomorpha superfamily imply a sister-group relationship between Tipuloidea and Trichoceridae, which is also corroborated by molecular dataset analysis[9,21]. The Culicomorpha is a large clade that includes the monophyletic Culicoidea (Corethrellidae, Culicidae, Chaoboridae, and Dixidae) and the sister-groups Simuliidae and Thaumaleidae, which are supported by both morphological and nucleotide-based

researches[9,15,22]. Psychodomorpha is made up of three families: Blephariceridae, Psychodidae, and Tanyderidae, all of which have freshwater larvae and several notions of "Psychodomorpha" have been offered in earlier research that recovered different affinities for all three of these families [21,23]. The researchers confirmed the relationship between these three families using nucleotide data from 28S rDNA and numerous nuclear genes, with Blephariceridae being the sister to the other two families, which is currently recognized as the strongest explanation for the location of these groups[9,15]. Bibionomorpha is a broad clade comprising 17 families, and its relationships are unclear. A new phylogenetic study based on nuclear and mitochondrial ribosomal gene segments by Ševčík et al. (2014) contradicts the phylogenetic ordering of Wiegmann et al. (2011), however support for links among the main bibionomorph clades is minimal in both studies[9,24]. However, nearly all phylogenetic analyses of Bibionomorpha support monophyly of the Sciaroidea[9,24,25].

It has been especially challenging to define the sister-group of the immensely diversified clade Brachycera. Earlier, Hennig (1968) offered evidence for a link between the Brachycera and the Bibionomorpha based on adult characteristics[26]. Further, Michelsen (1996) reported a sister-group relationship between the Brachycera and a revised Bibionomorpha based on morphological traits such as mature thoracic sclerites and musculature[27]. Recent studies of molecular and morphological datasets have further reinforced this Neodiptera group [9,15,28].

Lower Brachycera - a phylogenetic riddle: Establishing evolutionary links among the families and higher-level groupings that make up the Brachycera's early lineages, regardless of data or analytic techniques, still one of the most difficult aspects of fly phylogeny[29]. Lower Brachycera (= "Orthorrhapha") flies are predators or parasitoids as larvae and are typically large in size. They are grouped into three infraorders based on morphology: Xylophagomorpha, Tabanomorpha, and Stratiomyomorpha, as well as a number of superfamilies (Asiloidea, Empidoidea, Nemestrinoidea)[30]. Lower Brachycera also comprises the clades namely,

Nemestrinoidea (Acroceridae and Nemestrinidae), Bombyliidae, and Asiloidea (Apioceridae, Apsilocephalidae Asilidae, Evocoidae, Mydidae, Scenopinidae, and Therevidae). Over the last three decades, extensive comparative study based on morphological and molecular datasets on the relationships among and within these groupings has been undertaken, but most studies were unable to yield well-supported resolution for the higher-level relationships. Various phylogenetic studies on Therevidae, Bombyliidae, Asilidae, and Asiloidea, have offered new theories regarding the relationships of most of the families, although there is still substantial uncertainty[31–34]. A phylogenetic investigation of the lower Brachycera utilizing only nuclear 28S ribosomal DNA indicated paraphyly[35], an inference echoed by all morphological analyses of these species conducted over the last half century[30,36]. In contrast, Wiegmann et al. (2011) found substantial bootstrap support for Orthorrhapha monophyly in their more comprehensively sampled multigenic study, though it remains to be seen whether this finding will be supported further in larger dataset analyses[9].

The Empidoidea, or dance flies, and associated families (Atelestidae, Brachystomatidae, Dolichopodidae, Empididae, Homalocnemidae, Hybotidae, and Oreogetonidae), are a well-supported monophyletic group that has become the subject of extensive systematic morphological and molecular research[37–39].

Higher Brachycera and radiation of Schizophora: Empidoidea, together with Cyclorrhapha ("upper flies"), forms the monophyletic clade Eremoneura[9]. A relict lineage, Apystomyiidae consists of a single species, *Apystomyia elinguis*, has been well-supported in numerous molecular analyses as the sister-group to all higher flies[9,33]. Apystomyia is a morphologically unusual mixture of asiloid-, empidoid-, and cyclorrhaphan-like characteristics that makes classification challenging only using one or a few criteria. Although, a Quantitative analysis have rejected plausible alternatives of being located at the base of the Asiloidea or Empidoidea, indicating strong support for Apystomyia + Cyclorrhapha[28].

The primary Cyclorrhapha (= "Aschiza") first-branching lineages are divided into two superfamilies: Phoroidea and Syrphoidea (Syrphidae). Previously, the Pipunculidae (Big-headed flies) were regarded a sister group to the Syrphidae and categorized as a Syrphoidea[40]. Molecular evidence, on the other hand, consistently links the pipunculids to Schizophora, the order's next great monophyletic radiation[9,41].

Schizophora are subdivided into the monophyletic Calyptratae[13,14], with the remaining species most likely representing the paraphyletic acalyptrate grade[9,14,42]. The Calyptratae are sister to various subgroups of the acalyptrates, also supported by the majority of contemporary research[9,43,44]. Most scholars acknowledge at most ten well-defined superfamilies of putatively closely related families among acalyptrates[13]: Carnoidea, Conopoidea, Diopsoidea, Ephydroidea, Lauxanioidea, Nerioidea, Opomyzoidea, Sciomyzoidea, Sphaeroceroidea, and Tephritoidea. Except for the Ephydroidea, Lauxanioidea, Nerioidea, and Tephritoidea, virtually all super-families' relationships, classifications, and presence are not convincing either morphological synapomorphies or molecular phylogenetic evidence[9]. Whereas, several molecular phylogenetic analyses demonstrate Tephritoidea and Ephydroidea to be monophyletic clade[45], but accurately establishing the relationships within and among the rapid radiation of acalyptrate remains one of systematic entomology's most difficult tasks.

Resolving calyptrate phylogeny using morphological or small molecular data sets has proven to be difficult[46,47]. The majority of studies favor a Glossinidae and Hippoboscidae that branch earlier, a paraphyletic muscoid grade (Anthomyiidae, Fanniidae, Muscidae, and Scathophagidae), and a monophyletic Oestroidea (Calliphoridae, Mesembrinellidae, Mystacinobiidae, Oestridae, Rhiniidae Rhinophoridae, Sarcophagidae and Tachinidae)[9,19,43,46]. With a basal Fanniidae, a next-branching Muscidae, and a combined Anthomyiidae and Scathophagidae sister to the Oestroidea, the muscoid grade maintains a stable resolution. High diversity, short branch lengths, contradicting morphological evidence, and limited branch

support continue to complicate relationships within the Oestroidea. The conventional Calliphoridae family is no longer monophyletic[48,49], and two previous subfamilies, Rhiniidae and Mesembrinellidae, are increasingly recognized as complete family[48,50,51].

1.5 Tachinidae family:

The Tachinidae family is the most abundant group of Diptera flies, with almost 10,000 tremendously diverse species accounting for around 6.5% of entire Diptera diversity[52]. There are presently four recognized subfamilies within this family (Dexiinae, Exoristinae, Phasiinae, Tachininae)[52]. Tachinds flies are known for koinobiont, endoparasitism and their peculiar respiration inside wide range of arthropod hosts including caterpillars (Lepidoptera), true bugs (Hemiptera), adult and larval beetles (Coleoptera) centipedes (Chilopoda), and spiders (Arachnida) which is 80% of other insects[53,54]. Tachinid species mostly infest a wide range of phytophagous insects, with roughly 60% of them parasitoids of ditrysian Lepidoptera [52,55].

Tachinids' living strategy: Tachinids consume haemolymph and non-essential tissues and organs before moving on to crucial portions of their hosts; in most cases, tachinids kill their hosts before achieving larval development[53,56–58]. The majority of tachinids emerge from their hosts' pupal stage; no species has been known to attack pupae or their hosts' egg stage and most of tachinid species attack larval hosts, although a small proportion, possibly 5% to 10% of species, attack adults[55].

Tachinids, unlike parasitic Hymenoptera, lack a primary piercing ovipositor, therefore they lay their eggs on or near the host, and the newly hatched larvae must enter the host[55]. Because there is no ovipositor, paralytic poisons, mutualistic polyDNA viruses, and other accessory substances that paralyse the host and/or its immune system are not injected. For this reason, tachinids are classified as koinobiont parasitoids[59], meaning that they allow their host to continue feeding and developing as they grow inside of it rather than halting it (as do

idiobionts)[55]. Tachinid larvae, do not evade or destroy host hematocytes (as do hymenopteran parasitoids); instead, they are widely known for the formation of respiratory funnels derived from host defensive cells[55,60]. Most tachinids utilize this structure to keep their posterior spiracles connected to the host's external integument or major tracheal branches, allowing them to maintain a close connection with ambient air[55,61]. This capability to use the immunological response to form respiratory funnels allows tachinids to more easily explore new hosts, resulting in the dynamic evolution and variation of hosts[55]. Additionally, tachinids are relatively resistant to toxins ingested by their hosts, empowering for greater evolutionary adaptability in diverse host range[62,63]. This resilience might be attributed to pre-adaptations associated with the Oestroidea's ancestral saprophagous habits, wherein larvae exposed to severely toxic surroundings generated by bacteria and fungus developed tolerance to these toxins[55,64]. The location of immature larvae inside the host may be linked to tachinids' resilience to host physiological defenses. Instead of floating freely in the hemocoel[65], many tachinid larvae anchor themselves in specific tissues, and at least one extremely polyphagous species, *Compsilura concinnata*, spends the entire larval stage in the gut[66]. Tachinids' finely tuned lifestyle gives them an edge over other oestroids, which are frequently sarcophagous, coprophagous, parasitic on vertebrates, or kleptoparasitic on Hymenoptera [52,67].

Uzi flies: Uzi flies are Tachinids, responsible for infestation and death of commercially important silkworms. Four species of uzi flies are identified till date viz., the Japanese uzi fly, *Crossocosmia sericaria* (Rodani); the Hime uzi fly, *Ctenophora pavida* (Meigen); the Tasar uzi fly, *Blepharipa zebina* (Walker) and the Indian uzi fly, *Exorista sorbillans* (Wiedemann)[68]. The last two dipteran endo-parasites have wreaked havoc on India's sericulture sector (mulberry, muga, and tasar), inflicting financial damage to rural seri-based farmers[68–70]. The currently studied uzifly species, *Blepharipa sp.* (Chapter 2), found in Assam and Meghalaya,

attacks muga silkworm (*Antherea assamensis*) larvae during winter and post-winter season and has been accounted for around 80-90% yield loss in muga seed cultivating areas[71–73].

1.6 Molecular data in phylogenetics of Diptera:

The age of technical and theoretical advances in the utilization of genetic diversity in DNA and protein sequences rejuvenated fly phylogenetics and enabled to acquire evidence for several species across the order at all levels of analysis[74]. A crucial part of assessing molecular data is selecting genes for sequencing that will lead to significant variations at reasonable rates and that can be efficiently amplified and sequenced using standard laboratory techniques for the majority of the study species. Recently, it has been shown that small datasets of only a few genes may contain insufficient information led to conflicting result for complicated taxon radiations[75]. Multi-gene datasets are thus become norms to address phylogenetic issues in nearly all of the major and common fly families, and these studies provide fresh evidence for family-level relationships in the context of divergences within and across major fly clades[9,19,43].

Mitochondrial genome: One of the most essential eukaryotic organelles, the mitochondrion (mt), is descended from α-proteobacterium and as such retains a bacterial-like genome[76,77]. Since the release of the first insect mt-genome of *Drosophila yakuba* by Clary and Wolstenholme in 1985, the number of sequenced insect mt genomes has rapidly grew, and mt-genomes of all insect orders are now available[76,78]. The Diptera (flies) are one of the most widely sequenced orders amongst the Insecta, with 500 complete, partial mitogenomes in GenBank (as of 17 Feb 2022, including duplicate species) (Fig. 1.1 A). The mitogenome has been extensively employed as an estimator for phylogenetic investigations, owing to three factors: (1) their high copy number and easily obtainable conserved primer sets make them convenient to produce[79]; (2) they contain sufficient phylogenetic information to enable inference over large taxonomic ranges[18,43]; and (3) In comparison to many genes in the nuclear genome, mitochondrial genes evolve at a faster rate, and mitochondrial genomes are inherited

maternally[80]. Mitogenome-based phylogenetic studies have contributed and supported the overall pattern of relationships corresponding to known lineages in the higher phylogeny of Diptera[19,43,81,82]. Although, the majority of Diptera mitochondrial phylogenomic research have been under-sampled, focusing on model organisms or comparisons between published genomes and a few new updated species. Some of the recent studies, for example, Zhao et al. (2013), have integrated large taxon samples from published mitogenomes with newly sequenced taxa, or have done vast mitochondrial genome sequencing implementing modern sequencing techniques by Junquiera et al. (2016) and Zhang et al. (2016) [19,43,83].

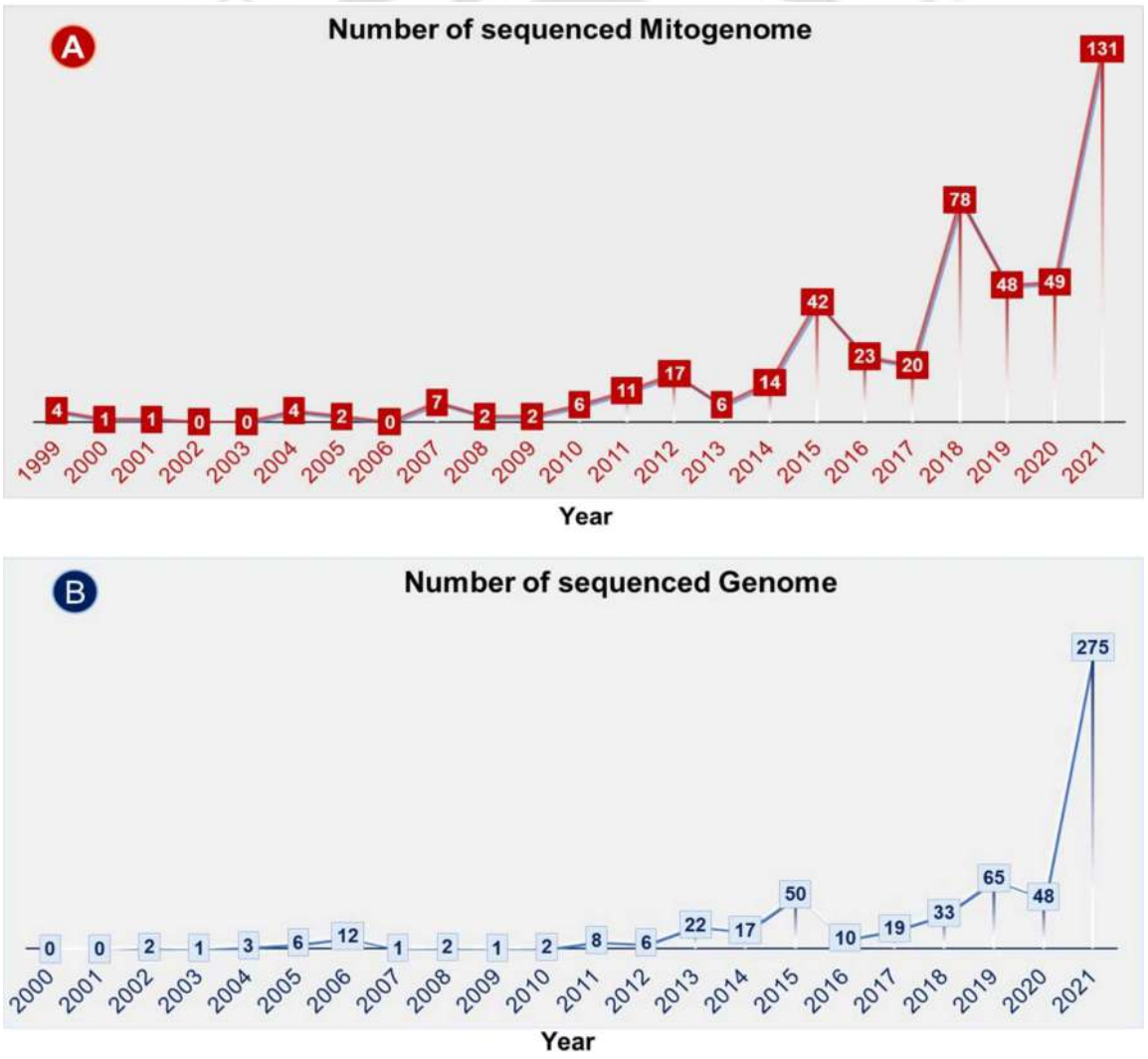


Figure 1.1: Trends of genome sequencing. (A) Mitochondrial genome sequencing of Diptera from 1999 to 2021. (B) Whole genome sequencing of Diptera from 2000 to 2021 (Data search from NCBI Genome browser).

Whole genome and transcriptome: Flies were among the first organisms (*Drosophila melanogaster*) to have their whole genome sequenced in the year 2000[84]. Since then, as an order Diptera include the largest number of insect species having complete or draft genomes that are publicly available[85] (Fig. 1.1 B). This number is increasing as better genomic sampling enables phylogenetic analyses to develop foundations for assessing genetic basis and evolution at the species and population levels[86,87].

The most contemporary and intriguing advance in fly phylogenetics is massive phylogenomic analysis of Diptera employing hundreds or even thousands of orthologous genes selected from the nuclear genome[29]. A novel approach, anchored hybrid enrichment, was first used to obtain vast amounts of genomic data for phylogenetic inference in Diptera, yielding a resolved and highly supported phylogeny of Syrphidae[41,88]. Transcriptome data can also be utilized for phylogenomic research since it is less expensive than whole genome sequencing and can provide a large number of molecular datasets. A massive phylogenomic analysis of over 1400 insects' transcriptome data for insect phylogeny reconstruction has revolutionized knowledge of fly relationships in terms of the entire insect phylogeny[89]. Without any doubt, as burgeoning amount of molecular data is generating by the advancement in sequencing technology, new resolution and ongoing challenges will be illuminated by a more completely resolved Fly Tree of Life.

1.7 Formulation of objectives:

Based on our literature review, the following objectives were formulated to gain a deeper understanding on the evolution of Diptera:

i) Mitochondrial genome sequencing of *Blepharipa sp.*, an endoparasite of Muga silkworm, and comparative codon usage analysis with other Oestroidea flies.

ii) Reconstruction of Diptera phylogeny with larger taxa.

iii) Reconstruction of Dipteran phylogeny and molecular dating using larger dataset (Orthologous genes) and comparative analysis.

iv) Classification of Diptera based on different dietary adaptation and the tempo and mode of their diversification.

v) Deciphering major diversification time through molecular dating and dynamics of speciation-extinction events of Diptera.

1.8 References:

1. Bradley, T. J. *et al.* Episodes in insect evolution. *Integr. Comp. Biol.* **49**, 590–606 (2009).
2. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–7 (2014).
3. Jarzembowski, E. A. Insects. in *Reference Module in Earth Systems and Environmental Sciences* (Elsevier, 2016). doi:10.1016/B978-0-12-409548-9.09735-9.
4. SEHNAL, F., ŠVÁCHA, P. & ZRZAVÝ, J. Evolution of Insect Metamorphosis. in *Metamorphosis* 3–58 (Academic Press, 1996). doi:10.1016/B978-012283245-1/50003-8.
5. Hammer, T. J. & Moran, N. A. Links between metamorphosis and symbiosis in holometabolous insects. *Philos. Trans. R. Soc. B* **374**, 20190068 (2019).
6. Moran, N. A. Adaptation and constraint in the complex life cycles of animals. *Annu. Rev. Ecol. Syst.* **25**, 573–600 (1994).
7. Skevington, J. H. & Dang, P. T. Exploring the diversity of flies (Diptera). *Biodiversity* **3**, 3–27 (2002).
8. Bertone, Matthew A., and B. M. W. *True fies (Diptera)*. *The timetree of life* (2009).
9. Wiegmann, B. M. *et al.* Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* **108**, 5690–5695 (2011).
10. Courtney, G. W., Sinclair, B. J. & Meier, R. Morphology and terminology of Diptera larvae. in (2000).
11. Groombridge, B. Status of the Earth's Living Resources. *Glob. Biodiversity. World Conserv. Monit. Center.* **1992**, (1992).

12. Thompson, F. C. Biosystematic Database of World Diptera. http://www.diptera.org/ (2004).

13. McAlpine, J. F. & Wood D. M. *Manual of nearctic Diptera. Vol 3* . http://www.barbau.ca/content/manual-nearctic-diptera-vol-3 (1989).

14. Yeates, D. K. & Wiegmann, B. M. Congruence and controversy: toward a higher-level phylogeny of Diptera. *Annu. Rev. Entomol.* **44**, 397–428 (1999).

15. Bertone, Matthew A., Gregory W. Courtney, and B. M. W. Phylogenetics and temporal diversification of the earliest true flies (Insecta: Diptera) based on multiple nuclear genes. *Syst. Entomol.* **33**, 668–687 (2008).

16. Courtney, G. W., Skevington, J. H. & Sinclair, B. J. Biodiversity of Diptera. in *Insect Biodiversity: Science and Society* (2017). doi:10.1002/9781118945568.ch9.

17. Gao, J., Watabe, H., Aotsuka, T., Pang, J. & Zhang, Y. Molecular phylogeny of the Drosophila obscura species group, with emphasis on the Old World species. *BMC Evol. Biol.* **7**, 87 (2007).

18. Nelson, L. A. *et al.* Beyond barcoding: A mitochondrial genomics approach to molecular phylogenetics and diagnostics of blowflies (Diptera: Calliphoridae). *Gene* **511**, 131–142 (2012).

19. Carolina, A. *et al.* Large-scale mitogenomics enables insights into Schizophora (Diptera) radiation and population diversity. *Sci. Rep.* **6**, 1–13 (2016).

20. Yeates, D. K. *et al.* Phylogeny and systematics of Diptera: Two decades of progress and prospects*. *Zootaxa* **1668**, 565–590–565–590 (2007).

21. Oosterbroek, P. & Courtney, G. W. Phylogeny of the nematocerous families of Diptera (Insecta). *Zool. J. Linn. Soc.* **115**, 267–311 (1995).

22. Borkent, A. The Pupae of Culicomorpha—Morphology and a New Phylogenetic Tree. *Zootaxa* **3396**, 1–98–1–98 (2012).

23. Amorim, D. de S. A new phylogeny and phylogenetic classification for the Canthyloscelidae (Diptera: Psychodomorpha). *Can. J. Zool.* **78**, 1067–1077 (2000).

24. Ševčík, J., Kaspřák, D., Mantič, M., Ševčíková, T. & Tóthová, A. Molecular phylogeny of the fungus gnat family Diadocidiidae and its position within the infraorder Bibionomorpha (Diptera). *Zool. Scr.* **43**, 370–378 (2014).

25. De Souza Amorim, D. & Rindal, E. Phylogeny of the Mycetophiliformia, with proposal of the subfamilies Heterotrichinae, Ohakuneinae, and Chiletrichinae for the Rangomaramidae (Diptera, Bibionomorpha). *Zootaxa* **1535**, 1–92–1–92 (2007).

26. Hennig, W. Kritische Bemerkungen über den Bau der Flügelwurzel bei den Dipteren und die Frage nach der Monophylie der Nematocera. *Stuttgarter Beiträge zur Naturkd. aus dem Staatl. Museum für Naturkd. Stuttgart* **193**, 1–23 (1968).

27. Michelsen, V. Neodiptera: New insights into the adult morphology and higher level phylogeny of Diptera (Insecta). *Zool. J. Linn. Soc.* **117**, 71–102 (1996).

28. Sinclair, B. J., Brooks, S. E. & Cumming, J. M. Male terminalia of Diptera (Insecta): a review of evolutionary trends, homology and phylogenetic implications. *Insect Syst. Evol.* **44**, 373–415 (2013).

29. Wiegmann, B. M. & David K. Yeates. Phylogeny of Diptera. Chapter 11. in *Manual of Afrotropical Diptera. Volume 1. Introductory Chapters and Keys to Diptera Families Suricata* (2017).

30. Yeates, D. K., Avid, D. & Eates, K. Y. Relationships of extant lower Brachycera (Diptera): a quantitative synthesis of morphological characters. *Zool. Scr.* **31**, 105–121 (2002).

31. Winterton, S. L. & Ware, J. L. Phylogeny, divergence times and biogeography of window flies (Scenopinidae) and the therevoid clade (Diptera: Asiloidea). *Syst. Entomol.* **40**, 491–519 (2015).

32. Trautwein, M. D., Wiegmann, B. M. & Yeates, D. K. Overcoming the effects of rogue taxa: Evolutionary relationships of the bee flies. *PLoS Curr.* **3**, (2011).

33. Trautwein, M. D., Wiegmann, B. M. & Yeates, D. K. A multi gene phylogeny of the fly superfamily

Asiloidea (Insec- ta): taxon sampling and additional genes reveal the sister- group to all higher flies (Cyclorrhapha). *Mol. Phylogenet. Evol.* **56**, 918–930 (2010).

34. Dikow, T. A phylogenetic hypothesis for Asilidae based on a total evidence analysis of morphological and DNA sequence data (Insecta: Diptera: Brachycera: Asiloidea). *Org. Divers. Evol.* **9**, 165–188 (2009).

35. Wiegmann, B. M., Yeates, D. K., Thorne, J. L. & Kishino, H. Time Flies, a New Molecular Time-Scale for Brachyceran Fly Evolution Without a Clock. *Syst. Biol* **52**, 745–756 (2003).

36. Hennig, W. *Handbuch der Zoologie: Diptera (Zweiflügler).. Beitr. 31* - . (1973).

37. Moulton, J. K. & Wiegmann, B. M. The phylogenetic relationships of flies in the superfamily Empidoidea (Insecta: Diptera). *Mol. Phylogenet. Evol.* **43**, 701–713 (2007).

38. Moulton, J. K. & Wiegmann, B. M. Evolution and phylogenetic utility of CAD (rudimentary) among Mesozoic-aged Eremoneuran Diptera (Insecta). *Mol. Phylogenet. Evol.* **31**, 363–378 (2004).

39. Sinclair, B. J. & Cumming, J. M. The morphology, higher-level phylogeny and classification of the Empidoidea (Diptera). *Zootaxa* **1180**, 1–172–1–172 (2006).

40. Rotheray, G. E. & Gilbert, F. Phylogenetic relationships and the larval head of the lower Cyclorrhapha (Diptera). *Zool. J. Linn. Soc.* **153**, 287–323 (2008).

41. Young, A. D. *et al.* Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC Evol. Biol.* **16**, 1–13 (2016).

42. Hennig, W. Neue Untersuchungen über die Familien der Diptera Schizophora (Diptera ... - Willi Hennig - Google Books. *Stuttgar? ter Beiträge zur Naturkd. aus dem Staatl. Museum für Naturkd. Stuttgart* **226**, 1–76 (1971).

43. Zhao, Z. *et al.* The Mitochondrial Genome of Elodia flavipalpis Aldrich (Diptera: Tachinidae) and the Evolutionary Timescale of Tachinid Flies. *PLoS One* **8**, 61814 (2013).

44. Vicoso, B. & Bachtrog, D. Reversal of an ancient sex chromosome to an autosome in Drosophila. *Nature* **499**, 332–335 (2013).

45. Han, H. Y. & Ro, K. E. Molecular phylogeny of the superfamily Tephritoidea (Insecta: Diptera): new evidence from the mitochondrial 12S, 16S, and COII genes. *Mol. Phylogenet. Evol.* **34**, 416–430 (2005).

46. Kutty, S. N., Pape, T., Wiegmann, B. M. & Meier, R. Molecular phylogeny of the Calyptratae (Diptera: Cyclorrhapha) with an emphasis on the superfamily Oestroidea and the position of Mystacinobiidae and McAlpine's fly. *Syst. Entomol.* **35**, 614–635 (2010).

47. Pape, T. Phylogeny of Oestridae (Insecta: Diptera). *Syst. Entomol.* **26**, 133–171 (2001).

48. Marinho, M. A. T. *et al.* Molecular phylogenetics of Oestroidea (Diptera: Calyptratae) with emphasis on Calliphoridae: Insights into the inter-familial relationships and additional evidence for paraphyly among blowflies. *Mol. Phylogenet. Evol.* **65**, 840–854 (2012).

49. Rognes, K. The Calliphoridae (Blowflies) (Diptera: Oestroidea) are Not a Monophyletic Group. *Cladistics* **13**, 27–66 (1997).

50. Pape, T., Blagoderov, V. & Mostovski, M. B. Order Diptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa* **3148**, 222–229–222–229 (2011).

51. Marinho, M. A. T., Wolff, M., Ramos-Pastrana, Y., de Azeredo-Espin, A. M. L. & Amorim, D. de S. The first phylogenetic study of Mesembrinellidae (Diptera: Oestroidea) based on molecular data: clades and congruence with morphological characters. *Cladistics* **33**, 134–152 (2016).

52. Cerretti, P. *et al.* Signal through the noise? Phylogeny of the Tachinidae (Diptera) as inferred from morphological evidence. *Syst. Entomol.* **39**, 335–353 (2014).

53. Dindo, M. L. Tachinid parasitoids: are they to be considered as koinobionts? *BioControl* **56**, 249–255 (2011).

54. Blaschke, J. D. Molecular Systematics of the Subfamily Phasiinae (Diptera: Tachinidae). *Masters Theses* (2013).

55. Stireman, J. O., O'Hara, J. E. & Wood, D. M. Tachinidae: Evolution, Behavior, and Ecology. *Annu. Rev. Entomol.* **51**, 525–555 (2006).

56. Herting, B. Biologie der westpaläarktischen Raupenfliegen Dipt. Tachinidae. in *Monographien zur angewandten Entomologie* vol. 16 1–188 (Parey, 1960).

57. Wood, D. M. Tachinidae. in *Manual of Nearctic Diptera,* (eds. J.F. McAlpine et al.) vol. 2 1193–1269 (Canadian Government Publishing Centre, Hull, 1987).

58. Mellini, E. Sinossi di biologia dei Ditteri Larvevoridi. in *Bollettino dell'Istituto di Entomologia della Universit`a degli Studi di Bologna* vol. 45 1–38 (1990).

59. Askew, R. & Shaw, M. Parasitoid communities: their size, structure and development. *Insect parasitoids* (1986).

60. Strand, M. R. & Pech, L. L. Immunological Basis for Compatibility in Parasitoid-Host Relationships. *Annu. Rev. Entomol.* **40**, 31–56 (1995).

61. Clausen, C. P. Entomophagous Insects. *Nature* **148**, (1941).

62. Gauld, I. D., Kevin J. Gaston & Daniel H. Janzen. Plant allelochemicals, tritrophic interactions and the anomalous diversity of tropical parasitoids: the" nasty" host hypothesis. *Oikos* 353–357 (1992).

63. Mallampalli, N., Barbosa, P. & Weinges, K. Effects of Condensed Tannins and Catalpol on Growth and Development of Compsilura concinnata (Diptera: Tachinidae) Reared in Gypsy Moth (Lepidoptera: Lymantriidae). *J. Entomol. Sci.* **31**, 289–300 (1996).

64. Eggleton, P. & Gaston, K. J. Tachinid host ranges: a reappraisal (Diptera: Tachinidae). *Entomol. s Gaz.* **43**, 139–143 (1992).

65. Belshaw, R. Life history characteristics of Tachinidae (Diptera) and their effect on polyphagy. *Parasit. Community Ecol.* 145–162 (1994).

66. Ichiki, R. & Shima, H. Immature Life of Compsilura concinnata (Meigen) (Diptera: Tachinidae). *Ann. Entomol. Soc. Am.* **96**, 161–167 (2003).

67. Feener, D. H. & Brown, B. V. Diptera as parasitoids. *Annu. Rev. Entomol.* **42**, 73–97 (1996).

68. Williams, K. A., Lamb, J. & Villet, M. H. Phylogenetic radiation of the greenbottle flies (Diptera, Calliphoridae, Luciliinae). *Zookeys* (2016) doi:10.3897/zookeys.568.6696.

69. Klong-klaew, T. *et al.* Observations on morphology of immature Lucilia porphyrina (Diptera: Calliphoridae), a fly species of forensic importance. *Parasitol. Res.* **111**, 1965–1975 (2012).

70. Chen, Y. *et al.* The complete nucleotide sequence of the mitochondrial genome of *Calliphora chinghaiensis* (Diptera: Calliphoridae). *Mitochondrial DNA Part B* **1**, 397–398 (2016).

71. Akbarzadeh, K., Wallman, J. F., Sulakova, H. & Szpila, K. Species identification of Middle Eastern blowflies (Diptera: Calliphoridae) of forensic importance. *Parasitol. Res.* **114**, 1463–1472 (2015).

72. Ren, L., Guo, Q., Yan, W., Guo, Y. & Ding, Y. The complete mitochondria genome of *Calliphora vomitoria* (Diptera: Calliphoridae). *Mitochondrial DNA Part B* **1**, 378–379 (2016).

73. Šuláková, H. & Barták, M. Forensically important Calliphoridae (Diptera) associated with animal and human decomposition in the Czech Republic: preliminary results. *Cas. slezskeho zemskeho Muz.* **62**, 255–266 (2013).

74. Yeates, D. K. & Wiegmann, B. M. The impact of the Manual of Nearctic Diptera on phylogenetic dipterology. *Can. Entomol.* **144**, 197–205 (2012).

75. Winkler, I. S. *et al.* Explosive radiation or uninformative genes? Origin and early diversification of tachinid flies (Diptera: Tachinidae). *Mol. Phylogenet. Evol.* **88**, 38–54 (2015).

76. Cameron, S. L. Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny mt:

mitochondria PCG: protein-coding gene. *Annu. Rev. Entomol.* **59**, 95–117 (2014).

77. Bevan, R. B. & Lang, F. B. Mitochondrial genome evolution: the origin of mitochondria and of eukaryotes. in *Mitochondrial Function and Biogenesis* (ed. Koehler, C.) (Springer, 2004).

78. Clary, D. O. & Wolstenholme, D. R. The mitochondrial DNA molecular of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22**, 252–271 (1985).

79. Simon, C. *et al.* Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers. *Ann. Entomol. Soc. Am.* **87**, 651–701 (1994).

80. Simon, C., Buckley, T. R., Frati, F., Stewart, J. B. & Beckenbach, A. T. Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA. (2006) doi:10.1146/annurev.ecolsys.37.091305.110018.

81. Beckenbach, A. T. Mitochondrial genome sequences of nematocera (lower diptera): Evidence of rearrangement following a complete genome duplication in a winter crane fly. *Genome Biol. Evol.* **4**, 89–101 (2012).

82. Beckenbach, A. T. & Stewart, J. B. Insect mitochondrial genomics 3: the complete mitochondrial genome sequences of representatives from two neuropteroid orders: a dobsonfly (order Megaloptera) and a giant lacewing and an owlfly (order Neuroptera). *Genome* **52**, 31–38 (2009).

83. Zhang, D. *et al.* Phylogenetic inference of calyptrates, with the first mitogenomes for Gasterophilinae (Diptera: Oestridae) and Paramacronychiinae (Diptera: Sarcophagidae). *Int. J. Biol. Sci.* **12**, 489–504 (2016).

84. Adams, M. D. *et al.* The Genome Sequence of Drosophila melanogaster. *Science (80-.).* **287**, 2185–2195 (2000).

85. Wiegmann, B. M. Genomes of Diptera. *Curr. Opin. Insect Sci.* **25**, 116–124 (2018).

86. Vicoso, B. & Bachtrog, D. Numerous Transitions of Sex Chromosomes in Diptera. *PLOS Biol.* **13**, e1002078 (2015).

87. Dikow, R. B., Frandsen, P. B., Turcatel, M. & Dikow, T. Genomic and transcriptomic resources for assassin flies including the complete genome sequence of Proctacanthus coquilletti (Insecta: Diptera: Asilidae) and 16 representative transcriptomes. *PeerJ* **2017**, e2951 (2017).

88. Lemmon, A. R., Emme, S. A. & Lemmon, E. M. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Syst. Biol.* **61**, 727–744 (2012).

89. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science (80-.).* **346**, 763–767 (2014).

CHAPTER 2

Mitochondrial Genome of *Blepharipa sp.*

> "Over the long term, symbiosis is more useful than parasitism. More fun, too. Ask any mitochondria."
>
> —Larry Wall

TH-2821_146106003

CHAPTER 2

Mitochondrial genome sequencing of *Blepharipa sp.*, an endoparasite of Muga silkworm, and comparative codon usage analysis with other Oestroidea flies

Abstract:

Uziflies (Family: Tachinidae) are dipteran endoparasites of sericigenous insects which cause major economic losses in the silk industries. The first complete mitogenome of *Blepharipa sp.* (Acc. No.: KY644698, 15080 bp, A+T = 78.41 %), a dipteran parasitoid of the Muga silkworm (*Antheraea assamensis*) found in the Indian states of Assam and Meghalaya, is presented here.This study has confirmed that *Blepharipa sp.* mitogenome gene content and arrangement is similar to other Tachinidae and Sarcophagidae flies of Oestroidea superfamily, typical of ancestral Diptera. However, Calliphoridae and Oestridae flies have undergone tRNA translocation and insertion, forming unique intergenic spacers (IGS) and overlapping regions (OL) and a few of them (IGS, OL) have been conserved across Oestroidea flies. Variation of tandem repeats and sequence coverage of the control region (CR) effect mitogenome size and tachinid flies have fairly smaller mitogenome owing to their small CR (*Blepharipa sp.*: 168 bp). Our research unveils those genes with a high AT content had a reduced effective number of codons, leading to high codon usage bias. The neutrality test shows that natural selection has a stronger influence on codon usage bias than directed mutational pressure. This study also reveals that longer PCGs (e.g., *nad5*, *cox1*) have a higher codon usage bias than shorter PCGs

(e.g., *atp8*, *nad4l*). The divergence rates increase nonlinearly as AT content at the 3rd codon position increases and higher rate of synonymous divergence than nonsynonymous divergence causes strong purifying selection. The phylogenetic analysis explains that *Blepharipa sp.* is well suited in the family of insectivorous tachinid maggots. It's possible that biased codon usage in the Tachinidae family reduces the effective number of codons, and purifying selection retains the core functions in their mitogenome, which could help with efficient metabolism in their endo-parasitic life strategy.

2.1 Introduction:

Insect mitochondria which arose from alpha-proteobacteria have its own circular mitogenome of about 14-20 kb[1–3]. The inner membrane of this organelle harbours five distinct protein complexes for efficient production of energy via oxidative phosphorylation (OXPHOS) process[4,5]. In general, the insect mitogenome has 13 protein-coding genes (PCGs), 2 ribosomal RNAs (rRNAs), 21 to 23 transfer RNAs (tRNAs)[6]. It also contains several non-coding regions with the longest being AT-rich control region (Table 2.1)[7]. A typical metazoan mitogenome is often small in size, maternally inherited, mutation prone, with low or no homologous recombination, conserved gene content, and high genetic polymorphism, making it a viable marker for barcoding, phylogeographic, phylogenetic, and molecular dating research[8,9,10]. However, there has been lack of investigation on mitochondrial codon change and its impact in environmental adaptation[10,11]. Work on differential mitochondrial codon usage has been seen mostly in vertebrates, but only a few parasitic Platyhelminthes, ribbon worms, and moths have been studied invertebrates to date [12–14].

Tachinidae is the biggest family of the Oestroidea superfamily, having around 10,000 incredibly diverse, koinobiont, internal parasitoid flies with indistinguishable phenotypic and morphology, making taxonomic categorization difficult[15–17]. The Tachinid larva hides, feeds, and respires inside the host larva before fast devouring it in the late larval or pupal stage, finally

killing it[1,2]. Tachinid flies have a wide range of hosts, including caterpillars, bugs, adult and larval beetles, and a variety of other arthropods and non-arthropods [16–18]. However, only about half of the species in this family have enough biological information, such as host range, necessary habitat, and mating mechanism[19,20]. Other Oestroidea flies have been rigorously studied in forensic science and as a myiasis-causing agent in humans and many domestic animals (Table 2.1). The Oestroidea flies rely on dead or living animals (necrophagous, sarcophagous, saprophagous) to complete early phases of metamorphosis[16]. Among Oestroidea, Tachinids adopt a different survival strategy in the larval phase in which they are surrounded by an oxygen-limited environment and are vulnerable to host immune systems[19,21,22]. Uzi flies are Tachinids, responsible for infestation and death of commercially important silkworms. Four species of uzi flies are identified till date viz., the Japanese uzi fly, *Crossocosmia sericaria* (Rodani); the Hime uzi fly, *Ctenophora pavida* (Meigen); the Tasar uzi fly, *Blepharipa zebina* (Walker) and the Indian uzi fly, *Exorista sorbillans* (Wiedemann)[23]. The last two dipteran endo-parasites (mulberry, muga, and tasar) had a severe impact on the Indian sericulture sector, inflicting economic losses to rural sericulture farmers in India[18,23,24]. The currently studied uzifly species, *Blepharipa sp.*, found in Assam and Meghalaya, causes the death of muga silkworm (*A. assamensis*) larva during winter and post-winter season and has been accounted for around 80-90% yield loss in muga seed cultivating areas[25–27].

Despite having the scientific importance of mitogenome and economic significance of Tachinid flies, only 4 mitogenomes of this family is available in the public databases till date (3 listed in Table 2.1). In this chapter, the complete mitogenome (mtDNA) sequence from Blepharipa genus (*Blepharipa sp.*) using next-generation sequencing (GenBank Acc No. KY644698) has been presented. A comprehensive comparison with other Oestroidea mitogenomes (Table 2.1) accessible at NCBI is also provided. Although, the sequenced mitogenome of *Blepharipa sp.* was acknowledged and used in a recent article on the full mitochondrial sequence of another Tachinidae fly, *Exorista japonica*, it was not included in this study[28]. Several mitogenome

TH-2821_146106003

physiognomies, including as size, nucleotide composition, and gene organization, have been explored in Oestroidea flies and other outgroups. This study also emphasized on mitochondrial codon usage pattern since every organism possesses a unique codon choice which is related to gene expression, translational efficiency, and further protein structure and function[29–32]. We found that whole mitogenome (WMG) and protein coding genes (PCGs) of Tachinid flies are highly AT biased in nature than other flies which is in agreement with the report of Zhao *et al.*[33]. In conjunction, the 3rd codon positions are AT-rich, resulting in the use of fewer effective number codons and maximum biased codons in the PCGs of this family. The substitution rate analysis of PCGs indicates that rate of synonymous divergence is higher than nonsynonymous divergence due to prevalence of purifying selection (dN/dS<1) in branch leading to *Blepharipa sp.* as well as in background branches. This study also ascertains that longer genes in mitochondria, such as *nad5*, *nad4*, *nad1*, and *cox1*, employ more biased codons than shorter genes (*nad4l*, *atp8*), which is also seen in intron-less prokaryotic protein-coding genes[34,35]. Neutrality test supports the role of natural selection in shaping codon choice in protein-coding genes. The regression analysis between nucleotide substitution rates and various codon usage indices suggests that a nonlinear model is more effective than a typical linear model in delineating relationships. It asserts that the rate of divergence rises with increasing AT concentration at the 3rd codon position along a nonlinear S-shape curve, and rate of synonymous divergence is higher than nonsynonymous divergence. The use of strongly biased codons by Tachinids leads to a reduction in the effective number of codons which may contribute to the efficient metabolism of endo-parasitic life strategies. Further, phylogenies of Oestroidea exhibited well-supported monophyly of Sarcophagidae and Calliphoridae family.

TH-2821_1461106003

Table 2.1: List of Diptera (n=42) and outgroup Lepidoptera (n=2) used in this study for comparative mitogenomics and phylogenetic analysis. (A= Ancestral mitogenome arrangement)

Sl No.	Accession No.	Family	Organisms	Mito genome (bp)/ AT%	CR (bp)/ AT %	rRNA (bp)/ AT%	tRNA (bp)/ AT%	PCG size (bp)/ AT%	Mito Genome Pattern	Common Name	Economic Importance	Lifestyle/ Food habit	References
1	NC_019632	Calliphoridae	*Chrysomya bezziana*	15,236	392	2116	1469	11,151	A+ *trnI* + duplication CR	Old World screwworm	Causes Myiasis in animal and human	Obligate Ectoparasite/ Necrophagous	4,6,36
				75.89	87.75	79.63	76.31	74.53					
2	NC_026996	Calliphoridae	*Aldrichina grahami*	14,903	89	2107	1475	11,118	A	Blow fly	Forensic insect, transmit human and animal pathogens	Necrophagous	77–9
				76.75	92.13	80.06	76.47	75.91					
3	NC_025338	Calliphoridae	*Chrysomya pinguis*	15,838	988	2114	1478	11,151	A+ *trnI*	Blowfly	Forensically Important	Ectoparasite/ Necrophagous	10–12
				76.06	88.25	79.75	75.98	74.11					
4	NC_019631	Calliphoridae	*Chrysomya albiceps*	15,491	657	2113	1472	11,151	A+ *trnI* + duplication CR	Hairy maggot blowfly	Cause secondary myiases	Necrophagous	44,13
				77.26	85.69	80.40	76.22	76.16					
5	NC_019636	Calliphoridae	*Protophormia terraenovae*	15,170	356	2112	1472	11,151	A	Northern blowfly	Myiasis pest of livestock	Ectoparasite/ Necrophagous	4,14
				75.87	90.73	80.06	76.01	74.44					
6	NC_019635	Calliphoridae	*Chrysomya saffranea*	15,839	994	2114	1472	11,151	A+ *trnI* + duplication CR	Steelblue blowfly	Forensic insect and causes myiasis in human beings and animals	Necrophagous	4,15
				76.45	88.12	79.84	76.08	74.63					
7	NC_019638	Calliphoridae	*Hemipyrellia ligurriens*	15,938	1119	2115	1473	11,157	A	Blowfly	Forensic insect, myiasis in goat, buffalo and bull, vector of pathogens	Parasite/ Necrophagous	4,16,17
				77.35	89.72	80.14	76.98	75.50					
8	NC_019637	Calliphoridae	*Lucilia porphyrina*	15,877	1047	2115	1470	11,157	A	Porphyrina blow fly/ Oriental blow fly	Forensic insect, myiasis in livestock, human	Ectoparasite/ Human or animal corpses/ Necrophagous	4,11,18,19
				76.26	88.92	79.57	76.46	74.26					
9	NC_019634	Calliphoridae	*Chrysomya rufifacies*	15,412	574	2114	1473	11,151	A+ *trnI* + duplication CR	Hairy maggot blowfly	Forensic insect, myiasis in livestock	Necrophagous	4
				77.20	84.84	80.22	76.57	76.18					
10	NC_019633	Calliphoridae	*Chrysomya megacephala*	15,273	428	2114	1472	11,151	A+ *trnI* + duplication CR	Oriental latrine fly	Forensic insect, myiasis in livestock	Necrophagous	20
				75.98	87.14	79.70	75.88	74.66					
11	NC_002660	Calliphoridae	*Cochliomyia hominivorax*	16,022	1175	2110	1470	11,157	CR A	New World screw-worm fly	Forensic insect, myiasis in mammals	Necrophagous	37
				76.90	90.80	79.81	76.59	74.72					
12	NC_002697	Calliphoridae	*Chrysomya putoria*	15,837	1008	2114	1471	11,154	A+ *trnI* + duplication CR	Tropical African latrine blowfly	Forensic insect, myiasis in mammals	Necrophagous	21
				76.70	88.59	79.99	76.13	74.91					

13	NC_031381	Calliphoridae	*Chrysomya phaonis*	15,831 76.09	992 88.10	2112 79.59	1472 75.81	11,151 74.23	A+ *trnI*	Blow flies	Medical and forensic importance.	Necrophagous	22
14	NC_029486	Calliphoridae	*Lucilia coeruleiviridis*	14,989 76.02	168 87.5	2110 80	1471 76.88	11,145 74.87	A + Partial CR	Green bottle fly	Forensic insect, myiasis in pig and other mammals	Ectoparasite/ Necrophagous	19,23
15	NC_029215	Calliphoridae	*Calliphora chinghaiensis*	15,269 76.75	441 84.35	2113 80.59	1463 76.82	11,190 75.55	Translocation of *trnS1*	Blue bottle flies	Forensic importance	Necrophagous	24,25
16	NC_028411	Calliphoridae	*Calliphora vomitoria*	16,134 77.55	1319 90.29	2110 80.33	1471 76.41	11,151 75.54	A	Blue bottle fly	Forensic importance and causes myiasis	Necrophagous	26,27,38
17	NC_028412	Calliphoridae	*Chrysomya nigripes*	15,832 76.92	966 88.09	2115 80.14	1476 76.01	11,154 75.24	A+ *trnI* + duplication CR	Blowfly	Forensic importance	Necrophagous	22,39
18	NC_028056	Calliphoridae	*Lucilia illustris*	15,954 77.42	1094 90.85	2153 79.74	1469 76.85	11,100 75.60	CR A	Green bottle fly	Forensic importance and myiasis in pet Animals	Ectoparasite/ Necrophagous	19,40
19	NC_028057	Calliphoridae	*Lucilia Caesar*	15,954 77.30	1117 90.59	2152 79.73	1469 76.78	11,121 75.39	CR A	Common greenbottle.	Forensic importance and facultative wound myiasis	Ectoparasite/ Necrophagous	19,41,42
20	NC_013932	Oestridae	*Hypoderma lineatum*	16,354 77.85	1493 87.54	2101 80.48	1453 77.70	11,136 75.85	A	Common cattle grub/ warble fly	Causes Myiasis in ruminants	Ectoparasite/ Sarcophagous/ Carnivore	43–46
21	NC_006378	Oestridae	*Dermatobia hominis*	16,360 77.81	1545 91.39	2112 81.43	1458 77.09	11,157 75.23	Insertion of *trnV* between *trnK-trnD*	Human botfly/ tropical warble fly	Causes Myiasis in human, cattle, dogs and forensically importance	Endoparasites of birds and mammals/ Carnivore	44,47–49
22	NC_029812	Oestridae	*Gasterophilus pecorum*	15,750 70.73	1001 80.81	2048 74.31	1461 75.56	11,103 68.46	A	Horse botfly	Gastrointestinal myiasis in equines and forensic science	Obligate intestinal parasites/ Carnivore	43,50
23	NC_029834	Oestridae	*Gasterophilus intestinalis*	15,660 70.16	875 80.8	2107 73.84	1470 74.69	11,103 67.88	CR A	Horse botfly	Gastric myiasis in horse, donkey	Obligate internal parasites /Carnivore	51–53
24	NC_026196	Sarcophagidae	*Ravinia pernix*	15,778 77.17	1750 84.34	2114 80.36	1470 76.32	11,154 75.46	A		Forensic importance, potential for myiasis	Endo-parasitoid/ Saprophagous	3,54
25	NC_026112	Sarcophagidae	*Sarcophaga melanura*	15,190 75.64	360 90.27	2108 80.07	1475 76.61	11,154 74.04	A	Flesh fly	Forensic importance, causes myiasis	Ectoparasite/ Saprophagous	55–58
26	NC_025944	Sarcophagidae	*Sarcophaga Africa*	15,144 75.74	338 89.34	2111 79.91	1469 76.31	11,151 74.32	A	Flesh fly	Intestinal myiasis and forensic science	Ectoparasite/ Saprophagous	58–60
27	NC_025574	Sarcophagidae	*Sarcophaga portschinskyi*	14,929 76.18	118 89.83	2109 80.41	1468 76.08	11,139 75.12	A	Flesh fly	Forensic importance	Ectoparasite/ Saprophagous	58,61
28	NC_025573	Sarcophagidae	*Sarcophaga similis*	15,158 76.36	354 87.57	2107 80.25	1461 76.11	11,139 75.21	A	Flesh fly	Responsible for myiasis and forensic importance	Ectoparasite/ Saprophagous	58,62–64
29	NC_023532	Sarcophagidae	*Sarcophaga peregrine*	14,922 74.97	123 87.80	2108 79.83	1470 76.12	11,139 73.61	A	Flesh fly	Responsible for myiasis and forensic importance	Ectoparasite/ Saprophagous	58,65,66
30	NC_017605	Sarcophagidae	*Sarcophaga impatiens*	15,169 74.76	359 88.30	2113 79.46	1469 76.37	11,154 73.08	A	Flesh fly	Forensic importance Carrion breeding	Ectoparasite/ Saprophagous	58,67
31	NC_026667	Sarcophagidae		15,420	613	2109	1484	11,153	A	Flesh fly			

			Sarcophaga crassipalpis	76.22	89.39	80.03	76.21	74.65			Forensic importance responsible of myiasis	Ectoparasite/ Saprophagous	58,68,69
32	NC_028413	Sarcophagidae	*Sarcophaga albiceps*	14,935 75.86	125 90.4	2111 79.77	1470 75.78	11,139 74.84	A	Flesh fly	Forensically important	Ectoparasite/ Saprophagous	58,70
33	**Current Study**	Tachinidae	***Blepharipa sp.***	**15,080** **78.41**	168 92.60	2143 82.54	1466 78.24	**11,166** **77.27**	**A**	Uzi Fly	Endoparasite of muga silkworm	Endoparasite/ Parasitoid	This Study
34	NC_019640	Tachinidae	*Rutilia goerlingiana*	15,331 77.70	568 91.07	2101 81.81	1451 77.18	11,131 76.11	A	Tachinid flies	Insect endoparasite	Endoparasite/ Parasitoid	4
35	NC_018118#	Tachinidae	*Elodia flavipalpis*	14,932 79.96	105 92.38	2120 83.49	1463 79.76	11,154 79.09	A	Tachinid flies	Natural enemies of the leaf-roller moths	Endoparasite/ Parasitoid	33
36	NC_014704#	Tachinidae	*Exorista sorbillans*	14,960 78.44	105 98.09	2117 81.76	1471 76.75	11,136 77.64	A	Uzi Fly	Endoparasite of mulberry silkworm	Endoparasite/ Parasitoid	71
37	NC_016713	Agromyzidae	*Liriomyza bryoniae*	16,183 79.26	1354 95.49	2111 82.42	1468 78.54	11,169 76.66	A	Tomato leaf miner	Pest species of Tomato and other vegetables (Cucurbitaceae and Solanaceae,)	Ectoparasite /Polyphagous/ Herbivore	72–74
38	NC_015926	Agromyzidae	*Liriomyza sativae*	15,551 77.53	741 92.98	2111 82.18	1465 76.99	11,160 75.59	A	Vegetable leafminer	Pest species vegetables	Ectoparasite /Polyphagous/ Herbivore	73,75
39	NC_014402	Tephritidae	*Bactrocera minax*	16,043 67.28	1140 77.63	2115 73.71	1466 72.30	11,151 64.21	A	Oriental citrus fly	Pest of citrus and related genera of Rutaceae	Phytophagous/ Herbivore	76,77
40	NC_029468	Tephritidae	*Bactrocera umbrosa*	15,898 70.48	944 86.22	2120 77.02	1465 74.12	11,157 67.19	A	Oriental fruit fly	Pest of Moraceae family	Phytophagous/ Herbivore	77,78
41	NC_015079	Culicidae	*Culex pipiens pipiens*	14,856 77.63	0	2118 82.24	1475 78.98	11,187 76.46	Inversion (*trnA-trnR => trnR-trnA*)	Culex Mosquito	Vector of multiple diseases (West Nile virus)	Free living/ Multivoltine	79
42	NC_027502	Culicidae	*Anopheles culicifacies*	15,330 78.44	498 92.57	2113 82.06	1474 78.56	11,199 77.04	Inversion (*trnA-trnR => trnR-trnA*)	Anopheles Mosquito	Vector of multiple diseases	Free living/ Multivoltine	80
43	NC 002355	Lepidoptera (Order)	*Bombyx mori*	15,643 81.32	499 95.39	2158 84.80	1468 81.40	11142 79.50	*trnM-trnI-trnQ*	Mulberry Silkworm	Economically beneficial in silk and textile	Phytophagous/ Herbivore	81
44	KU379695	Lepidoptera (Order)	*Antheraea assamensis*	15,272 80.18	328 91.15	2123 84.26	1465 80.75	11175 78.75	*trnM-trnI-trnQ*	Muga Silkworm	Economically beneficial in silk and textile	Phytophagous/ Herbivore	82

3 # Mitogenomes were used for manual curation of *Blepharipa sp*

2.2 Materials and Method:

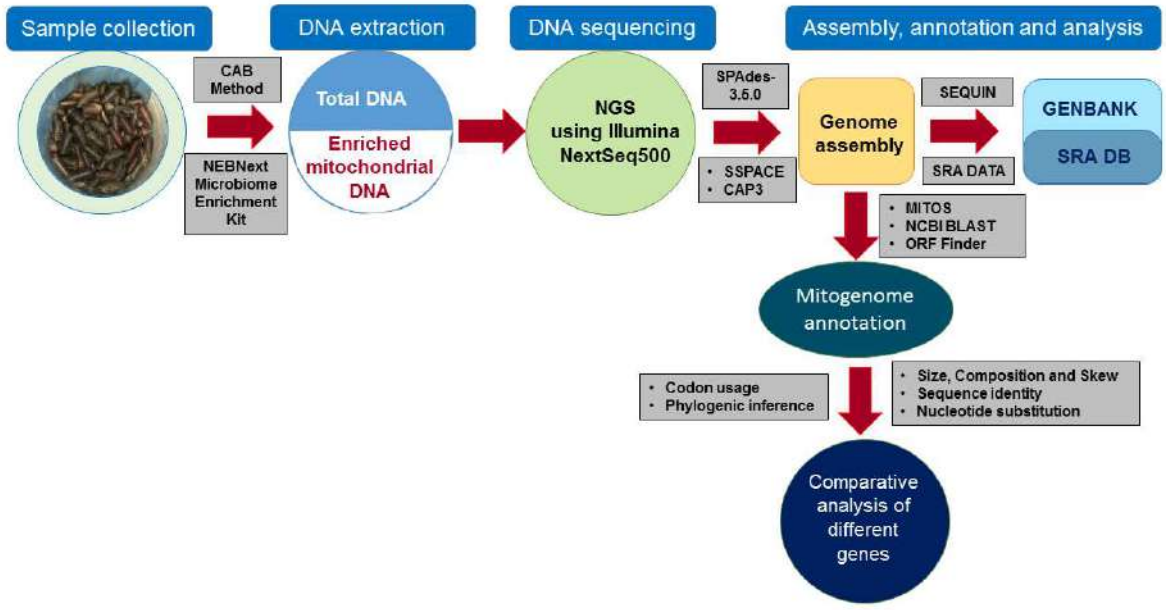


Figure 2.1: Complete workflow of Uzifly (*Blepharipa sp.*) sample collection, mitochondrial genome (mtDNA) sequencing, assembly, annotation and analysis

2.2.1 Sample collection, processing, sequencing, and assembly:

The fully grown *Blepharipa sp.* pupa was obtained from the Central Muga Eri Research and Training Institute (CMER&TI), Jorhat, Assam, India (Lat: 26 ° 47'49.1"N Lon: 94 ° 19'35.0"E) with the Sample ID- CMERI-Uzi-001. The pupa was dissected, chopped, and stored in 95% absolute ethanol at -80°C freezer. The steps involving mitochondrial DNA isolation, library preparation to sequencing, and assembly were carried out at the Genotypic Technology Pvt. Ltd. Bangalore, India (http://www.genotypic.co.in/) and are briefly discussed here. Total DNA was extracted from tissues using CTAB (Cetyl Trimethyl Ammonium Bromide) based method and filtered by silica column (Genotypic Technology Pvt. Ltd. Bengaluru, India). The quality, quantity, and purity of isolated purified DNA was tested using agarose gel electrophoresis, light absorption, and fluorescence spectroscopy.

The library preparation was performed by using Illumina-compatible NEXTFlex DNA library protocol (Cat #5140-02). Mitochondrial DNA was preferentially enriched through NEBNext

microbiome DNA enrichment kit (New England Biolabs, USA) which selectively removed CpG-methylated eukaryotic nuclear DNA. The enriched mitochondrial DNA obtained was sheared to produce fragments of about 200-400 bp in Covaris microTube with the S220 system (Covaris, Woburn, Massachusetts, USA) through focused ultra-sonication. The fragment size distribution was determined using Agilent Tape Station with D1000 DNA Kit (Agilent Technologies, Santa Clara, California, USA). The resulting fragmented DNA was cleaned up by HighPrep magnetic beads (MagBio Genomics, Inc, Gaithersburg, Maryland) to remove salts, primers, primer-dimers, dNTPs, etc. The fragments were subjected to end-repair, A-tailing, and ligation of the Illumina multiplexing adaptors using the NEXTFlex DNA Sequencing kit (Catalogue # 5140–02, BioScientific), followed by purification of adaptor-ligated DNA sequence through HighPrep beads and amplification through PCR. The PCR cycling conditions followed include, the initial denaturation at 98℃ for 2 min; 10 cycles of denaturation at 98℃ for 30 sec; annealing at 65℃ for 30 sec followed by extension at 72˚C for 60 sec; and a final extension at 72˚C for 4 min employing the primers supplied by NEXTFlex DNA Sequencing kit. Further, the amplified PCR product was purified via HighPrep beads, quantified using Qubit fluorometer (Thermo Fisher Scientific, MA, USA) and the fragment range was assessed using Agilent D1000 Tape (Agilent Technologies). Finally, the sequencing was performed using Illumina NextSeq500 (Illumina Inc, Sandiego, USA) through 2×150 bp paired-end chemistry. The raw paired-end reads were de-multiplexed using Bcl2fastQ (V2) and the quality was assessed with FastQC v2.2 tool[83]. The Illumina raw reads were processed by in-house Perl script (ABLT-Scripts (no version available), Genotypic technology, Bangalore India) for the removal adapters and low-quality bases (Q<30) towards 3'-end. The SPAdes-3.6.0 (St. Petersburg genome assembler) was used for *de novo* assembly of reads[84,85], scaffolding of assembled contigs and clustering were carried out with SSPACE (v 2.0) and CAP3 (Version Date: 10/15/07) programs[86,87]. The closest reference species was identified by BLAST (online blast was used) analysis of assembled scaffold against NCBI nr (non-

redundant) database and the alignment of scaffold against reference sequence was done through Bowtie2 (v 2.2.7)[88]. The aligned data was processed using SAMtools (last used July, 2016) for generating reference assisted consensus sequences[89]. Final scaffolding was done in SSPACE using that reference assisted consensus sequence along with spades assembly-based scaffold to correct the regions having N's in the initial scaffold. All tools were run on default parameters. The assembly was then validated using a PCR-based technique on two regions: *nad6* (protein coding gene) and the control region (AT rich region), followed by Sanger sequencing. According to previous reports on Tachinids, NGS sequencing had significantly lower coverage in the control region (CR) compared to other species groups, which was attributed to AT rich bases, lowering the correctness and completeness of Tachinid mitochondrial genome assemblies[33,71]. Hence, primer sets was designed as per Bronstein, O. *et al.* 2018 for targeting the CR of *Blepharipa sp.*[90]. In addition, Bowtie2 (v 2.4.4) was used to check the depth of coverage over the control area using Illumina reads mapped to the assembly [88]. The complete workflow is shown in Figure 2.1.

2.2.2 Mitogenome annotation and documentation:

The assembled scaffolds were annotated using MITOS WebServer [91] (last accessed April 2017). The PCG boundaries (start and stop codons) were determined through NCBI ORF Finder (last accessed April-May 2017) based on the invertebrate mitochondrial genetic code[92]. Additionally, gene boundaries, overlapping and intergenic spacer regions were estimated through NCBI BLAST (last accessed April-May 2017), BioEdit v. 7.2, and ClustalW program of MEGA 7.0 software using reference sequences from other published Dipteran mitogenomes[93–95]. The control region (CR) was confirmed by comparing it with the available sequences in GenBank[96]. The secondary structures of tRNAs were predicted through MITOS Server and confirmed using tRNAscan-SE tool [97]. The secondary structures of mitochondrial rRNAs were examined by using Mfold Web Server[98] (last accessed May 2017). Finally, the annotated file of *Blepharipa sp.* mitogenome was prepared through the NCBI Sequin tool and

SRA data along with the sequin file were submitted to NCBI GenBank (Acc No.: KY644698)[99]. Additionally, for comparative analysis, mitogenome sequences and annotations of other 43 species were downloaded from NCBI (Table 2.1). It is visible from Fig. 2.3B that *Culex pipiens pipiens* (0 bp) and *Ravinia pernix* (1750 bp) display anomalies in their CR size. However, it may be due to an error in NCBI annotation as the associated literature of R. pernix had documented the CR size as 965 bp[3].

2.2.3 Sequence alignment and phylogenetic inference:

To obtain the molecular phylogeny of Oestroidea flies, especially among 4 four distinct families (Calliphoridae, Sarcophagidae, Oestridae, and Tachinidae) listed in Table 2.1, were selected to use in phylogenetic analysis, including 2 species from each of the Tephritidae, Agromyzidae, Culicidae family, and 2 species from the order Lepidoptera (*B. mori* and *A. assamensis*) as an outgroup. The translated nucleotide sequences of each PCGs were aligned using MAFFT v. 5 algorithm in TranslatorX server (http://translatorx.co.uk/; last accessed July 2017) which were again back translated[100,101]. The rRNAs were aligned via Clustal Omega and tRNAs were aligned via Clustal W[102]. After that, individual aligned PCGs (rRNAs and tRNAs not included) were concatenated using the nexus module of Bio-python programme[103]. Substitution model optimization for the dataset was performed in jModelTest 2.1.7[104]. The Bayesian analysis of the dataset was conducted with MrBayes v3.2.6 based on the Markov chain Monte Carlo (MCMC) method for 2,000,000 generations[105]. Two independent runs with four chains (one cold and three heated chains) were sampled every 1,000 MCMC steps. A 50% majority-rule consensus tree was built after discarding the initial 10% as burn-in and node supports were analyzed based on posterior probabilities (PP). Other parameters like effective sample size (ESS > 200) and potential scale reduction factor (PSRF) were evaluated for stationary using Tracer v1.6[106]. The Maximum Likelihood analysis was executed using RAxML 8.2.x with 5,000 bootstrap replicates and the rapid bootstrap feature (random seed value 12345)[107]. The individual gene trees for 13 PCGs also estimated similarly through

RaxML 8.2.x with 5,000 bootstrap replicates. Finally, the consensus phylogenetic trees for the dataset were visualized and edited using iToL v3.6.1 tool[108]. To create a contour map, RaxML cladogram tree was generated using Figtree v1.4.4 (https://www.softpedia.com/get/Science-CAD/FigTree-AR.shtml/) and used as a reference tree for contMap function in the R v. 4.0.2 environment using package Phytools[109].

2.2.4 Nucleotide content, skew and substitution analysis:

The nucleotide composition of the whole mitochondrial genome, concatenated and individual PCGs, tRNAs, rRNAs, intergenic spacers, and control region was calculated using MEGA 7.0 software[95]. The base composition skewness was also calculated for all the regions of mitogenome using the formula (Eqn. i and Eqn. ii)[21].

$$AT\ skew = (A-T)/(A+T) \qquad (i)$$

$$GC\ skew = (G-C)/(G+C) \qquad (ii)$$

where A, T, G, and C denote the frequencies of respective bases.

Further gene alignments, consensus species tree, and individual gene trees were used for the investigation of molecular evolution. The analysis was constrained only to the branch of interest and we used a gene-level approach based on the ratio (ω) of nonsynonymous (dN) to synonymous (dS) substitutions rate (ω = dN/dS) to detect possible diversifying selection, via likelihood ratio tests through CODEML algorithm from the PAML package[110]. We tested branch-specific models M0, the simplest model, which has a single ω ratio for the entire tree. Furthermore, for both types of trees, two-ratio models with separate ratios for background and foreground lineage were used, with Blepharipa sp. lineage serving as a foreground branch (gene tree and species tree). The significance level for these LRTs (likelihood ratio test) was measured using a $\chi 2$ approximation, where twice the difference of log-likelihood between the models ($2\Delta lnL$) would be asymptotic to a $\chi 2$ distribution, with the number of degrees of

freedom corresponding to the difference in the number of parameters between the models. Lineage-specific ω value was estimated for each branch through Model=1. Synonymous and non-synonymous divergence rates (dS and dN) was calculated as pairwise manner implementing F3X4 codon frequencies.

The comparison of the control region (CR), overlapping region (OL), and Intergenic spacer (IGS) of *Blepharipa sp.* was carried out with the selected organisms based on the nucleotide identity, length, and location annotation from NCBI. The multiple sequence alignment was performed using Clustal Omega (the online version) and the conserved regions, repeats, and indels in these regions were visualized using BioEdit[94,102].

2.2.5 Codon usage indices calculation and analysis:

Initially, We calculated relative synonymous codon usage (RSCU) of amino acid using MEGA 7.0; which was further confirmed and batch calculation were carried out by DAMBE 6.4.67[95,111]. The cluster analysis of RSCU values was done using CIMminer web tool[112] (last accessed August 2017). Principle component analysis of RSCU values was carried out in R v. 4.0.2 environment using ggfortify package (https://cran.r-project.org/web/packages/ggfortify/index.html/).

Different codon usage indices related to nucleotide composition namely, total of Guanine and Cytosine of any gene (GC), Average of GC at 1st and 2nd codon positions (GC12), GC at 3rd codon position (GC3), and GC content at 3rd codon position for the synonymous codons (GC3s) were calculated. The GC, GC12, GC3 were measured using MEGA 7.0[95], and GC3s was estimated through CodonW (version 1.4.2, http://codonw.sourceforge.net/).

To measure the effective number of codons (ENc), We have followed the calculation of ENc from the study of Sun, X. et al. in (2012) and estimated through DAMBE 6.4.67 software[111,113]. ENc designates the degree of codon bias for genes; where it computes deviation from uniform

codon usage without any prior dependency over the sequence length or specific information of preferred codons[114]. The ENc values range between 20 to 61 and in general, values lesser than 35 signifies strong codon bias[115,116]. To detect different influencing factors of codon usage pattern among the genes in different organisms ENc vs GC3s (ENc-plot) graph was plotted using R v. 3.4.4[113,115]. The standard curve shows the functional relation between ENc and GC3s was under mutation pressure rather than selection[117].

The neutrality test is a plot of GC12 against GC3 (GC12 vs GC3) for demonstrating the relationship between GC12 and GC3, and then investigating the mutation-selection equilibrium in forming the codon usage bias (CUB)[118,119]. The synonymous mutation frequently happens in the 3rd position of codons without changing the amino acid, whereas less frequent nonsynonymous mutations occur in 1st and 2nd positions[117]. Therefore, mutation in the 3rd position of codon is neutral and change in GC content at 1st or the 2nd positions would be correlated with the 3rd codon position if the mutation rate is similar in GC3 and GC12. This indicates that without any external pressure, the occurrence of mutations would be random rather than in a certain direction under the condition of pressure toward higher or lower GC content[118]. Thus, the base composition is similar and there is no variation across three codon positions; but, in the presence of external selection pressure, the base preferences would differ at individual codon positions [117,118]. In the neutrality plot, each gene is represented by discrete points, and when the points are placed along the diagonal line (slope of unity), GC12 is equally neutral to selection as GC3. It means that there will be no significant difference in the rate of mutation between three codon positions due to strong directional mutational pressure and lacks or only a weak external selection pressure[117,120]. Alternatively, as the regression slope of the plot approaches zero or parallel to the horizontal axis, the correlation between GC12 and GC3 declines due to the low mutation rate in GC12 [117,121]. Therefore, the Neutrality plot would be crucial in determining the neutral degree while evaluating evolutionary factors.

2.2.6 Regression modelling between substitution rates and codon usage indices:

To demonstrate the correlation between various substitution rates (dS, dN, and ω) and codon usage indices (GC3, GC3s, GC12, ENc) regression analysis namely linear model (LM), polynomial model (PM), and generalized additive model (GAM) were fitted on a univariate model. All statistical analysis was done using R v. 4.0.2.

Linear regression model forms a straight line between the dependent and independent variables[122]:

$$E(Y) = \beta_0 + \beta_1 X + \varepsilon \quad \text{(iii)}$$

Where Y is the dependent variable, E(Y) is the expected value of Y, β_0 is the intercept, β_1 is the coefficient of X (predictors) and ε is the residual.

Polynomial regression models use the approach of polynomial least squares to fit a non-linear relationship between the dependent and independent variables as an nth degree polynomial[123]:

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X_2 + \ldots + \beta_n X_k + \varepsilon \quad \text{(iv)}$$

Where Y is the dependent variable, E(Y) is the expected value of Y, β_0 is the intercept, β_1, β_2, β_n is the coefficient of X (predictors), k is the degree of polynomial and ε is the residual. We used the poly_degree function from the npbr package in R v. 4.0.2 (https://cran.r-project.org/web/packages/npbr/index.html) for choosing optimal polynomial degrees via the BIC and AIC criterion.

GAM is an additive modelling technique that employs a sum of smoothing functions to represent the predictor variables, and it was fitted using the package mgcv (https://cran.r-project.org/web/packages/mgcv/index.html)[124,125]:

$$g(E(Y)) = a + f_1(X_1) + f_2(X_2) + \ldots\ldots + f_n(X_n) + \varepsilon \quad \text{(v)}$$

where Y is the dependent variable, E(Y) is the expected value of Y, g(Y) is a link function, a is the intercept, f_(1) (X_(1))+f_(2) (X_(2))+… ….+ f_(n) (X_(n)) is the smooth function of predictors, and ε is the residual. Here, we utilized thin plate regression splines (default in mgcv) as a smoothing function and the default gaussian family with the identity link function. All models were plotted using ggplot2 package (https://cran.r-project.org/web/packages/ggplot2/index.html) in R v. 4.0.2.

2.3 Result and Discussion:

2.3.1 Outcome of DNA sequencing, assembly:

Total DNA was isolated from finely chopped, full-grown pupa of *Blepharipa sp.* and its concentration was found to be optimum by NanoDrop spectrophotometer (1294 ng/µl) and by Qubit fluorometer (732.8 ng/µl) followed by mitochondrial DNA enrichment. Tape Station profile showed that the size of the fragments of mitogenomic library was in the range of 250 to 550 bp. The complete insert size distribution ranged from 130 to 430 bp, with the combined adapter size being ~120 bp with mitogenome fragments. The appropriate distribution of fragments and their concentrations (~27.1 ng/µl) were also found to be suitable for sequencing. Sequencing through Illumina NextSeq500 yielded 4402752 raw reads of which around 3663404 high-quality reads were retained after post-quality filtering. The final scaffolding and assembly of contigs generated 15,080 bp single scaffold MtDNA in *Blepharipa sp* (N50 = 15,080).

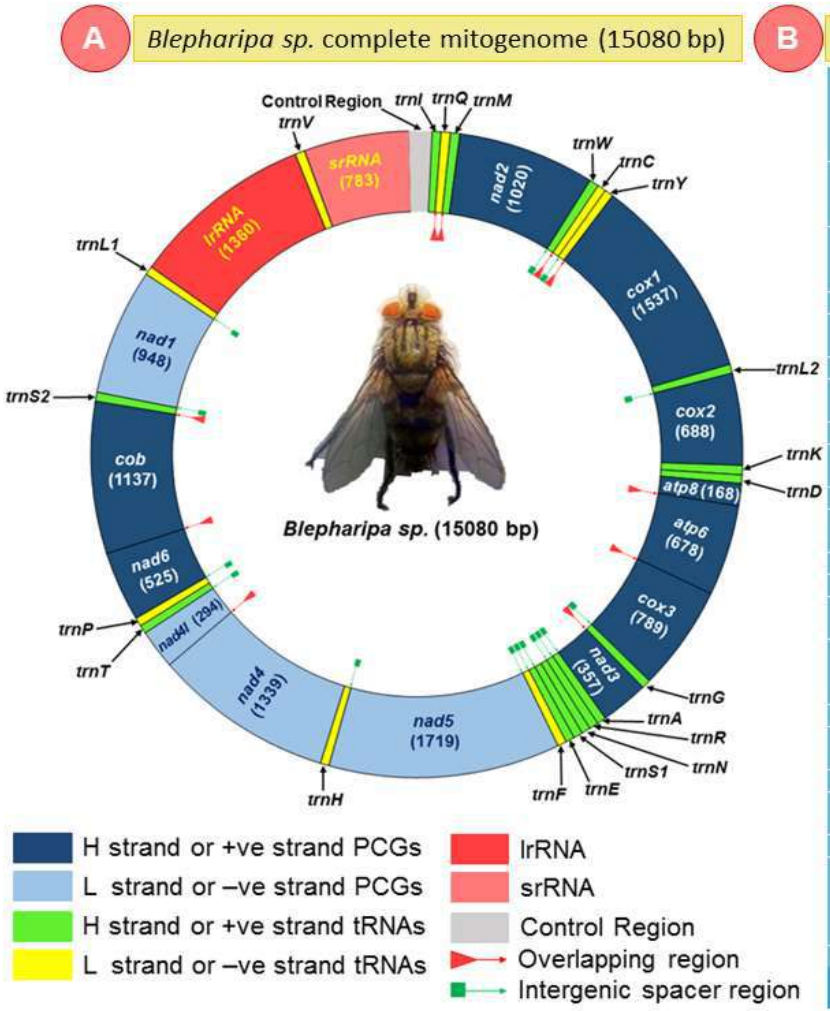
2.3.2 Mitogenome organization and structure of *Blepharipa sp.*:

The newly sequenced mitochondrial genome of *Blepharipa sp.* is closed circular and has a size of 15080 bp which falls within the typical insect mitogenome size (14 to 20 kb)[126–128]. Similar to other sequenced bilaterian mitogenomes the *Blepharipa sp.* mitogenome has conventional gene content, a total of 37 genes (viz. 13 PCGs, 22 tRNAs, 2 rRNAs) and an AT rich control region (CR) (Fig. 2.2A)[129–132]. Among these, 23 genes are present on the major strand (J strand

or +ve strand), while the remaining 14 genes are present in the minor strand (N strand or –ve strand). The intron-less 13 PCGs are also separately encoded by these two strands, 9 PCGs (*nad2, cox1, cox2, atp8, atp6, cox3, nad3, nad6, cytb*) from the J strand and 4 PCGs (*nad5, nad4, nad4l, nad1*) from N strand covering 6899 bp and 4300 bp respectively constituting around 74.31 % of the entire mitogenome (Fig. 2.2). The largest PCG present in this organism is *nad5* (1716 bp) and the smallest one is the *atp8* (165 bp). Excluding stop codons, the J strand has 2237 codons and N strand has 1430 codons respectively. Apart from *cox1* (TCG) and *nad1* (TTG), 11 PCGs follow the canonical "ATN" start codon. Ten PCGs of this mitogenome have "TAA or TAG" as their stop codon except for *cox1*, *cox2*, and *nad4*, where they end with an incomplete stop codon, a single T (Fig. 2.2)[133]. A total of 22 tRNAs are interspersed all over the entire mitogenome ranging from 63 bp (*trnT*) to 72 bp (*trnV*) in size. The J and N strands have 14 tRNAs and 8 tRNAs respectively, with 928 bp and 528 bp nucleotides. Typical clover-leaf shaped secondary structures of tRNAs have been observed with a few exceptions where *trnC*, *trnF*, *trnP*, *trnN* lack stable TΨC loop see Supplementary Fig. S7 online). Two N-strand rRNAs with nucleotides of 1360 bp and 783 bp are transcribed individually for *rrnL* and *rrnS* (Fig. 2.2B). This mitogenome has 10 gene boundaries where genes overlap with adjacent genes, varying from 1 to 8 bp in length, which are altogether 35 bp. Likewise, the cumulative length of the intergenic spacer sequences (excluding the control region) is 139 bp at 15 gene boundaries. The length of the total intergenic spacer varies between 1 to 40 bp and the largest one is located between *trnE* and *trnF*. The longest overlapping sequence of 8 bp is present over *trnW* and *trnC* genes. In this organism, eleven pairs of genes are located discreetly but adjacent to each other, and any PCG adjacent to tRNA, ending with an incomplete stop codon (*cox1-trnL2*, *cox2-trnK*). The control region's length of this Dipteran fly is 168 bp and the nature of this region is extremely biased towards A+T content (Fig. 2.2).

TH-2821_146106003

TH-2821_146106003



B Annotation and mitogenome organization of *Blepharipa sp.*

Gene name	Strand	Start Codon	Stop Codon	Anti-codon	Start	End	OL/IS	Length	AT skew	AT content
Control Region					1	112	0	112	0.203	91.070
trnI	+			GAT (142-144)	113	178	-3	66	0.115	78.790
trnQ	-			TTG (212-214)	176	244	-1	69	-0.123	82.610
trnM	+			CAT (274-276)	244	313	0	70	0.000	71.430
nad2	+	ATT	TAA		314	1333	5	1020	-0.108	81.690
trnW	+			TCA (1369-1371)	1339	1408	-8	70	0.185	77.140
trnC	-			GCA (1432-1434)	1401	1463	6	63	0.067	71.430
trnY	-			GTA (1501-1503)	1470	1534	-2	65	0.059	78.460
cox1	+	TCG	T-		1533	3069	0	1537	-0.093	69.550
trnL2	+			TAA (3099-3101)	3070	3135	7	66	0.021	71.210
cox2	+	ATG	T-		3143	3830	0	688	-0.019	76.740
trnK	+			CTT (3861-3863)	3831	3900	0	70	-0.021	67.140
trnD	+			GTC (3931-3933)	3901	3964	0	64	0.036	87.500
atp8	+	ATC	TAA		3965	4132	-7	168	0.007	85.120
atp6	+	ATG	TAA		4126	4803	-1	678	-0.126	76.990
cox3	+	ATG	TAA		4803	5591	9	789	-0.119	72.500
trnG	+			TCC (5633-5635)	5601	5667	-3	67	0.154	77.610
nad3	+	ATA	TAA		5665	6021	2	357	-0.088	82.910
trnA	+			TGC (6053-6055)	6024	6090	6	67	-0.115	77.610
trnR	+			TCG (6126-6128)	6097	6165	5	69	0.042	69.570
trnN	+			GTT (6201-6203)	6171	6235	0	65	0.137	78.460
trnS1	+			GCT (6261-6263)	6236	6303	19	68	-0.115	76.470
trnE	+			TTC (6353-6355)	6323	6388	40	66	0.000	87.880
trnF	-			GAA (6460-6462)	6429	6495	4	67	0.127	82.090
nad5	-	ATT	TAG		6500	8218	15	1719	-0.168	79.810
trnH	-			GTG (8265-8267)	8234	8297	0	64	-0.038	81.250
nad4	-	ATG	T-		8298	9636	-7	1339	-0.191	78.860
nad4l	-	ATG	TAA		9630	9923	2	294	-0.207	82.310
trnT	+			TGT (9956-9958)	9926	9988	0	63	0.020	80.950
trnP	-			TGG (10020-10022)	9989	10052	2	64	0.000	81.250
nad6	+	ATT	TAA		10055	10579	-1	525	-0.107	85.710
cob	+	ATG	TAG		10579	11715	-2	1137	-0.111	74.580
trnS2	+			TGA (11742-11744)	11714	11780	16	67	0.037	80.600
nad1	-	TTG	TAA		11797	12744	1	948	-0.283	76.790
trnL1	-			TAG (12777-12779)	12746	12809	0	64	-0.094	82.810
rrnL	-				12810	14169	0	1360	0.031	83.310
trnV	-			TAC (14206-14208)	14170	14241	0	72	0.069	80.560
rrnS	-				14242	15024	0	783	-0.016	81.230
Control Region					15025	15080		56	0.203	91.070

Figure 2.2: Complete mitochondrial genome structure of *Blepharipa sp.* A. Circular Map B. Annotation and genome organization of mitogenome. tRNAs are represented as trn followed by the IUPAC-IUB single letter amino acid codes e.g., *trnM* denote tRNAMet

2.3.3 Size comparison of Oestroidea mitogenome and their genes:

The *Blepharipa sp.* whole mitogenome size (15080 bp) is 416 bp smaller than the average size of Oestroidea mitogenome. Whereas *D. hominis* (human bot fly), an Oestridae fly has the largest mitogenome (16360 bp), and *A. grahami*, a Calliphoridae fly in this superfamily has the shortest (14903 bp). The average mitogenome size of tachinid flies (~15,076 bp) are smaller than the average size of the other flies in this superfamily and the Oestridae flies have a relatively larger mitogenome (~16,031 bp) on average. The average length of total PCGs, tRNAs, and rRNAs was observed to be 11,145 bp, 1482 bp, and 2113 bp, respectively (Fig. 2.3A, green, yellow, and blue lines, Table 2.1). The difference in mitogenome size in insects can be attributed to variations in the length of non-coding regions, especially the control region that differs in length as well as the pattern of sequences (Fig. 2.3B)[82,134]. In addition, based on mtDNA sequence similarity among all the Oestroidea flies, *Blepharipa sp.* has high resemblance with the Tachinid Fly *E. flavipalpis* (87.83%) followed by the two hairy maggot blowflies, *Chrysomya albiceps* (85.51%) and *C. rufifacies* (85.44%). Another well-studied uzi fly, *E. sorbilans* displayed 84.82% similarity and Gasterophilus horse botfly noted the lowest sequence similarity (~77 %) with *Blepharipa sp.*.

2.3.4 Gene content and arrangement:

This study identified that the Oestroidea mitogenome represents the reserved gene arrangement of Ecdysozoan for which it can be easily distinguishable from other bilaterians (Lophotrochozoa and Deuterostomia)[131]. The mitogenome of *Blepharipa sp.* and other Oestroidea have three core tRNA clusters including (1) *trnI-trnQ-trnM,* (2) *trnW-trnC-trnY* and (3) *trnA-trnR-trnN-trnS1-trnE-trnF,* as depicted in Fig. 2.2 and Fig. 2.3C. A comparative study revealed that Oestroidea superfamily has 4 different kinds of mitogenome arrangement (Fig. 2.3C). The majority of the Oestroidea flies (25 out of 36) of this study have ancestral (A) dipteran type mitogenome sequence (Table 2.1)[135]. Minor inconsistencies exist in the Calliphoridae family (blowflies), such as the insertion of extra tRNAs (*trnI* in the genus

TH-2821_146106003

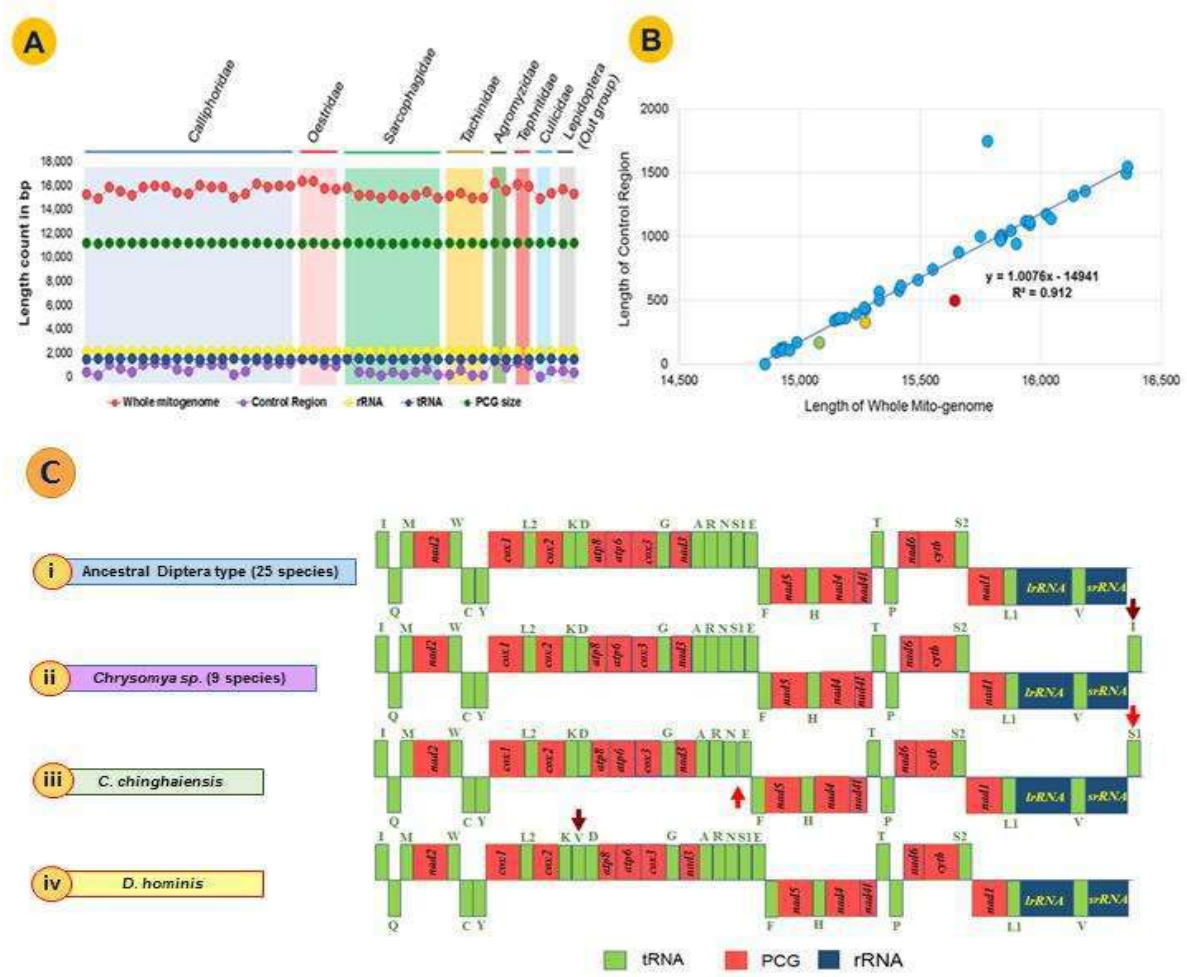


Figure 2.3: A) Whole mitogenome (WMG), Protein coding genes (PCG), tRNA, rRNA and Control region (CR) length variation among Ostroidea Superfamily. B) Relation between WMG and CR length (R^2= .912 p<0.001). Green bubble = *Blepharipa sp.*, Yellow bubble = *Antheraea assamensis*, Red bubble = *Bombyx mori*, the isolated bubble represents *Ravinia pernix* and the only bubble on the X axis represents *Culex pipiens pipiens*. C) Gene arrangement of *Blepharipa sp.* mitogenome (i), a common Diptera type with respect to other selected exceptional arrangement of Oestroidea superfamily (ii, iii, iv). Downward brown arrow= Insertion of tRNA; Upward-downward red arrow= translocation of tRNA. The J strand genes were shown in upward direction and the N strand genes were downward direction.

Chrysomya and *trnV* in *D. hominis*) or the translocation of tRNA (*trnS1* in *C. chinghaiensis*) (Fig. 2.3C)[21,24]. Apart from that, all species, including *Blepharipa sp.*, have 37 genes in their respective mitogenomes and follow a standard dipteran gene order (insertion of tRNA into the genus Chrysomya and *D. hominis* raises gene count) (Fig. 2.3C(i)(ii),Table 2.1). In case of dipterans other than the Oestroidea superfamily, species like gall midge (Cecidomyiidae),

TH-2821_146106003

mosquitos (Culicidae), and crane flies (Tipulidae) exhibit various rearrangements in mitochondrial tRNA, such as the absence, inversion, translocation, and extreme truncation of certain genes[136,137].

2.3.5 Comparison among tRNAs:

The mitogenome of *Blepharipa sp.* has 22 tRNAs with a total length of 1466 bp, the relative positions of which are presented in the Figure 2.3C. These tRNA genes ranged in length from 63 to 72 bp and were found all across the mitogenome of Oestroidea superfamily. There are 14 tRNAs on the +ve strand and 8 tRNAs on the (-) strand among the 22 tRNAs. Here in this section, we have extensively compared different tRNAs from different families of Oestroidea flies in terms of size (bp), match-mismatches in base pairs, sequence identity and Secondary structure. Additional tRNAs exit in few species were exempted from the analysis.

tRNA size, nucleotide content and identity comparison: We observed that *trnV*, which is situated between two rRNAs on the -ve strand, is the overall longest tRNA, with an average length of 71.97 bp; nevertheless, except for *C. chinghaiensis*, the highest length is 72 bp for 35/36 species (71 bp). The smallest tRNA found in Oestroidea mitogenomes is tRNA-Ala of *C. chinghaiensis* (56 bp); else, *trnR* has a general minimum length of 64.35 bp and 22/36 species have *trnR* lengths of 64 bp, while *trnR* of *S. melanura* has a length of 73 bp, making it one of the longest tRNAs found in this superfamily. The mean AT content of tRNAs of Oestroidea superfamily reveals that 11 tRNAs with high AT content (>75%) and out of them 6 tRNAs transcribed on +ve strand including the tRNAs with highest AT content were *trnE* (90.72%) and *trnD* (86.57%). In contrast, *trnK* (68.65%) and *trnM* (69.71%) also located in +ve strand and have the lowest AT content.

The pattern of nucleotide conservation at tRNA genes was clearly +ve strand or J strand skewed, and several tRNAs, such as *trnL2*, *trnK*, *trnG*, t*rnW*, reached 100% similarity for the Calliphoridae and Sarcophagidae families, and even *trnA*, *trnM*, *trnS2*, *trnL2*, *trnW*, *trnK*

TH-2821_146106003

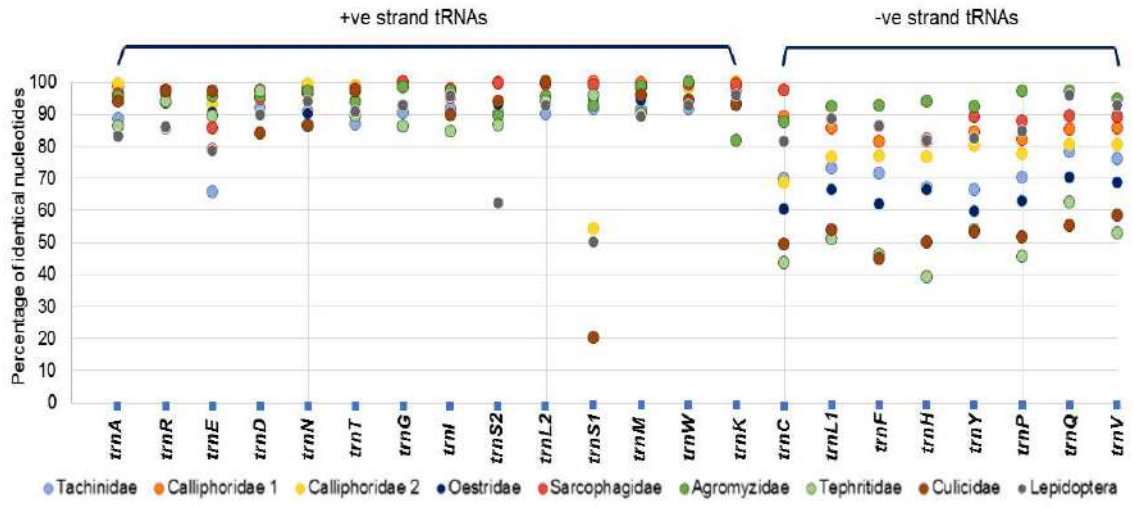


Figure 2.4: The percentage of identical nucleotides for each mt-tRNAs of different Oestroidea fly families.

showed overall %INUC > 95%. Except for *trnS1*and *trnE*, other 12 tRNAs of +ve strand overall exhibited over 90% identity. In contrast, out of 8 -ve strand tRNAs, 7 exhibited %INUC < 80% except trnQ; whereas, trnH (74.58%), which is transcribed from -ve strand, is one of the least conserved tRNA in Oestroidea as well in the other insect orders[138]. Other less conserved tRNAs (75%<%INUC>80%) are *trnF, trnY, trnP, trnC, trnL1, trnV* and all of these tRNAs transcribed on -ve strand. Only 3/22 tRNAs show %INUC ranges (80-90) %, include *trnS1, trnE, trnQ;* and *trnS1, trnE* located at the +ve strand just upstream of first -ve strand encoded PCGs (Fig. 2.3C, 2.4); where *trnQ* located between two +ve strand tRNAs (*trnI, trnM*) without using any intergenic spacer. The initial tRNA gene, *trnI*, showed overall 94% identity in Oestroidea flies, while the ending tRNA gene, *trnV*, showed overall 79 % identical; these two tRNAs, which are closest to the control region, were found to be among the most and least conserved tRNAs, respectively. Amongst different families in Oestroidea superfamily, Sarcophagidae family had the most conserved tRNAs, with 14/22 showing %INUC > 90%, whereas the Tachinidae family had just 5/22 showing more than 90% identity, lowest between the families of this superfamily. It is apparent from Figure 2.4, that -ve strand tRNAs of different family presented huge variation in their %INUC while in +ve strand except for *trnE, trnS1, trnS2*, showed strict

TH-2821_146106003

conservation among the species of any family. These 3 exceptional +ve strand tRNAs are located immediately upstream of -ve strand of the dipteran mitogenome (Fig. 2.4). Therefore, %INUC scores indicate a clear distinction between the nucleotide variety seen in Oestroidea flies, and various families exhibit different degrees of identity in their corresponding tRNAs.

tRNA structure comparison: Regardless of the level of conservation, some of the tRNAs offered mismatched pairs in their stems (e.g., A-A or U-U mismatches). These mismatches are common in many arthropod mitogenomes and might be adjusted through different types of editing processes[139]. These include: cytidine to uridine conversion, cytidine or uridine insertion, template-dependent editing of the first three nucleotides at the 5′ ends of tRNAs, and template-independent editing at the 3′ ends of tRNAs[140]. We identified 22 such mismatched base pairs in tRNA genes of *Blepharipa sp.* distributed over the AA stem, TΨC stem, AC and DHU regions of 15 tRNAs (Fig. 2.5). 13 mismatches were observed in weak G-U combination while 7 mismatches were observed in U-U combinations and 1 mismatch was found in each

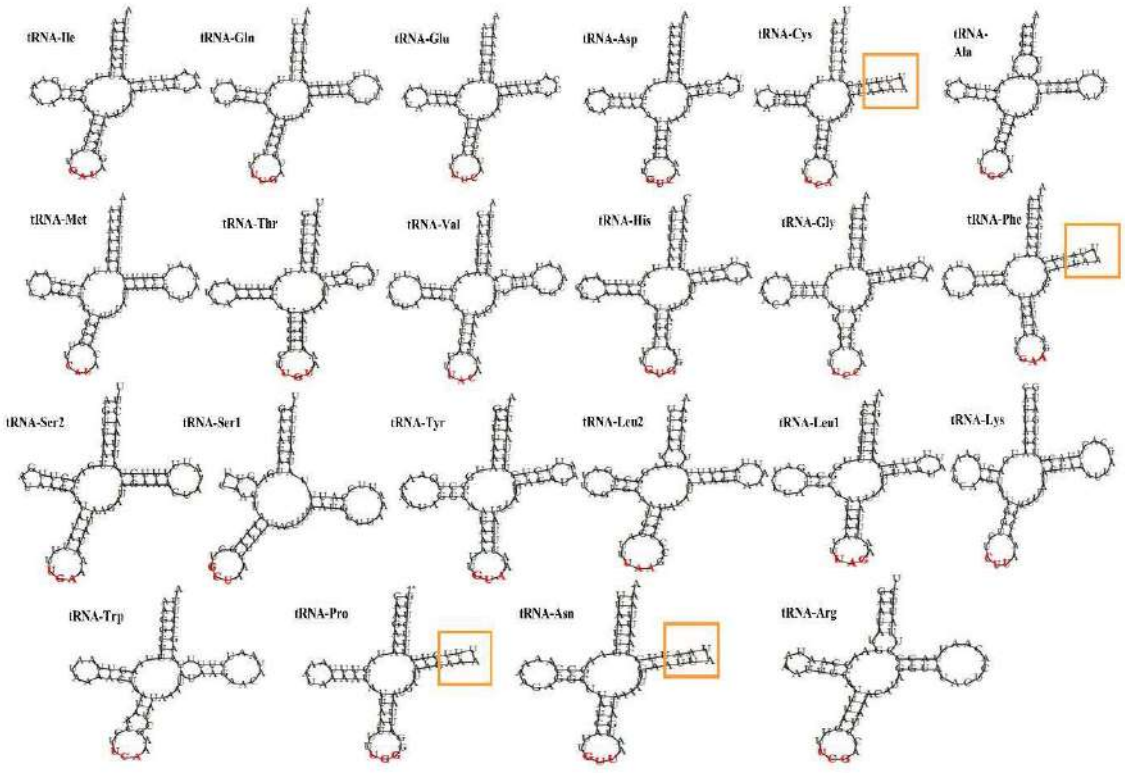


Figure 2.5: 22 tRNA structures encoded by *Blepharipa sp.* mitogenome. Red colour three letter signifies anticodon site and *trnC*, *trnF*, *trnP*, *trnN* lack stable TΨC loop denoted by Yellow box.

TH-2821_146106003

G-A and A-C combinations respectively. Maximum 3 mismatches found in acceptor stem of *tRNA-Ala* including 2 U-U mismatches and a G-U mismatch.

Comparative analysis of tRNA secondary structures with other two tachinid flies (*E. flavipalpis* and *E. sorbilans*) suggest that few of the mismatches, regardless of bond strength, are found to be conserved among the species of this family. Most common of them is G-U pair and it is visible at DHU arm of *trnQ*, *trnY, trnG, trnH, trnP, trnV* and at anticodon stem of *trnT* (not present in *R. goerlingiana*). While U-U, a strong mismatch pair is evident at the anticodon arms of *trnG*, acceptor arms of *trnR* and TΨC stem of *trnV* genes of Tachinidae family. Further comparison with other selected dipteran species supports that the G-U mismatch at DHU stem of *trnQ, trnY, trnG* (also present in both out-group Lepidoptera)*, trnH, trnP* (present in both out-group Lepidoptera) and also the U-U mismatch at acceptor arms of *trnR* (except *A. culicifacies*); TΨC stem of *trnV* are common and conserved among the selected dipteran flies.

The tRNA secondary structures of this uzifly mitogenome displayed a typical clover-leaf structure with the exception of *trnS1* that lacks a proper DHU stem. However, loss of this stem in *trnS1* is a common feature to many insect mitogenomes including other dipteran species[78,141]. Some other variation in the tRNA structures of *Blepharipa sp.* was also detected such as tRNAs carrying Cys, Phe, Pro and Asn amino acids do not have proper TΨC loop. The incomplete TΨC loop of *trnN* is also evident in Oestroidea flies like *R. goerlingiana* and non-Oestroidea flies like *B. minax*. Similarly, *trnF* genes of *E. sorbillans, D. hominis*, *S. albiceps*, *B. minax* and *B. mori* does not have complete TΨC loop.

2.3.6 Control region (CR) of *Blepharipa sp.* and comparison with Oestroidea:

This region in the metazoan mitogenome is a single large non-coding sequence that holds essential regulatory elements for the initiation of transcription and replication and is therefore named as control region (CR)[142,143]. Similar to other Diptera, CR of *Blepharipa sp.* is also flanked by *rrnS* and *trnI-trnQ-trnM* gene cluster (Fig. 2.2). Sequence similarity with other

TH-2821_146106003

Oestroidea superfamily species indicates that this segment is extremely variable in nature owing to the lack of coding constraints[144]. The CR sequence of *Blepharipa sp.* 75.49% similar to another tachinid fly *E. flavipalpis* followed by *C. bezziana* (71.15%). Despite the region's overall high nucleotide variation, this region necessarily contains a variety of repetitions (e.g., tandem repeat, inverted repeats)[71,145] and conserved structures namely Poly-T stretch (15 bp), [TA(A)]n-like, G(A)nT-like stretches, and poly A tail (15 bp)[146–148](Fig. 2.6A). Another conserved motif "ATTGTAAATT" we found in the CR of *Blepharipa sp.* and *E. flavipalpis* (Fig. 2.6A). Such conserved structures are considered to be involved in the regulation process of transcription or replication, and to maintain the initiating mode after binding with RNA polymerase by inhibiting the transition to elongation mode without altering its open-complexity[149,150].

The CR is also known as the AT-rich region because it has the highest proportion of A/T nucleotides (91.4 % for *Blepharipa sp.*) of any region in the mitogenome. The Tachinidae family has higher A+T content relative to other groups with the highest levels at Mulberry uzi fly, *E. sorbillans* (98.10%) and AT poor CR regions identified in *G. intestinalis* (80.80%) and *G. pecorum* (80.82%) (Oestridae)[71]. In this study CR of thirteen species have above 90% A+T content, and the top 3 are the tachinid flies, led by *A. grahami*, *D. hominis* and *Blepharipa sp.* consecutively. The CR is prone to high mutation and yet, the substitution rate is low due to high A+T content and directional mutation pressure[134,148]. This region varies greatly in length among insects, ranging from 70 bp to 13 kb and it accounts for most of the variation in mitogenome size[147]. The CR size of 36 Oestroidea flies ranges from 89 bp to 1750 bp, of which 16, 12 and 8 species can be categorized as large (> 800 bp), medium (200-800 bp), and small (< 200 bp) CR respectively, and *Blepharipa sp* (168 bp) falls under small category. while the shortest CR is present at *A. grahami* that might explain its small mitogenome size which is the smallest in this superfamily (Fig. 2.3B). The mean GC content of <200 bp CR is 8.84% which is less than (medium: 11.83%, large: 12.04%) the species with longer CR length (Table 2.1).

TH-2821_146106003

The average GC content of Tachinid flies' CR is 6.46%, with a mean CR length, 234 bp, and two tachinids, *E. sorbilans*, and *E. flavipalpis* have 105 bp long CRs which is relatively smaller than other reported flies, and their GC contents 1.9 % and 7.9%, respectively[33].

2.3.7 Overlapping sequence (OL) and intergenic spacer (IGS) regions:

The overlapping sequences (OL) and intergenic spacers (IGS) are frequently found in the mitogenome of dipterans. These two regions could vary in length and location from species to species during the course of evolution[151]. We identified 10 overlapping sequences in muga uzifly mitogenome with longest 8 bp OL between *trnW* and *trnC* genes (Figure 2.2). Other significant OSs are between *atp8* or *atp6* genes and *nad4* or *nad4l* genes of both 7 bp length are very common in insect phylum as well[152,153]. It has been previously reported that an "ATTATAA" motif present between *nad4* and *nad4l* in the –ve strand of the most of the insect mitochondrial genome, exceptions also exist, because of the direct adjacency of *nad4* and *nad4l*. Another characteristic 7 bp overlap motif of "ATGATAA" present between *atp8* and *atp6*, located on the +ve strand appears in most species of Symphyta and other insects[154]. These three overlapping sequences exist 35/36 Oestroidea flies. The region between *trnW* or *trnC* genes of *G. pecorum* and the regions between *atp8* or *atp6* and *nad4* or *nad4l* genes of C. *vomitoria* do not have OL (Fig. 2.6B). We noticed that total 30 types of OLs were present in different gene boundaries in mt-DNA of 36 flies and the count of OLs are varied from 4 to 21. Where, *C. vomitoria* have the minimum number of OLs and *S. crassipalpis* have the highest and total nucleotides in overlapping regions varies from 16 bp in *C. vomitoria* to 102 bp in *D. hominis*. The mitogenome arrangement reveals that the largest overlapping sequence of 64 bp situated over inserted *trnV* between *trnK-trnD* cluster at +ve strand of *D. hominis* (Figure 2.3C).

Mitogenome of *Blepharipa sp.* have 15 IGSs, which are distributed through PCGs, tRNAs and rRNAs. It has only one major IGS of more than 20 bp, the IGS 1 between *trnE* and *trnF* (40 bp); three medium size IGS of more than 10 bp, IGS 2 between *trnS1* and *trnE* (19 bp), IGS 3 at *trnS2-NAD1* (16 bp), and IGS 4 at *nad5-trnH* (15 bp); length of the rest 11 IGSs are below

TH-2821_146106003

10 bp. To a lesser extent, with a 5 bp conserved motif (ATCWW) at IGS 1 and 7 bp conserved motif (TWYTTMA) at IGS 4, present in several dipteran insects. In addition, IGS 3 has been reported to comprise a 7 bp conserved motif (ATACTAA) across Lepidoptera and 5 bp (TACTA) motif conserved across Coleoptera.

The analysis of comparative variation in intergenic spacer (IGS) regions reveals that IGS' varies across the whole Oestroidea superfamily in terms of length, location, and number of occurrences. A total 29 kinds of IGSs were found in 36 gene boundaries of 36 Oestroidea fly species. The number of intergenic spacer regions in any mitogenome ranges from 9 to 18, with *Lucilia coeruleiviridis* having the most and *Sarcophaga crassipalpis* having the fewest. The length of IGS between *trnS2* and *nad1* of *Hypoderma lineatum* or Common cattle grub is 102 bp that is found to be longest IGS present in mitogenome of this superfamily and this species also have the highest number of total nucleotides (174 bp) in its spacer region. The entire size and numbers of spacer regions are unevenly distributed over the mitogenomes of Oestroidea flies. We found that 11/19 Calliphoridae, 3/4 Oestridae, 4/9 Sarcophagidae, and 2/4 Tachinidae have total IGS with more than 100 bp length. We also found a unique 70 bp IGS in between *trnN* and *trnE* of *C. chinghaiensis*, we speculate that it might be the reason of sudden translocation of *trnS1* gene which lead to a creation of a long spacer region on that organism (Figure 2.3C).

Among the all IGSs, only 5 of them (*trnL2-cox2*, *cox3-trnG*, *trnE-trnF*, *nad5-trnH*, *trnS2-nad1*) were present in every Oestroidea species considered in this analysis. The average length of these 5 IGS regions are 5.30 bp, 6.22 bp, 18.63 bp, 14.33 bp, 18.52 bp respectively for 36 Oestroidea flies. Moreover, 4 IGSs of them form conserved motif, *trnE-trnF,* (ACTAAHWWWAATTMHHWA), *nad5-trnH* (WGAYADATWYTTCAY), *trnS2-nad1* (TACTAAAHHHHAWWMH), *cox3-trnG* (HTAAYT) in Oestroidea superfamily and were observed to be in similar location of other Diptera (Fig. 2.6C)[76]. The *trnS2-nad1* spacer is a typical feature of insect mitogenomes and is likely to contain the DmTFF, a bidirectional

TH-2821_146106003

transcription termination factor[155,156]. We also found seven unique IGSs that are seen in any single Oestroidea flies: *nad2-trnW*, *trnK-trnD*, *trnD-atp8*, *trnN-trnE*, *trnH-nad4*, *trnT-trnP*, and *trnV-srRNA*. This study shows that Chrysomya genus contains one extra copy of *trnI* at their CR region where *C. chinghaiensis* and *D. hominis* witnessed translocation and insertion of *trnS1* and *trnV* respectively which also leads to the formation of a unique spacer and overlapping region at their consecutive area.

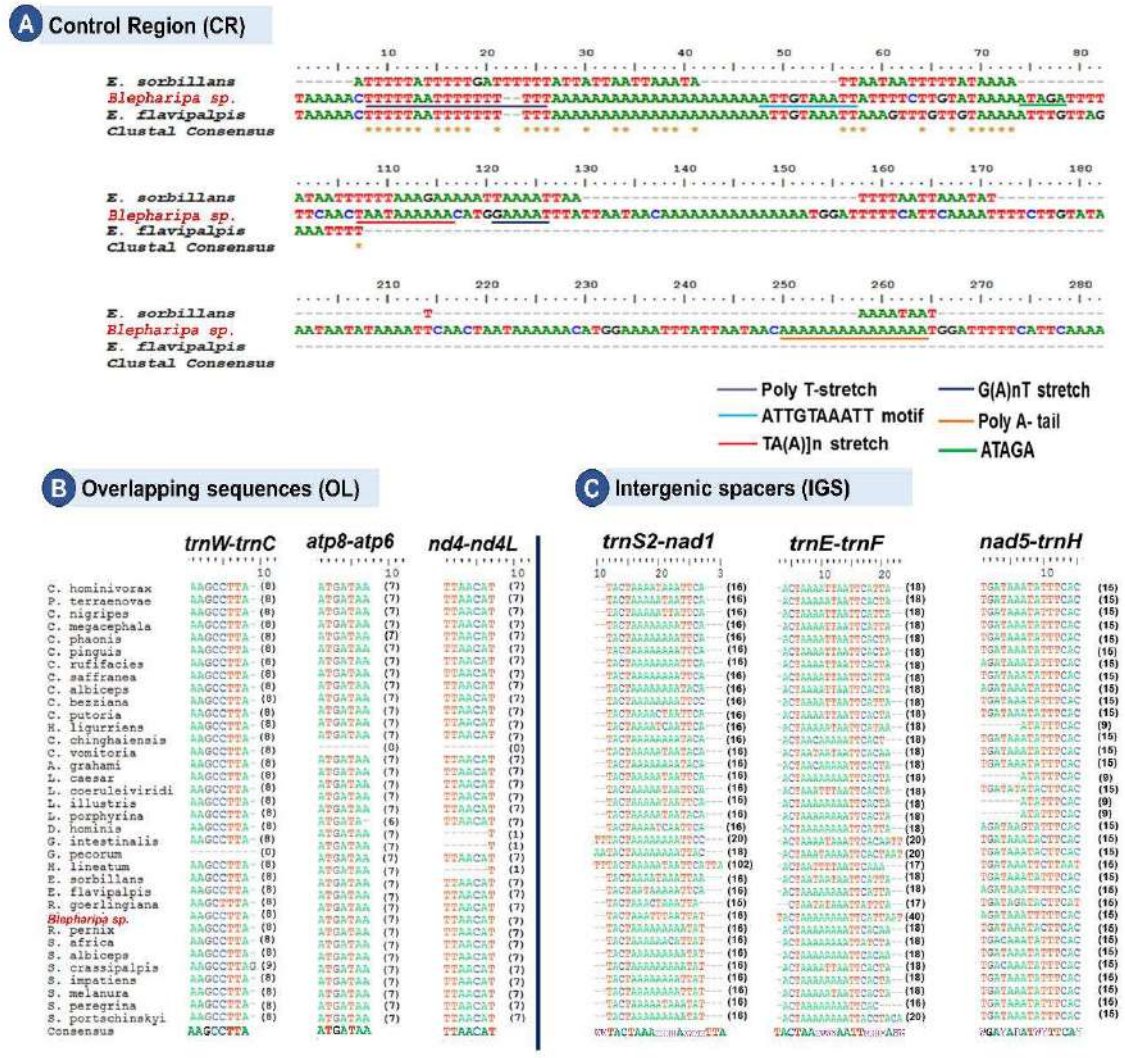


Figure 2.6: A) AT rich control region Alignment of *Blepharipa sp.* with other two Tachinidae species. B) Three alignments of the common overlap region between *trnW-trnC*, *atp8-atp6* and *nad4-nad4l*. C) Three alignment of the consensus gap region between *trnS2-nad1* (TACTAAAHHHHAWWMH), *trnE-trnF* (ACTAAHWWWAATTMHHWA), *nad5-trnH* (WGAYADATWYTTCAY) genes of all 36 Oestroidea mitogenome (Where, W= A/T, H=A/T/C, Y=T/C, D=G/T/A, M=A/C).

TH-2821_146106003

2.3.8 A comparison among Oestroidea mitochondrial protein coding genes (PCGs):

The *Blepharipa sp.* entire mitogenome contained recognized all of the 13 PCGs which spanned 11,166 bp constituting around 74.31 % of the total mitogenome. The total size of PCGs and their individual sizes were found to be similar to the other Oestroidea species. The PCGs were distributed over both the strands of double stranded mitogenome, 9 PCGs in the +ve strand and 4 PCGs in the –ve strand and it is also true for other Oestroidea flies.

Start codon and stop codon: The initiation of total 11 PCGs of *Blepharipa sp.* mitogenome follow standard "ATN" start codon except cox1 and nad1, translation of these two proteins started by TCG and TTG start codon respectively. The "ATN" family start codon found in this organism include ATA (*nad3*), ATT (*nad2*, *nad5*, *nad6*), ATG (*cox2*, *cox3*, *cytb*, *atp6*, *nad4*, *nad4l*), ATC (*atp8*). TCG used as very common additional start codon for *cox1* gene of Diptera as well as in Oestroidea species which was previously accepted as the canonical start codons for invertebrate PCGs (Fig. 2.7A)[157]. Other than *Blepharipa sp.*, TTG as start codon only presence in *nad1* gene of another tachinid fly *E. flavipalpis* and human bot fly *D. hominis*[33]. This study revealed that there are a total of eight alternative start codons present in Oestroidea PCGs, with 90.40% of them beginning with the four conventional "ATN" codons and the remaining 9.6% beginning with non-canonical codons such as TCG, TTA, TTG, and GAG (Fig. 2.7A). Among the Oestroidea mitogenomes sequenced to date, ATG (*Met*) is the most often used start codon followed by ATT (*Ile*). ATG (*Met*) is exclusively used in the *cox2*, *cytb*, *atp6*, *nad4l*, *nad4* and *cox3* genes of almost all Oestroidea species (ATT served as start codon of nad4 gene of *E. sorbillans* and *cox3* gene of *C. chinghaiensis* and *L. illustris*). The start codon ATT frequently used in PCGs like *nad2*, *nad6*, and *nad5* of Oestroidea flies (Fig. 2.7A). Although non-canonical start codons are not rare in insect mitogenomes, GAG was reported as the start codon for the *nad1* gene for the first time in *R. goerlingiana* (Tachinidae)[4,33,157–159]. Hence, the genes like *nad1*, *atp8*, *nad3*, and *nad6* utilize three to six distinct start codons. The remaining PCGs, begin with no more than two, and typically just one, start codon (Fig. 2.7 A).

TH-2821_146106003

The most commonly used stop codons in *Blepharipa sp.* are TAA (*atp6, atp8, cox3, nad1, nad2, nad3, nad4l, nad6*), TAG (*cytb, nad5*) remaining *cox1*, *cox2*, and *nad4* genes of this newly sequenced mitogenome do not have complete stop codon, instead they have "T-" which are supposed to be completed by the addition of 3' A residue to the mRNA during post-transcriptional polyadenylation (Fig. 2.2)[160]. The existence of such incomplete stop codon in these three genes also presence in other fly from this family such as *R. goerlingiana* and *E. flavipalpis*. In addition, our comparative analysis supported the idea that incomplete stop codons (denoted as T or TA) are accepted for those PCGs where a tRNA is immediately flanked at 3' end [161]. For instance, *trnL2* gene is adjacent to *cox1* gene of 34 flies and created overlap

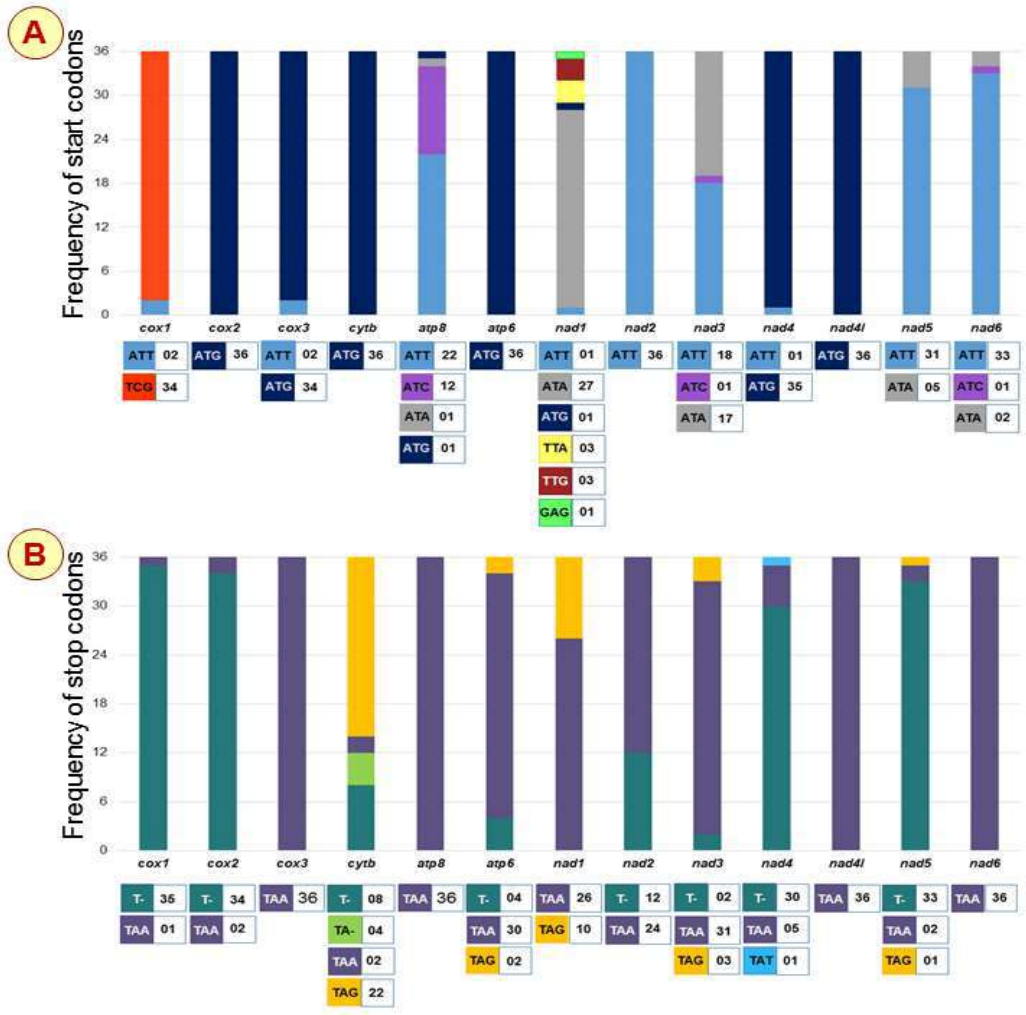


Figure 2.7: Usage of start and stop codons in complete Oestroidea mitogenomes. (A) Start codons usage of 13 PCGs in Oestroidea. (B) Stop codons usage of 13 PCGs in Oestroidea.

TH-2821_146106003

sequence in one fly and finally ended with incomplete stop codon T-, whereas, *E. sorbillans* forms a 2 bp spacer between these two genes and terminated with a compete TAA stop codon. Similar scenario could be observed in other PCGs as well, 34 flies of Oestroidea have incomplete stop codon at *cox2* and 33 of them are just adjacent to *trnK, L. caesar* forms a 3bp overlapping region; while *E. sorbillans* and *H. lineatum* have 5 bp and 4 bp spacer region and accordingly have proper stop codon of *cox2.* This kind of occurrences also observed in our sequenced mitogenome of *Blepharipa sp.* (Fig. 2.2B). In this study, we also revealed that T-, the partial stop codon, is the second (33.76 %) most commonly used stop codon after TAA (57.05 %), and TAA is only found in genes like *cox3*, *atp8*, *nad4l*, and *nad6*. Other than that, a number of other stop codons like TAG, TA-, TAT is also used in different genes like *cytb, nad1, nad3, atp6* and *nad1* of this superfamily (Fig. 2.7B).

Nucleotide composition and comparison: The PCGs of Tachinid fly, *E. flavipalpis* has the highest A+T content (79.09%) followed by the uzi flies *E. sorbillans* (77.64%) and *Blepharipa sp*. (77.28%) (Table 2.1). However, the nucleotide bias in individual PCGs has moved towards higher use of Thymine rather than Adenine, and this trend is observed in Diptera as well as Oestroidea flies' PCGs. The J strand PCGs and N strand PCGs show that both the gene sets are moderately T skewed (-ve AT skew); while the J strand gene set are moderately C skewed, N strand is strongly G skewed and, a similar kind of pattern is also observed in other insects [162]. The 4-fold degenerate codons do not influence the amino acid selection. Whereas, 2-fold degenerate codons are restricted to change their 3rd position for the presence of two-fold redundant codon positions. Codon redundancy arises due to change in a nucleotide in 2nd codon position accounts for 6 fold codon degeneracy[163]. We calculated A+T/G+C content and skew for all codon positions of Oestroidea flies (Fig. 2.8). We found that 3rd codon positions are rich in A+T content (Highest mean in *nad4l*: 92.85 ± 4.17%, lowest mean in *cytb*: 88.08 ± 4.59 %; n=36) than other positions (1st codon position AT%: highest mean in *atp8*: 79.18 ± 4.18%, lowest mean in *cox1*: 57.45 ± 1.55%; 2nd codon position AT%: highest mean in *nad6*: 75.63 ±

TH-2821_146106003

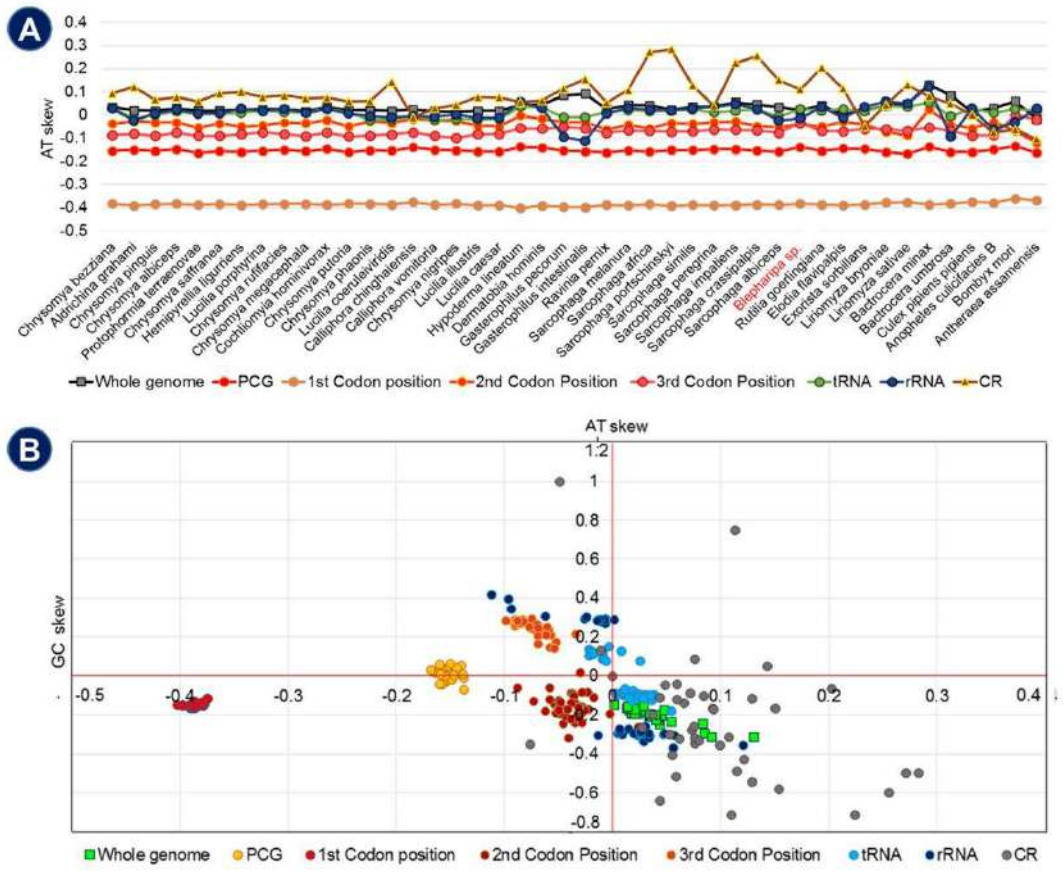


Figure 2.8: (A) Trend of AT skew across the Oestroidea superfamily and outgroups. (B) AT skew vs GC skew of different genetic position of 44 organisms (CR shows maximum variation and 1st codon position shows least variation).

1.37%, lowest mean in *cox1*: 59.57 ± 0.41%). In the case of Tachinidae (n=4) A+T content in different codon positions are (3rd codon position AT%: highest mean in *nad3*: 94.87 ± 2.1%, lowest mean in *atp8*: 91.67 ± 5.73%; 1st codon position AT%: highest mean in *atp8*: 86.12 ± 2.28%, lowest mean in *cox1*: 58.84 ± 1.09%; 2nd codon position AT%: highest mean in *nad4l*: 60.45 ± 0.41%, lowest mean in *cox1*: 59.57 ± 0.41%) enriched with higher A+T content than other species and so the 3rd codon positions and has also documented by other research[164]. It is also clear that the standard deviation of the 2nd codon position is quite low, whereas the standard deviation of the 3rd codon position is the highest among other codon positions. This may reflect the prevalent 4-fold degeneracy of codon and the frequency of codon usage variation in different species. The PCGs have the most conserved AT and GC skewness in the sample set,

TH-2821_146106003

eventually forming four distinct clusters for complete PCGs, 1st, 2nd, and 3rd codon positions. It indicates the lowest AT skew (-ve) at the 1st codon position, which is consistent across and beyond the Oestroidea superfamily (Fig. 2.8B). It also suggests that the abundance of Ts and Cs (Pyrimidine) is higher in the 1st and 2nd codon positions than As and Gs (Purine) respectively, and the 3rd codon location shows the abundance of As over Ts and Cs over Gs. It appears to apply to all Oestroidea flies. According to the GC skew analysis, the –ve GC skew value is fairly consistent with other dipteran insects, except for a few lower Diptera[165].

Synonymous Codon usage pattern: In general, synonymous codons that code for the same amino acid do not present at the same frequency in protein-coding genes[166,167]. Differences in synonymous codon usage bias can be seen in a wide variety of species, ranging from prokaryotes through unicellular and multicellular eukaryotes[168–170]. Since various genomes possess typical patterns of synonymous codon usage, thus the comparative RSCU analysis facilitates the understanding of evolution and adaptation of living organisms[171,172]. The genetic code of mitochondria as well often differs from the standard genetic code[173]. We know that the pattern of codon usage changes over evolution, but we don't know how. A thorough comparative study was conducted to interpret the pattern of codon usage in the mitogenome of Oestroidea flies; six additional Diptera species and two Lepidoptera moths were also described.

The RSCU vs amino acid and codon graph reveals that the mitochondrial protein coding genes of *Blepharipa sp.* are biased towards A/U ending codons, which account for about 92.07% of all codons. Codon UCU responsible for coding of amino acid serine is most frequent codon (4.40%, RSCU= 2.73) and CUG does not code amino acid in any mitochondrial protein coding genes of this organism. Assessment with Oestroidea species shows a stringent favor towards codons containing A or U at 3rd position but the ranges of biasness vary largely from 97.31% (*E. flavipalpis*) to 75.41% (*G. intestinalis*). Gasterophilinae subfamily of Oestroidea flies (*G. intestinalis* and *G. pecorum*) had the lowest A+T concentration at the third codon position, with 75.4 and 76.9 %, respectively. According to this study, CGA and UCU are the most often

TH-2821_146106003

utilised codons by PCGs of the Oestroid flies, but PCGs of the Sarcophagidae family solely use UCU as the most common codon for Serine, and the Oestridae family employs UCA together with UCU for coding the same amino acid. Except for *H. ligurriens* (UCU), the Calliphoridae family employs CGA as the most common codon for Arginine coding. However, the most used codon in mitochondrial PCGs of the Tachinidae family varies; the Uziflies employ UCU, whilst the other two tachinid flies (*R. goerlingiana* and *E. flavipalpis*) prefer CGA and CUU for Arginine and Leucine coding, respectively. The relative synonymous codon use (RSCU) values of nearly half of the codons (29/62) suggest that they are commonly utilised codons (RSCU>=1) (except for termination codon). The comparison of RSCU values of 62 sense codons shows that there's some difference in the RSCU values of total sense codons from 42 dipteran mitochondrial proteins, but the overall trend is relatively similar and there is a clear separation of A/U and G/C ending codons visible in the cluster analysis (Fig. 2.9A). This demonstrates that substantially similar species have stable codon usage patterns.

When the codon usage value for a certain codon increase, the synonymous codons decrease, indicating the presence of a greater bias. This demonstrates that similarly related species preserve the stability of codon usage behavior; as the use of one particular codon grows, the use of other synonymous codons decreases, reflecting a stronger bias. For instance, Lysine (K) encoded by AAA and AAG in insect mitochondrion. The Tachinidae flies show more favoritism towards AAA codon, while other Oestroidea flies are biased towards AAG for coding the same amino acid. Moreover, Tachinidae have eighteen such predominant A/U ending codons (GUU, ACA, UGA, CAA, UGU, UUA, AUU, UUU, AAU, AUA, AAA, CAU, UAU, GAU, AGA, GCA, GGU, CGU) which show higher codon usage than other families and seven out of them consist completely A/U nucleotides (Fig. 2.9B). Thus, similar to other invertebrate species, the individual RSCU evaluation of all thirteen PCGs reveals a general tendency towards codons with A or U at 3rd place[82].

TH-2821_146106003

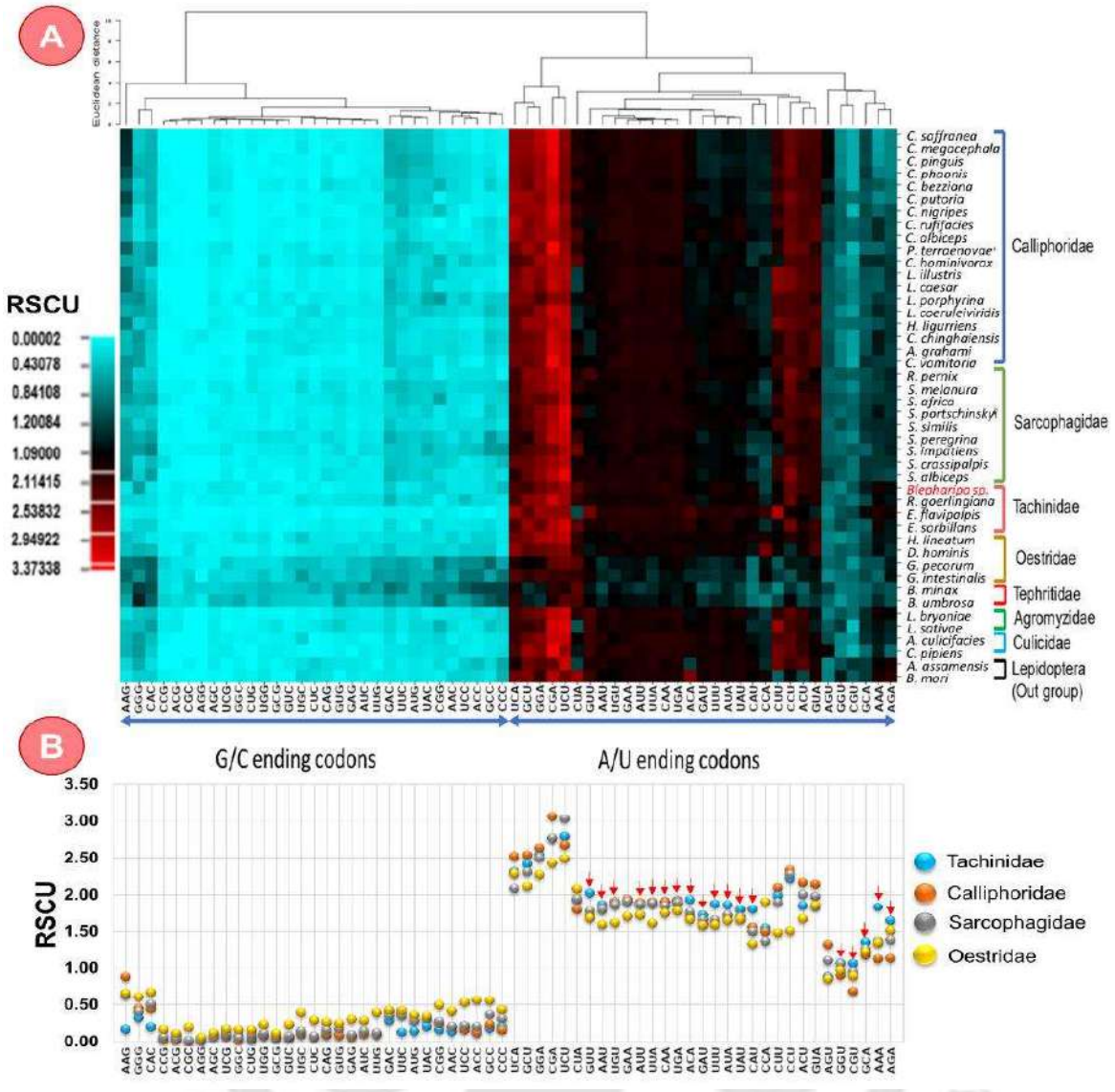


Figure 2.9: (A) RSCU Cluster analysis of 36 species from Oestroidea Superfamily, 6 organisms from other Diptera and 2 organisms from out group (Lepidoptera). Termination codons are excluded. The heat-map was drawn with CIMminer. Bigger RSCU values, suggesting more frequent codon usage, are represented with darker shades of red. (B) Family wise Average RSCU value plot of 62 codons.

Exploration of more trends of codon usage in each gene, we measured effective number codons (ENc) in all PCGs of our test species. The ENc values range from 20 (just one codon allocated to each codon family), which indicates extreme codon bias, to 61 (equal usage of all synonymous codons), which indicates no codon bias at all[115]. In mitochondrial context, every PCG is essential and in the absence of adequate evidence on gene expression, ENc plays a valuable role in determining codon bias[174]. Analysis shows ENc values of each PCGs (n=13*36) of Oestroidea flies varied from 30.11 (*nad5* gene of *E. sorbilans,* strong bias) to

TH-2821_146106003

49.19 (*atp8* gene of *G. intestinalis,* weak bias). If the ENc value of any gene is closer to 20, it implies that the gene has a very strong codon bias, and many studies have shown that ENc < 35 indicates a mostly high codon bias[115,116]. On the other hand, if the ENc value of any gene nearer to 61 denotes extremely weak bias so we believe that ENc > 45 should denote relatively weak codon bias. The family-wise mean ENc value of mitochondrial PCGs is depicted in Figure 2.10B. The most biased gene observed in this superfamily is *nad5* (Mean ENc: 34.15 ± 2.36) followed by *nad4* (Mean ENc: 34.62 ± 2.12), *nad1* (Mean ENc: 35.56 ± 1.79), and *cox1* (Mean ENc: 36.02 ± 2.31) with relatively strong codon bias for every family except families like Oestridae of Oestroidea superfamily and Tephritidae (Fig. 2.10B). The least biased gene is *atp8* (Mean ENc: 45.20 ± 1.96) followed by *nad3* (Mean ENc: 41.34 ± 2.02) and *nad4l* (Mean ENc: 42.45 ± 1.85) genes of every family of Oestroidea including Tachinidae family exhibit relatively weak codon bias (ENc > 40). In addition, the mean ENc values of each mitochondrial PCG of Tachinidae flies are lower than those of other Oestroidea flies, suggesting that the mitogenome of this family has a higher codon bias (Fig. 2.10B).

Relation between nucleotide composition and codon usage: The nucleotide composition has a strong correlation with codon usage in the Oestroidea superfamily as well as other dipteran mitochondria. The Tachinidae family exhibits lower mean ENc values across the PCGs, indicating a greater codon bias at the gene level (Fig. 2.10B). As evidenced by RSCU analyses, all 13 PCGs are skewed toward A/T, resulting in codon usage biases (Fig. 2.9A). Correlation among 3rd codon position and relative synonymous codon usage value pointed out that total RSCU value of the codons with A/U at 3rd codon position is inversely proportional to the GC3 content and directly proportional to the total codon usage value when G/C at third nucleotide position ($p<0.001$). For example, *G. intestinalis*, a horse botfly shown in orange colour, has the highest GC3 content and has less biased codon usage among the PCGs. *E. flavipalpis* have the lowest GC3 content among the Oestroidea flies and display relatively stronger codon bias

TH-2821_146106003

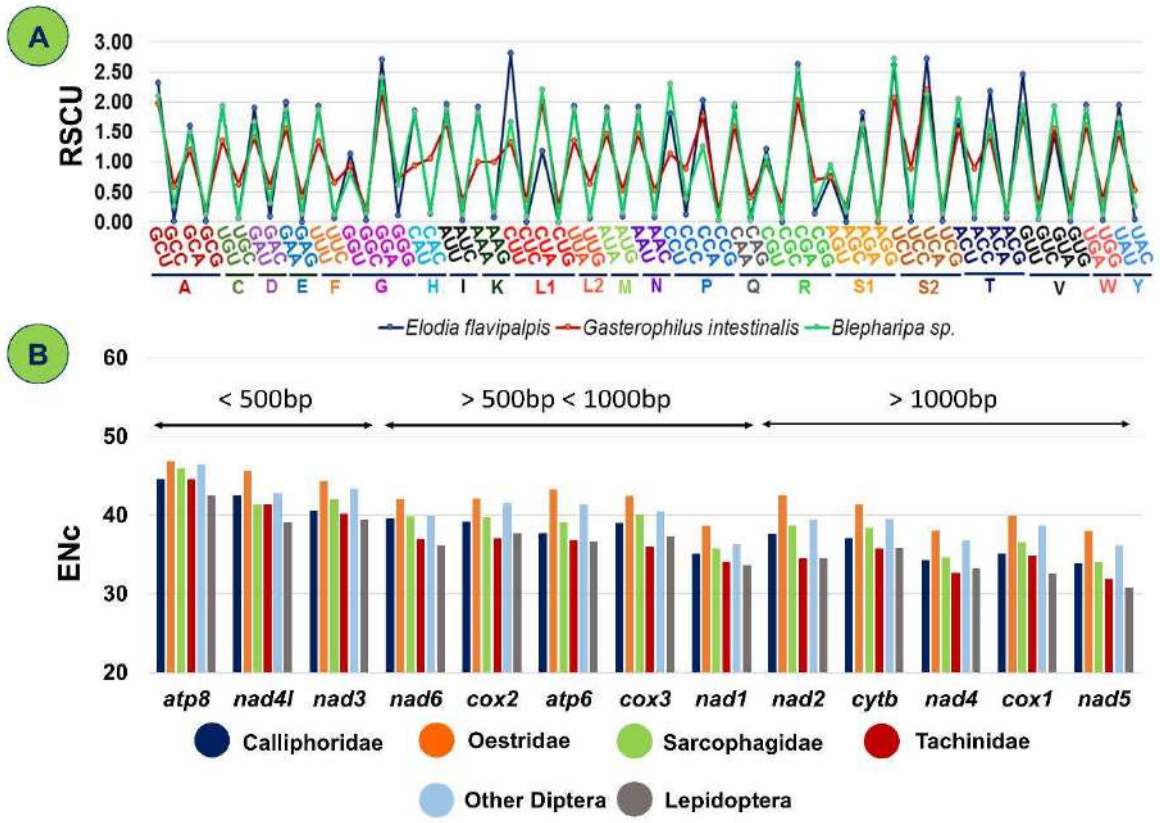


Figure 2.10: (A) RSCU value comparison between *E. flavipalpis* (Maximum A/U at 3rd codon position), *G. intestinalis* (Minimum A/U at 3rd codon position) and *Blepharipa sp.* (B) Average Effective codon number (ENc) of 13 PCGs of different families of Oestroidea flies and out groups.

and low ENc value (Fig. 2.10A). Similarly, *Blepharipa sp.* mitogenome also shows very less GC3 content and has comparatively stronger codon bias (Fig. 2.10A). The Pearson correlation results reveal that ENc has a significant positive correlation with the GC content at 3rd codon positions of PCGs (GC3, R = 0.374, p <0.01 and GC3s, R = 0.374, p <0.01), and on the other hand other codon positions, particularly GC1 and GC2, have a weak but significant negative correlation with ENc (GC1, R = -0.121, p<0.01 and GC2, R = -0.112, p<0.01). This indicates that by increasing GC content at 3rd codon position the ENc values of the genes also increase and as a consequence codon usage bias will decrease in Oestrodea mitogenome since insect mitochondrial genomes are rich in AT content (Fig. 2.9)[33,71].

TH-2821_146106003

ENc-plot for determining the factors of codon usage bias: For a better understanding of nucleotide composition and codon usage bias, ENc values were plotted against the GC3s values in ENc-plot where the standard curve demonstrates the functional relationship between ENc and GC3s is under mutation pressure rather than selection[117]. The plot suggests that if the codon usage bias is entirely dependent on GC3s, all of the points would be precisely on the standard curve (corresponding to the ENc values)[117,121]. In this study, the majority of the points in this plot do not lie close to the standard curve, indicating that the role of GC3s in mutation bias is not the key factor for the formation of the codon bias (Fig. 2.11A). The ENc-plot depicted that some points lie on or nearer to the curve (on or above: *atp6, cox2, cox3*, *nad6*; both sides of the curve: *nad2* and on or below the curve: *cox1, cytb, nad1, nad4, nad5*) but some of the points situated far away from the curve (above the curve: *atp8, nad3*, *nad4l*) indicative of variation in codon usage bias and their causes. While PCGs like *cox1, cytb, nad1*, and *nad2* of *Blepharipa sp.* are located nearer to the curve; *atp6, cox2, cox3, nad6* genes are situated slightly above the curve; *nad4, nad4l* positioned below the curve and *atp8, nad3*, *nad6* located far above the curve. Therefore, this outcome infers that along with mutation pressure for shaping codon usage bias in different species, some independent factors, like natural selection strongly influence the bias pattern and these factors are more dominant than mutation pressure[175].

Neutrality test for determining the factors of codon usage bias: In order to measure the degree to which directional mutation pressure is neutral to selection in the codon usage bias of mitogenome, the neutrality test was carried out as ENc-GC3s could not estimate precisely which of mutation pressure or natural selection was more essential[118,121]. According to the theory, nucleotide heterogeneity is the effect of bidirectional mutation pressures between G/C and A/T pairs, and this pressure induces directional changes more in neutral parts than in functionally significant parts[118,176]. In this analysis (GC12 vs GC3) regression slopes of 13 PCGs substantially deviated from the diagonal line (regression coefficient < 1; lowest 0.1149

TH-2821_146106003

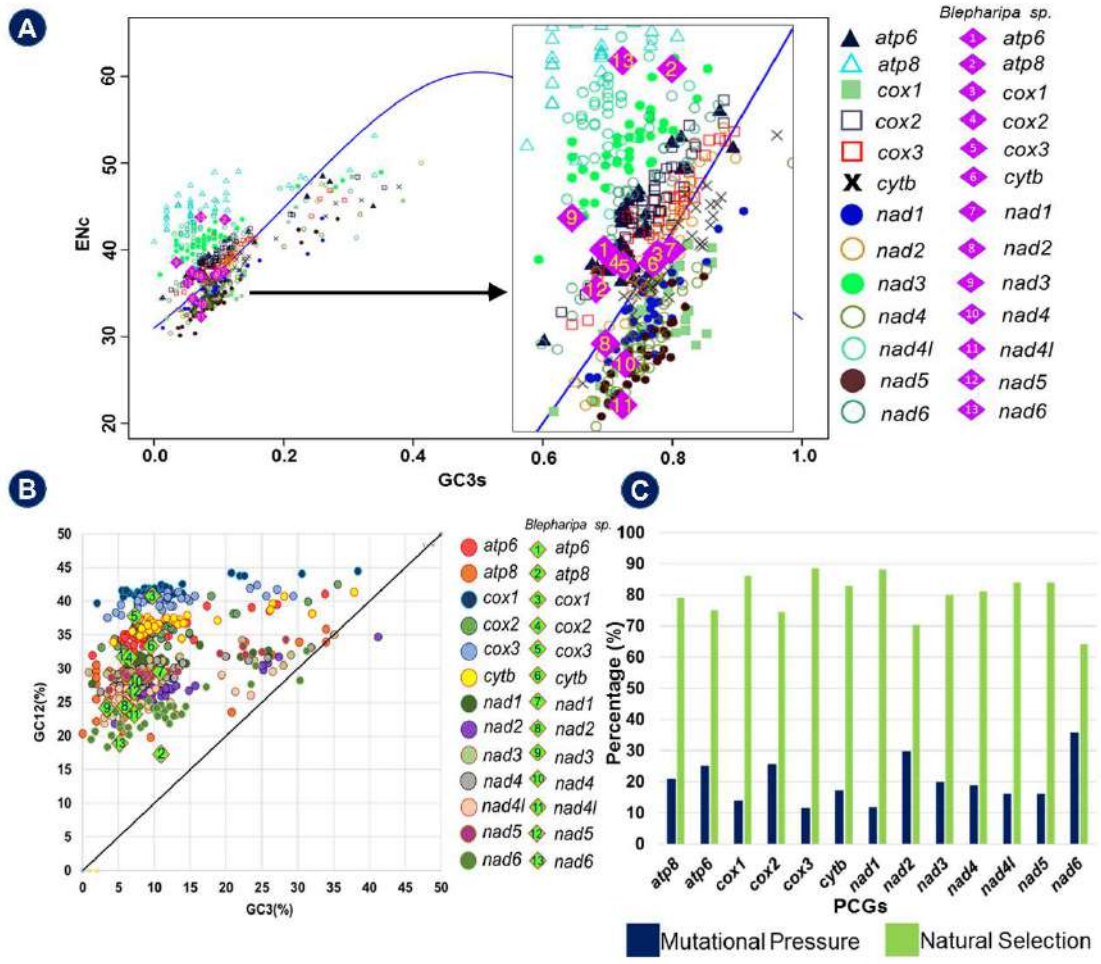


Figure 2.11: (A) The ENc vs. GC3s plots of Oestroidean mitochondrial protein coding genes. The standard curve ENc = 2+GC3s +29/ [GC3s2+(1-GC3s)2] represents the expected ENc to GC3s. (B) Neutrality plots (GC12 vs. GC3) of 13 PCGs of 42 species. GC12 stands for the average value of GC content in the first and second position of the codons (GC1 and GC2). While GC3 refers to the GC content in the third codon position (each dot signifying a gene). (C) Probability of selection pressure on each PCGs of Oestroidea. The regression line of all PCGs denoted by y= mx+c (Where, Mutational Pressure (M) =m*100, Natural Selection (N) =100-M).

(*cox3*) to highest 0.3563 (*nad6*)) by contributing a significant but weak positive correlation ($R^2 < 0.9$; P-values < 0.01) between observed GC12 and GC3 (Fig. 2.11B, data not shown). The plot suggests that relative neutrality of GC12 varies from 11.5% in *cox3* to 35.6% in *nad6* as compared to GC3 (100% neutrality or 0% constraint) in the mitogenome of Oestroidea superfamily[120]. It also indicates that intensity of mutation pressure is weakest in *cox3*, accounted for only 11.5% and highest in *nad6* accounted for 35.6% towards neutrality. It was observed in this study that low and narrow distribution GC content of Oestroidea varies from

TH-2821_146106003

20.03% to 29.83% in WMG and 20.9% to 32.1% in PCGs and it has never exceeded 50% of the total nucleotide content. The variation and scarcity of GC content in 3rd position of codon (e.g., GC3 of *cox3*: 3.43%-29.38% and of *nad6:* 1.72%-30.28%), and narrow distribution of GC12 content (e.g., GC12 of *cox3*: 37.59%-41.41% and GC12 of *nad6:* 18.4%-28.28%) also observed (Fig. 2.11C, data not shown). It was reported that selection against mutational bias can cause a narrow distribution of GC content and poor correlation between GC12 and GC3[177,178]. The predominance of natural selection together with some other minor factors accounted for almost 88.5% in *cox3* (highest) and 64.3% (lowest) in *nad6* relative constraint. Thus, it appears that the mitogenome of the Oestroidea superfamily retains a low and restricted distribution of GC contents owing to the selection against mutation bias[117,178].

As the Oestroidea mitogenomes were highly AT-rich (highest for *E. flavipalpis*, WMG: 79.96%, PCG: 79.06%; lowest for *G. intestinalis*, WMG: 70.16%, PCG: 67.88%) and prevalence of A/T ending codons (highest for *E. flavipalpis*, 3rd position: 72.72%, lowest for *G. intestinalis*, 3rd position: 63.41%) observed in this study. This is in line with the theory that the strong bias of the Oestroidea mitogenome's codon usage towards a large representation of NNA and NNT codons is due to mutational bias towards A/T, which was also documented for other mitochondrial genomes [117,177,179].

Relation between gene length and codon usage: Longer genes need more energy to improve accuracy by selecting such favorable codons which are able to minimize the proofreading costs, maximize the rate, and accuracy of translation[34,180]. This study shows that the smallest gene (*atp8*, mean length: 161.75 bp) has the highest mean ENc (45.20) and the longest gene (*nad5*, mean length: 1719.16 bp) has the lowest mean ENc (34.15) (Fig. 2.10B) among 13 mitochondrial PCGs of Oestroidea. The Pearson correlation statistics show a satisfactory and significant negative correlation of ENc with gene length ($R = -0.742$, $p<0.001$) (data not shown). It indicates that the length of mitochondrial genes in Oestroidea flies is inversely related to the effective number of codons (ENc), which ensures that as gene length increases,

TH-2821_146106003

ENc reduces and, as a result, codon usage bias increases. Thus, longer mitochondrial genes show stronger codon bias than smaller genes. This has also been found by Lei Wei et al. (2014) while studying *B. mori* mitogenome. Lei Wei et al. also argued that mitochondrial gene length and codon usage bias related to their expression level[117]. It has been widely known that highly biased codons are mainly observed in highly expressed genes and in mitochondria longer genes are highly expressed as well [34,117,180]. Our findings are in accord with previous studies in which prokaryotes like *E. coli* and *Yersinia pestis* exhibits a common trend of elevated codon usage bias for longer genes, unlike nuclear genes of multicellular eukaryotes namely Yeast and *Drosophila*, where smaller genes appear to be more biased than longer genes[34,117,180].

2.3.9 Phylogenetic inference:

Phylogenetic relation of Oestroidea superfamily: Phylogenetic relationship through 13 mitochondrial protein coding genes represents very similar topology in both Bayesian Inference (BI) and Maximum Likelihood (ML). It established a link among major clades with very good support from Bayesian posterior probability and moderate bootstrap support from ML analysis. Adjacent grouping of *Blepharipa sp.* and *E. flavipalpis* with 100 percent bootstrap support and congruent support from Bayesian posterior probability (1.00) is evident within the monophyletic clade of *E. sorbilans* (BI/ML: 1.00/69) (Fig. 2.12 A, B). This study revealed that the two families namely, Sarcophagidae (1.00/100), and Calliphoridae (1.00/100) belong to the monophyletic group of the Oestroidea superfamily. The Calliphoridae family is distributed in two different clades, wherein a single clade *Chrysomya sp.* along with *P. terraenovae* (1.00/79) separated from other Calliphoridae flies (1.00/100) as found by other research as well [181]. While the Oestridae and Tachinidae families were unable to recover as monophyletic, they do form a paraphyletic connection with the rest of the tree. Though taxonomically *H. lineatum* belongs to the Oestridae family but our inference using both the method exhibits polyphyletic relation with Oestridae flies and clustered with *R. goerlingiana* from Tachinidae with 50% bootstrap

TH-2821_146106003

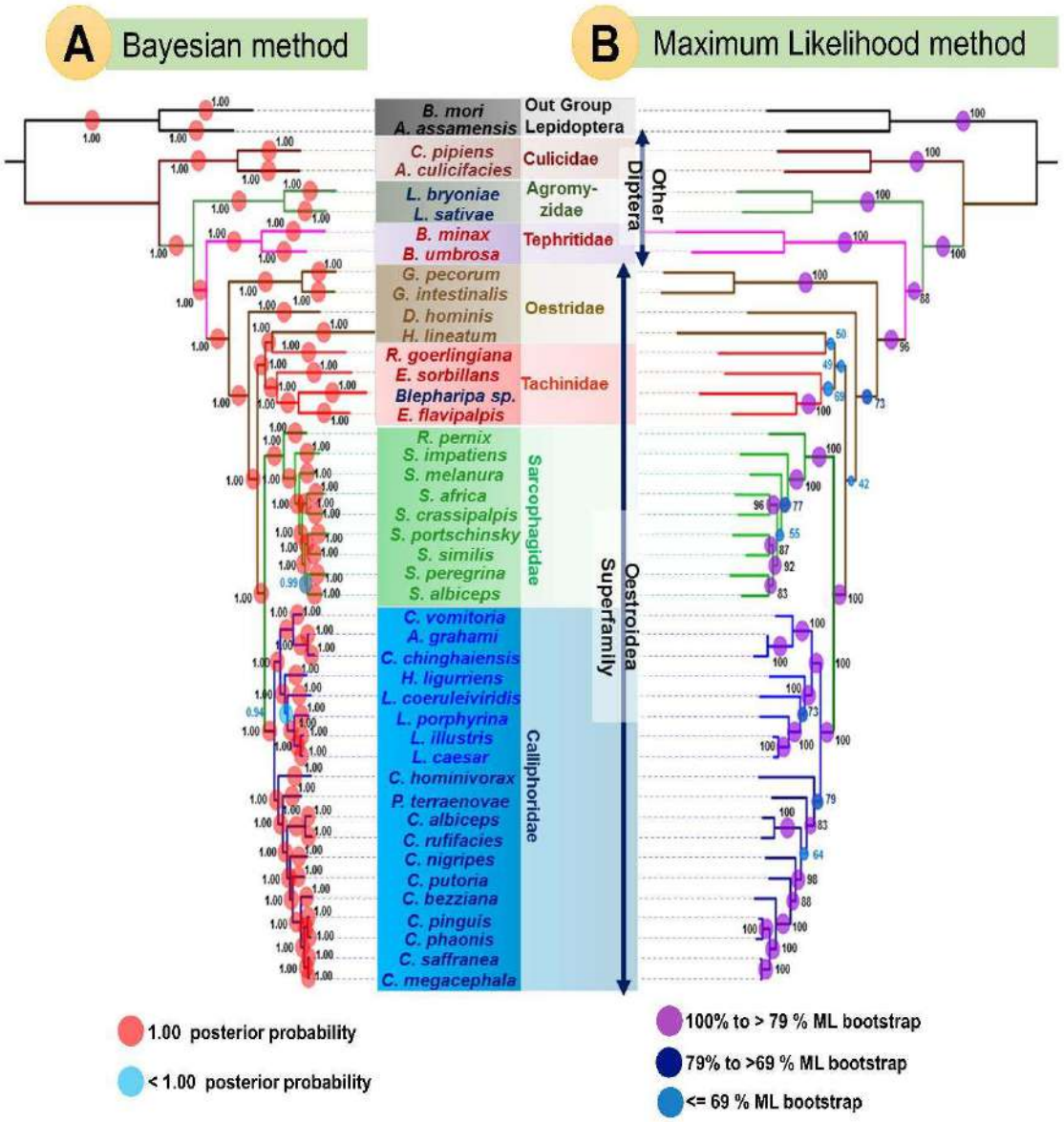


Figure 2.12: (A) Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 44 organisms (36: Oestridea superfamily, 6: other Diptera and 2: Out group Lepidoptera) using maximum likelihood (ML) method in RaxML 8.2.x (5000 bootstrap replicates). (B) Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 44 organisms (36: Oestridea superfamily, 6: other Diptera and 2: Out group Lepidoptera using Bayesian inference (BI) method in MrBayes v3.2.6.

support[182]. Therefore, with the exception of Calliphoridae our analysis establishes the monophyletic status of the Sarcophagidae, and Oestridae is established as the sister group of remaining Oestroidea flies via both ML and BI methods [181]. Both Lepidoptera sequences group together and are represented as outgroup for this analysis.

Location of Tachinidae at Oestroidea phylogeny: The relationship between the different oestroid lineages is still a point of controversy in Diptera phylogeny. Wiegmann et al. presented that the speciation of Oestroidea and closely related lineages were linked with higher rates of diversification[183]. This was made it hard to resolve relationships among these taxa, particularly concerning the origin of the Tachinidae family. According to the morphological and molecular evidence, nearly every other family of Oestroidea has been assigned as a potential sister clade of Tachinidae [32,191,192,193,194,195]. The common nature of the internal parasitism of the arthropods and subscutellum development, some poorly defined families (e.g. Rhinophoridae (not in this study)) were also been proposed as a sister group of tachinids[184,186]. However, this is less convincing than the reality that certain representatives of Calliphoridae, Sarcophagidae, and Oestridae do have sclerotized subscutellum[187]. Some sarcophagids are parasitoid with insects and other arthropods, while certain calliphorids are parasitoid with snails and earthworms[186]. However, greater diversity of feeding habits and breeding environments, including hematophagous parasitism of birds and mammals has been evident from these groups of species[183,186].

Tachinidae is the morphologically most heterogeneous subgroup of this superfamily, lacking clear morphological synapomorphy and usually serving as a dumping place for taxa with confusing characteristics[188]. Cerrati et al. in their morphological study present Tachinidae as polyphyletic, and the bulk of their subfamily exist as paraphyletic[189]. The several morphological or character synapomorphic states of these families and their classification in the tree leads us to conclude that it currently reflects the monophyly Sarcophagidae and Calliphoridae but we still haven't been able to recover the monophyly of Tachinidae and Oestridae as a clade. This discordance from conventional knowledge may be attributed to long branching of two genera and insufficient sampling of Oestridae and Tachinidae taxa. In other ways, this issue may indicate that, since these families seem to have experienced a significant

variation in both molecules and morphology, which has contributed to extremely developed parasitic behaviour and made other comparisons in characters' tough[181].

In this study, all flies exhibit parasitism in diverse forms, with the Oestridae family parasitizing mammals and the Tachinidae family parasitizing insects (Table 2.1). The phylogenetic inference reveals very little about the monophyly of Oestridae and Tachinidae using the combined 13 mitochondrial genes yet having substantial support from bootstrap and posterior probability may be due to phylogenetic inertia playing a major role in resolving true relationship. According to the physical law of inertia, an inertially moving body subjected to various forces will move in the direction of 'least resistance.' Similarly, the biological world obeys the same rule of inertia as the inorganic world, with evolutionary lineages following the path of least resistance, implying that evolution will continue in the direction of previously acquired adaptations despite environmental perturbations[190–192]. For example, the failure of birds to evolve viviparity[193], high altitude behavior in a valley population of a South American rodent despite half a million years of isolation[194]. In this scenario, we can say that parasitism might occur before formation of families like Oestridae or Tachinidae, and persistence of resemble characters or traits among species hinder in distinguishing phylogenetic relationship.

2.3.10 Nonsynonymous substitution:

The PAML package was used in this study to ensure whether there was any beneficial adaptation that occurred in the PCGs. Two different trees (gene tree and species tree) were used to estimate nonsynonymous to synonymous rate ratios (ω = Ka/Ks or dN/dS) through the maximum-likelihood method in each gene. dN/dS > 1 specifies positive selection, dN/dS =1 neutrality, and dN/dS <1 negative selection. First, a very simple model known as the one-ratio model (M0) was used, it allows a single ω ratio for all branches. The ω ratios that we estimated from 13 individual PCGs are all less than 1 for both the trees, facilitating enough support for the occurrence of negative selection acting on the mitochondrial genes. In this study, the gene *atp8* shows the highest ω value (gene tree: 0.11541, species tree: 0.12904), and *cox1* shows the

TH-2821_146106003

lowest ω value (gene tree: 0.02328, species tree: 0.02035) among the 13 mitochondrial PCGs (Table 2.2.Table 2.3). To retain the important mitochondrial functions in energy metabolism strong purifying selection plays an important role in the evolution of the mitogenome of Oestroidea flies.

Since insect endo parasitism was acquired only in tachinid flies thus, we assumed that there may have been some evolutionary pressure on this lineage. Therefore, the lineage belongs to *Blepharipa sp.* of the Tachinidae family with other members (if available in the same clade) considered as foreground lineage for branch specific two ratio model in two different trees. The two-ratio model using gene tree showed except *cytb* (ω0 = 0.03005, ω1=0.00010), ω for other 12 genes on the foreground branch (ω1) is greater than the background lineages (ω0) but not more than 1. The gene *nad5* (ω0= 0.04559, ω1= 0.92099) and *atp8* (ω0 = 0.11541, ω1= 0.93957) have maximum ω value for lineage of interest in the foreground branch. Through reference species tree the *nad2* gene (ω0 =0.08155, ω1=0.04346) exhibits low ω value at foreground branch than background branches and *nad4* (ω0 = 0.04972, ω1 = 0.20965) and *atp8* (ω0 = 0.05145, ω1= 0.20860) show maximum ω1 value. The log-likelihood difference, $2\Delta lnL = 2(l_1-l_0)$ between the one-ratio and two-ratio model presents that the two-ratio model fits better than the one-ratio model. Using gene tree, we obtained maximum 2ΔlnL from *nad5* gene (2ΔlnL = 27.01) with significant level $0.001 < p$ and df =1 and minimum from *atp8* gene (2ΔlnL = 0.00007199), which is comparatively very less significant ($p < 0.995$) than other genes. In case of species tree, we got maximum 2Δl from *nad6* gene (2ΔlnL = 1560) with significance level $0.001 < p$ and df =1 (Table 2.2,Table 2.3). Overall, in each tree's foreground or background branches, the ω ratio never exceeded 1, considering the fact that the branch leading to the uzi flies' common ancestor has gained more nonsynonymous mutations than synonymous mutations, and therefore putting more selective pressure on it than other branches. However, the possibility of relaxed selection cannot be excluded and thus assessment does not support positive selection on the foreground branch. Positive selection normally operates on a few sites

TH-2821_146106003

for a brief amount of evolutionary time, but the signal for positive selection is usually drowned out by the continuous negative selection that occurs on the majority of sites in a gene sequence[195]. The branch leading to the common ancestor of uziflies (Tachinidae) have seldom nonsynonymous mutations, indicating that most have been occupied by purifying selection.

Purifying selection cannot generate better genes it is only responsible for preserving the function of a gene[195]. The mitochondrial protein products are crucial for survival; thus, their activities are more restricted. Hence, it can be inferred that the numerous selection constraints present in codons effect their evolution through influencing transcription and translation efficiency[179].

2.3.11 Correlation between nucleotide substitution rates and codon usage indices

Regression analysis was performed to establish the correlation between nucleotide substitution rates (dS, dN and ω) and codon usage indices (GC3, GC3s, GC12, and ENc). Univariate models based on general additive model (GAM) fitted better compared to linear model (LM) and polynomial model (PM) where the coefficient of determination R^2 always performed better except for few cases where R^2 of LM and PM is equivalent to GAM. Other model evaluation factors, such as RMSE, Residual standard error, and AIC, are mostly attributed to improved GAM fit. The synonymous divergence rate (dS) fit better for predictor variables such as GC3, GC3s, and ENc, but the nonsynonymous divergence rate, dN fits better for GC12. This trend is observed in all regression models and is similar even after the logarithmic transformation of the response variable (Fig. 2.13, 2.14).

The ω does not fit well as compared to dS and dN but like dN, it fits better for GC12 than other predictor variables. In addition, model fitting improved after log transformation of dS and dN, but model fitting degraded after transformation of ω. The ratio of nucleotide substitution at nonsynonymous and synonymous sites is defined as dN/dS or ω, and certain genes display a very low nonsynonymous substitution rate than the corresponding synonymous substitution

TH-2821_146106003

TH-2821_146106003

Table 2.2: Branch-specific assessments of selective pressure on the common ancestor of *Blepharipa* sp. for 13 PCGs using species tree

Gene	atp6		atp8		cox1		cox2		cox3		cytb		nad1		nad2		nad3		nad4		nad4l		nad5		nad6	
Model	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio:	One ratio: ω0 = ω1	Two ratio:	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio
p	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
lnL	-10728.41	-10728.27	-2846.23	-2846.1	-18325.59	-18324.83	-8540.17	-8539.99	-10261.12	-10260.18	-16291.46	-16290.17	-11861.26	-11859.95	-17686.9	-17686.56	-5547.05	-5546.88	-19670.25	-19666.66	-3689.06	-3689.05	-26541.42	-26536.9	-9940.92	-9160.92
2ΔL	0.27		0.26		1.51		0.37		1.86		2.58		2.62		0.66		0.35		7.17		0.01		9.04		1560	
k	1.2428	1.2428	1.4452	1.4441	2.0646	2.0646	1.8173	1.8171	1.7784	1.7784	1.6482	1.6477	1.2662	1.2654	1.1891	1.1893	1.0915	1.0915	1.0996	1.0994	0.9595	0.9594	1.3272	1.3262	0.9082	0.9705
ω0	0.0417	0.0416	0.1290	0.1283	0.0204	0.0203	0.0333	0.0332	0.0366	0.0363	0.0349	0.0347	0.0324	0.0322	0.0814	0.0816	0.0492	0.0491	0.0501	0.0497	0.0479	0.0479	0.0519	0.0515	0.0809	0.0616
ω1	ω0	0.0632	ω0	0.1949	ω0	0.0445	ω0	0.0561	ω0	0.0962	ω0	0.0996	ω0	0.1206	ω0	0.0435	ω0	0.0941	ω0	0.2097	ω0	0.0530	ω0	0.2086	ω0	0.1558

Table 2.3: Branch-specific assessments of selective pressure on the common ancestor of *Blepharipa* sp. for 13 PCGs using gene trees

Gene	atp6		atp8		cox1		cox2		cox3		cytb		nad1		nad2		nad3		nad4		nad4l		nad5		nad6	
Model	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio:	One ratio: ω0 = ω1	Two ratio:	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio	One ratio: ω0 = ω1	Two ratio
p	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
lnL	-10190.90	-10190.18	-2616.65	-2616.65	-18758.10	-18755.80	-8441.42	-8441.26	-10452.84	-10452.15	-15331.56	-15330.10	-11526.89	-11526.70	-16834.79	-16827.77	-5148.74	-5144.74	-19292.43	-19288.11	-3619.28	-3618.77	-25675.98	-25662.47	-9162.08	-9160.92
2ΔL	1.43		7.2E-05		4.59138		0.30588			1.380836		2.915314		0.388064		14.0383		7.999406		8.647452		0.5		27.01589		2.319474
k	1.2491	1.2494	1.5980	1.5980	1.9602	1.9596	1.8929	1.8926	1.7613	1.7612	1.6743	1.6740	1.2693	1.2692	1.2044	1.2040	1.2007	1.1994	1.1077	1.1072	0.9940	0.9942	1.3515	1.3519	0.9707	0.9705
ω0	0.0376	0.0375	0.1154	0.1154	0.0233	0.0230	0.0339	0.0338	0.0402	0.0400	0.0299	0.0301	0.0302	0.0301	0.0665	0.0651	0.0417	0.0403	0.0481	0.0476	0.0511	0.0503	0.0464	0.0456	0.0622	0.0616
ω1	ω0	0.0939	ω0	0.9396	ω0	0.0780	ω0	0.0464	ω0	0.0983	ω0	0.0001	ω0	0.0464	ω0	0.1546	ω0	0.1090	ω0	0.2090	ω0	0.0785	ω0	0.9210	ω0	0.1558

kappa (Ts/Tv, Transition/ Transversion) = κ

omega (dN/dS, nonsynonymous/synonymous) = ω

One ratio model = One nonsynonymous/synonymous rate ratio (ω) for all lineages of reference phylogenetic tree;

Two ratio model = Two ω value, ω1 for foreground lineage of interest and ω0 for background linages of reference phylogenetic tree;

lnL = Log likelihood of the model estimation; p = Number of parameters

rate, resulting in very small ω (0.0001) for those genes. This makes a separation of data by extremely small and substantially larger values of ω, and following log transformation, such data creates bimodal distribution. As a result, the data distribution deviated from normality and eventually worsen the model fitting.

In general, 3rd positions of 4-fold degenerate codons act as a silent site or synonymous site where a change in nucleotides does not change the resultant amino acids. The codon usage indices like GC3, GC3s denote GC content at 3rd codon positions of all codons and 4-fold degenerate codons respectively. It has also been observed that GC3, GC3s, and divergence rate at silent sites are negatively correlated, and according to GAM, this association is not linear. It means that the reducing synonymous divergence rate, dS, and increasing GC content at 3rd codon positions is not uniform across mitochondrial genes of various species. GAM also depicts nonsynonymous divergence rate (dN), which declines with increasing GC content at 3rd codon positions in a manner almost similar to dS, although model fitting of dN with GC3 and GC3s is inferior to dS. After log transformation of dS and dN, the relationship curve with ENc seems nearly similar, and the same is true for GC3 and GC3s, indicating that the relationship pattern with those codon usage indices does not change after log transformation of nucleotide substitution rate. The GC12 represents the average GC content of 1st and 2nd codon positions, which are typically regarded as nonsynonymous sites where nucleotide alterations influence the amino acid composition. According to GAM, both dS and dN decrease as GC12 increases, but dN has a persistent wiggle in its pattern despite fitting GC12 better than dS. The log transformed dN displays a decreasing but wiggly relationship with GC12, whereas the log transformed dS shows a decreasing trend with little wiggle at higher GC12. This implies that when GC12 grows, both dS and dN decrease, but dS rate decreases more smoothly than dN.

TH-2821_146106003

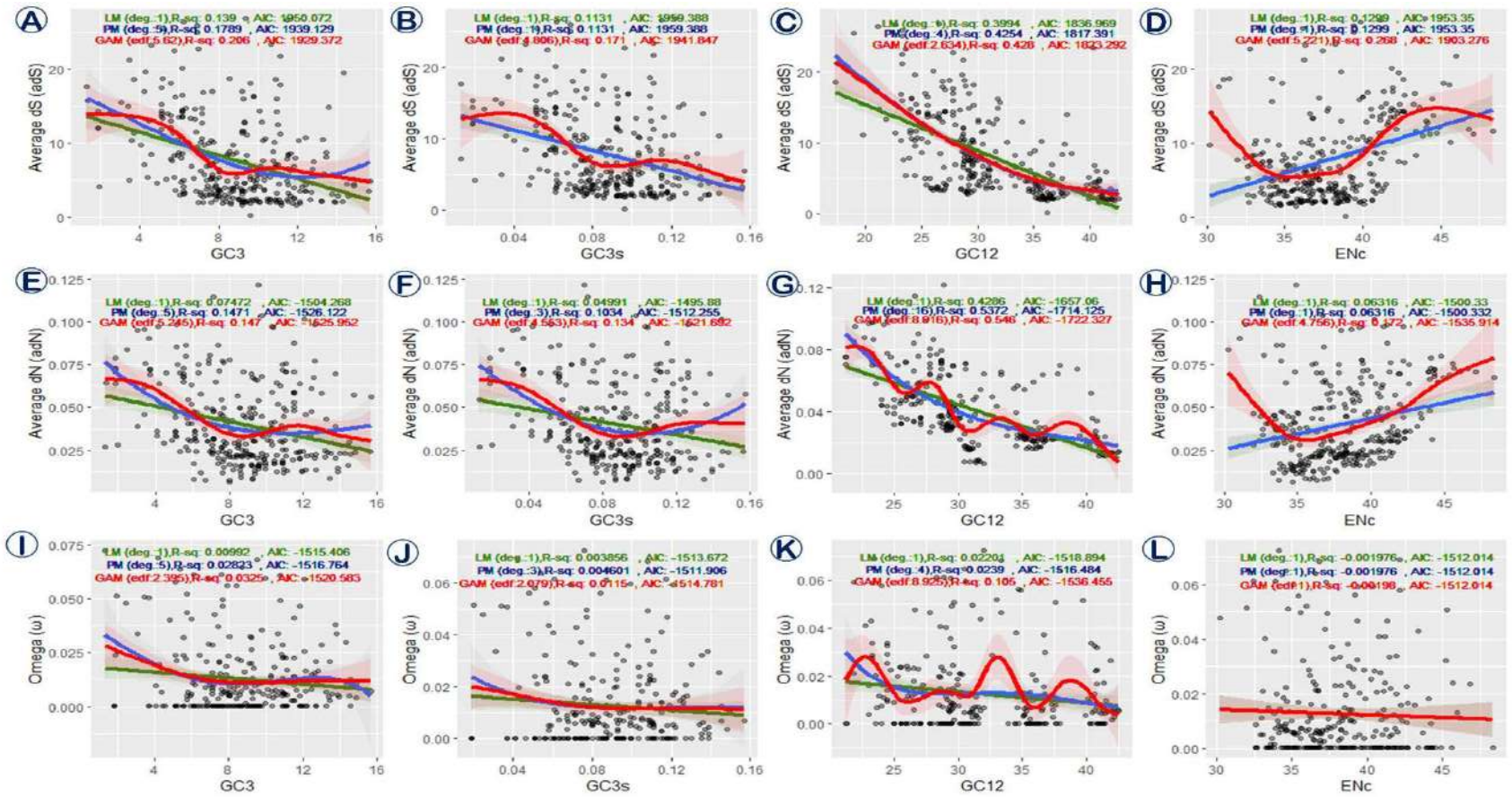


Figure 2.13: Univariate regression model fitting between response variables, divergence rate (dS, dN and ω) and predictor variables, codon usage indices (GC3, GC3s, GC12, ENc) of training datasets; (A-D): average synonymous divergence (adS) rate vs codon usage indices; (E-H): average nonsynonymous divergence rate (adN) vs codon usage indices; (I-L): omega ratio (ω) vs codon usage indices; Green: Linear Model (LM), Blue: Polynomial Model (PM), Red: Generalized Additive Model (GAM), light colour represent the 95% confidence interval; R-sq (R^2): Coefficient of determination; AIC: Akaike information criterion; deg: degree, edf: effective degrees of freedom.

TH-2821_146106003

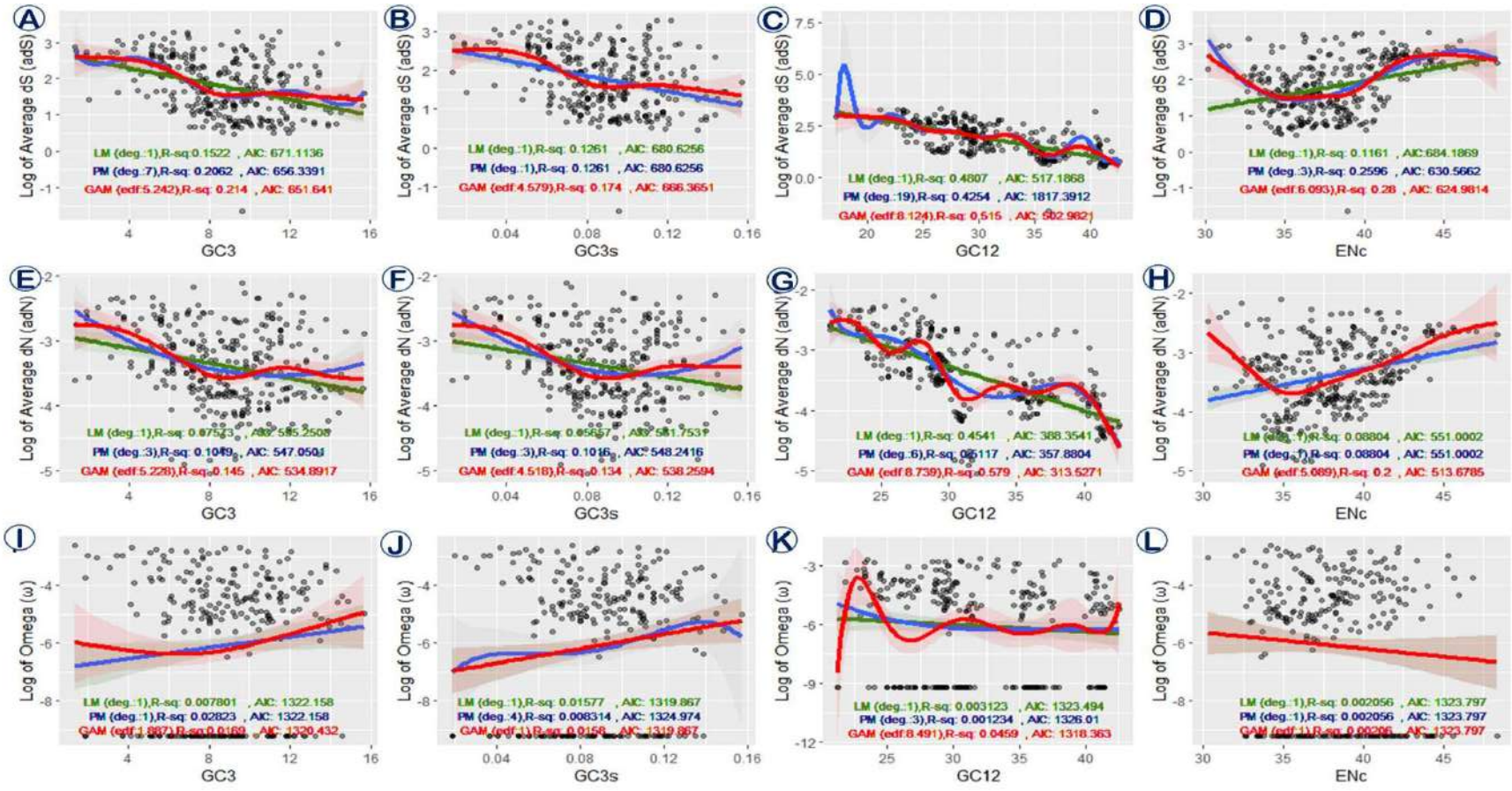


Figure 2.14: Univariate regression model fitting between logarithmic response variables, divergence rate (dS, dN and ω) and predictor variables, codon usage indices (GC3, GC3s, GC12, ENc) of training datasets; (A-D): log of average synonymous divergence (adS) rate vs codon usage indices; (E-H): log of average nonsynonymous divergence rate (adN) vs codon usage indices; (I-L): log of omega ratio (ω) vs codon usage indices; Green: Linear Model (LM), Blue: Polynomial Model (PM), Red: Generalized Additive Model (GAM), light colour represent the 95% confidence interval; R-sq (R^2): Coefficient of determination; AIC: Akaike information criterion; deg: degree, edf: effective degrees of freedom.

Overall, dS drop with increasing GC3 and GC3s with a significant decline at 4-8% GC content at the 3rd codon position and dS is considerably higher than dN (Fig. 2.13). Both dS and dN appear to create curves that are almost opposite of the S-curve with GC3 and GC3s. Since mitochondrial genes and their 3rd codon positions are strongly AT biased, divergence rates might increase with rising AT concentration at 3rd codon positions. Although the relationship would not be linear, it will follow an S-shaped curve, with a spike in divergence rate occurring at genes with 92-96 % AT content at 3rd codon positions. The ENc designates the effective number of codons of any gene when ENc increases codon usage bias decreases. The GAM of both dS and dN depicts a valley shape curve, implying that the rate of nucleotide change at synonymous and nonsynonymous sites initially declines, remains steady, and then slowly increases as ENc increases. It indicates that when codon usage bias reduces, dS and dN drop drastically at first, then stabilize for a time before gradually increasing.

The independent and uneven distribution of codon usage indices of mitochondrial genes across species complicates formal analysis using standard statistical linear models. Simpler linear correlations are generally employed with minimal consideration for assumption violations and do not offer estimates of the magnitude of change, instead of focusing on whether or not there is a linear or monotonic trend[196]. Alternative techniques have been expanded to allow for more complicated nonlinear trends by having response variables rely on predictor polynomials. The fully parametric model has some flaws, most notably poor fitting or overfitting of the data and the behavior of the fitted trend at the beginning and end of the observed series[197,198]. Whereas, GAMs employ automated smoothness selection methods to establish the complexity of the fitted trend objectively, and they allow for potentially intricate, non-linear trends as well as adequate accounting of model uncertainty[196].

2.3.12 Codon usage bias and parasitism:

According to our findings, Tachinidae is the only obligate insect endoparasite family in the Oestroidea superfamily with significantly AT biased PCGs and a high concentration of AT in

codons of their mitogenome. Interestingly, the Gasterophilinae tribe of the Oestridae family, which is an internal parasite of mammals, has the lowest A+T content, and its clade is phylogenetically split before other Oestroidea flies diverged (Fig. 2.12,2.15)[181]. Tachinidae flies, a sister clade of the Oestridae family, as well as two other Oestridae flies, *H. lineatum* and *D. hominis*, have AT-rich genes. Despite being in the same lineage as *R. goerlingiana*, *H. lineatum* is an external parasite, whereas *D. hominis* is an endoparasite. Apart from that, other Oestroidea flies included in this study all show ecto-parasitism (Table 2.1). Consistent with the tendencies seen in the base composition, we found that the Tachinidae's codon usage was biased toward high AT content (Fig. 2.9, 2.10). As a result, a link between AT ending codons and ENc is found in the Tachinidae family, where the codon usage of AT ending codon is higher, but the effective number of codons (ENc) is lower, as shown in the contour map colour scheme (Fig. 2.15A). We also conducted a principal component analysis of concatenated 13 PCGs RSCU values using covariance matrix and correlation matrix where Tachinids are distinguishable from the rest of Oestroidea flies through the first two principal components (PC1 and PC2) of both the matrices (Fig. 2.15 B, C).

It has been frequently reported that synonymous changes in codons do not alter the protein sequence but it can have a substantial impact on protein levels, folding, translation efficiency, and gene expression of other organisms[29,31,199–201]. The ENc and neutrality plots revealed that directional mutations and selection forces had a role in shaping Oestroidea's mitogenome during evolution. The AT-rich mitochondrial PCGs of endo-parasite tachinid flies are the consequence of mutational bias towards A/T ending codons, whereas natural selection has shaped biased codon usage in PCGs by constraining GC content (Fig. 2.9, 2.11). Although the narrow distribution GC3, the deviation of points from the ENc standard curve in the ENc-plot, and the ω value of less than 1 in non-synonymous substitution analysis all imply that mutation bias is not the primary factor driving codon bias[117,179,195]. This study also infers that PCGs have high purifying selection due to a higher synonymous divergence rate than nonsynonymous

TH-2821_146106003

divergence rate. Tachinidae flies have a much higher AT concentration in the 3rd codon position than other species, with the lowest AT level being 87.9% in *nad6* of *R. goerlingiana* and the highest being 100% in *atp8* of *E. flavipalpis*, resulting in a less effective number of codons (data not shown). Purifying selection appears to be the major evolutionary force, as it efficiently eliminates harmful adaptive changes in amino acids and reduces the effects of adaptive selection pressures at the codon level, despite the fact that directional mutations caused considerable AT-usage bias. According to a previous study on mitogenome evolution, strongly locomotive species rapidly eliminate detrimental non-synonymous substitutions, indicating that they were subjected to intense purifying selection to maintain effective respiratory-chain activity[202]. During this process, Synonymous substitutions were maintained, and directional mutations resulted in the use of specific types of codons being used more frequently[179,202]. Altogether it leads to high usage of synonymous codons in Tachinids than that of other Oestroidea flies. Hence, efficient energy production by mitochondria of Tachinids needs fewer codons, that might assists in maintaining their gene expression, translation, or further protein folding and function[28-31,203,204].

Various hosts react differently to parasitic infections, and while arthropods or insects do have innate immunity, they show very little adaptive immune response in comparison to mammals[205]. Larva of other Oestroidea flies included in this study are generally necrophagous, saprophagous or sarcophagus ectoparasite and feed on carrion and carcass of a broad range of vertebrates (Table 2.1)[206]. On the other hand, tachinid flies are typical internal parasitoids with specialized in their host choice (insect)[207,208]. During larval stages, the feeding maggots constantly tackle the host defense mechanism that builds up a highly stressful environment for the larvae[145,207]. Additionally, Tachinids have to thrive as endo-parasites in highly dioxic environments by adopting a unique respiration strategy[2,71]. We postulate that Tachinidae flies have naturally selected for limited GC content and purifying selection to preserve

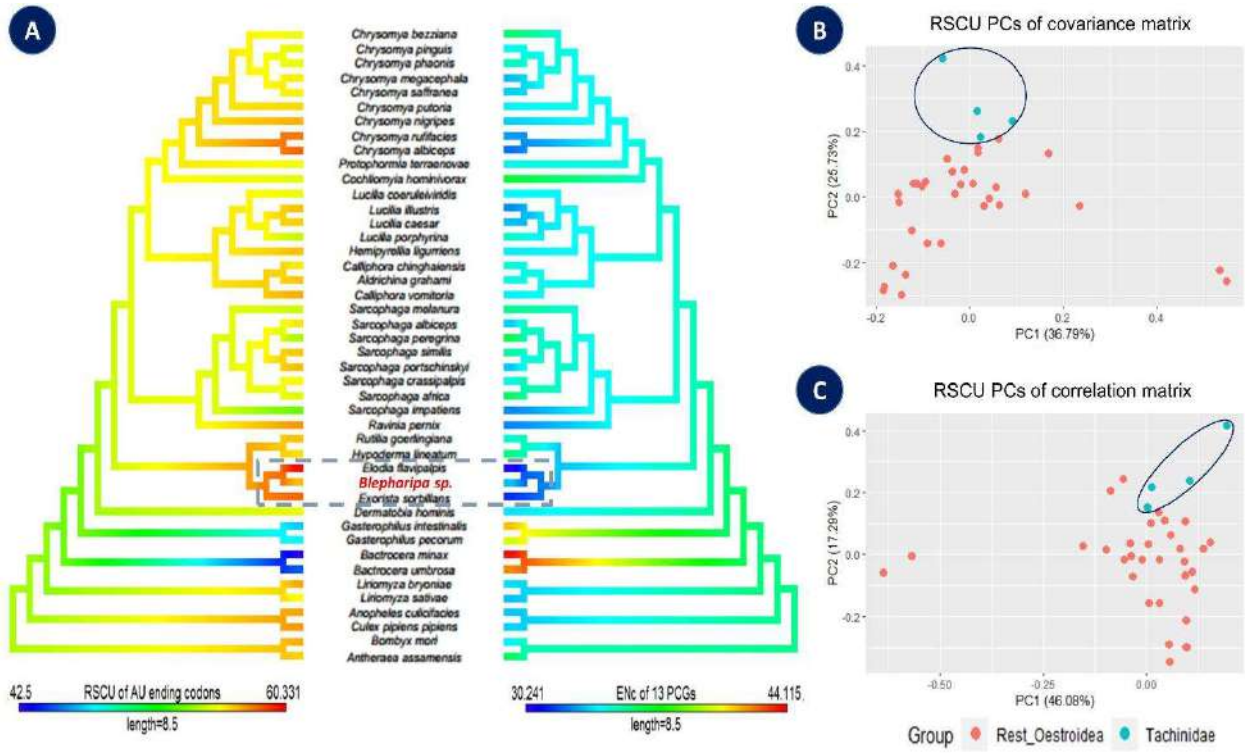


Figure 2.15: (A) RSCU value of AU ending codons and ENc of 13 concatenated PCGs. Contour map phylogeny shows the estimated evolutionary history of codon usage, and corresponding variation of ENc produced via contMap function in the R package Phytools. Note that the Tachinidae clade has evolved a AT content that is higher than the rest of the ingroup that is reflected in ENc. (B) and (C) Principal components analysis of RSCU across the Oestroidea. The Tachinidae groups are distinguishable from rest of the Oestroidea insects.

mitochondrial functions, as well as mutational pressure towards biased AT content to reduce the number of effective codons, resulting in a higher rate of energy synthesis at a lower cost[179,202,203,209,210]. This, in turn, provides a selective energetic advantage to the Tachinids in surviving the hostile environment of the host[211].

2.4 Conclusion:

The complete mitogenome of *Blepharipa sp.* has been sequenced and annotated to describe its characteristics at the molecular level. This study deliberates on gene orders, gene length, noncoding regions (control region, intergenic spacers), nucleotide composition, and codon usage of *Blepharipa sp.* mitogenome. In general, the features found in the mitogenome of *Blepharipa sp.* are similar to other previously studied tachinid flies[33,71]. The mitogenome

arrangement among Sarcophagidae and Tachinidae is consistent with ancestral type but some of the members of Calliphoridae and Oestridae have undergone tRNA rearrangements which have further led to the formation of a unique intergenic spacer and the overlapping region at their adjoining areas. Tachinid flies have a shorter mitogenome than other Oestroidea flies since the control region might not have been adequately covered with current sequencing and assembly methods due to the presence of extreme AT richness and repetitive sequences at CR. One important finding of the current comparative study is that *Blepharipa sp.* along with its family Tachinidae contain a relatively higher proportion of A+T nucleotides in their mitogenome and consequently, possess AT biased codons in their protein coding genes. The role of natural selection is found to be a major factor in determining organisms' synonymous codon usage bias rather than mutation pressure, as proven by other studies[117]. Within mitochondria, the longer genes (*nad5, nad4, nad1, cox1*) possess the most biased codons than the shorter genes and this phenomenon is equally observed in intron less genes of prokaryotes[34,35]. The significant usage of AT-rich codons by Tachinids is shown in this study, which limits the use of other codons. Tachinidae are also distinguished from the rest of the Oestroidea insects by principal component analysis of RSCU values. Further, the phylogenetic analysis based on protein-coding genes (PCGs) shows well-supported monophyly of the Sarcophagidae and Calliphoridae family, whereas Tachinidae and Oestridae encountered some irregularities and non-monophyly of taxa. Additional mitogenome sequencing data and a wider taxon sample are necessary to get an absolutely resolved Oestroidea phylogeny, particularly for the Tachinidae family, as it is one of the largest families of species existing on Earth.

The lineage wise nucleotide substitution analysis shows strong purifying selection on mitochondrial genes although the branch leading to the uzi flies' common ancestor has gained more nonsynonymous mutations than synonymous mutations, and therefore putting more selective pressure on it than other branches. We believe that the signal for positive selection is usually drowned out by relaxed selection because positive selection often occurs on a few sites

for a short period of evolutionary time, and therefore the value of ω is always less than 1.0 in either foreground or background branches. This study also shows that the nonlinear model fitted better to deduce the relationship between divergence rate and codon usage indices. Where, synonymous and nonsynonymous divergence rates exhibit opposite S-curve-like relationships with GC3 and GC3s, respectively, and we argue that both divergence rates will eventually form an S-curve with AT3. The divergence rate forms a valley-shape relation with ENc where the rate of divergence first decreases rapidly then again gradually increases, although the intensity of the synonymous divergence rate is higher than the nonsynonymous divergence rate.

Overall, the mitogenome reported here will serve as a useful dataset for studying the genetics, systematics, and phylogenetic relationships of many species, the Tachinidae family, in particular, and uzi flies flies, in general. Therefore, along with the completion of *Blepharipa sp.* mitogenome sequencing and documentation; a series of these extensive comparative analyses with related Oestroidea flies can open new aspects of insect mitogenome research.

2.5 References:

1. Smith, M. A., Wood, D. M., Janzen, D. H., Hallwachs, W. & Hebert, P. D. N. *DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists*. www.pnas.org/cgi/content/full/ (2007).
2. Dindo, M. L. Tachinid parasitoids: are they to be considered as koinobionts? *BioControl* **56**, 249–255 (2011).
3. Guo, J., Xie, K., Che, K., Hu, Z. & Guo, Y. The complete mitochondria genome of *Ravinia pernix* (Diptera: Sarcophagidae). *Mitochondrial DNA* 1–2 (2014) doi:10.3109/19401736.2014.982560.
4. Nelson, L. A. *et al.* Beyond barcoding: a mitochondrial genomics approach to molecular phylogenetics and diagnostics of blowflies (Diptera: Calliphoridae). *Gene* **511**, 131–42 (2012).
5. Signes, A. & Fernandez-Vizarra, E. Assembly of mammalian oxidative phosphorylation complexes I-V and supercomplexes. *Essays Biochem.* **62**, 255–270 (2018).
6. Szpila, K., Hall, M. J. R., Wardhana, A. H. & Pape, T. Morphology of the first instar larva of obligatory traumatic myiasis agents (Diptera: Calliphoridae, Sarcophagidae). *Parasitol. Res.* **113**, 1629–1640 (2014).
7. Zhu, Z. *et al.* The complete mitochondria genome of *Aldrichina grahami* (Diptera: Calliphoridae). *Mitochondrial DNA Part B* **1**, 107–109 (2016).
8. He, L., Wang, S., Miao, X., Wu, H. & Huang, Y. Identification of necrophagous fly species using ISSR

and SCAR markers. *Forensic Sci. Int.* **168**, 148–153 (2007).

9. Núñez-Vázquez, C., Tomberlin, J. & García-Martínez, O. First Record of the Blow Fly *Calliphora grahami* [1] from Mexico. *Southwest. Entomol.* **35**, 313–316 (2010).
10. Yan, J., Liao, H., Xie, K. & Cai, J. The complete mitochondria genome of *Chrysomya pinguis* (Diptera: Calliphoridae). *Mitochondrial DNA Part A* **27**, 3852–3854 (2016).
11. Monum, T. *et al.* Forensically Important Blow Flies Chrysomya pinguis, C. villeneuvi, and Lucilia porphyrina (Diptera: Calliphoridae) in a Case of Human Remains in Thailand. *Korean J. Parasitol.* **55**, 71–76 (2017).
12. Satou, A., Nisimura, T. & Numata, H. Reproductive competition between the burying beetle Nicrophorus quadripunctatus without phoretic mites and the blow fly Chrysomya pinguis. *Entomol. Sci.* **3**, 265–268 (2000).
13. Carvalho, L. M. L., Linhares, A. X. & Trigo, J. R. Determination of drug levels and the effect of diazepam on the growth of necrophagous flies of forensic importance in southeastern Brazil. *Forensic Sci. Int.* **120**, 140–144 (2001).
14. Protophormia terraenovae : Blackbottle | NBN Atlas | NBN Atlas. https://species.nbnatlas.org/species/NBNSYS0100004890.
15. Abd-Algalil, Zambare, S. P. & Mashaly. First record of Chrysomya saffranea (Diptera: Calliphoridae) of forensic importance in India. *Trop. Biomed.* **33**, 102–108 (2016).
16. Bunchu, N. *et al.* Morphology and Developmental Rate of the Blow Fly, *Hemipyrellia ligurriens* (Diptera: Calliphoridae): Forensic Entomology Applications. *J. Parasitol. Res.* **2012**, 1–10 (2012).
17. Sinha, S. K. Sarcophagidae, Calliphoridae and Muscidae (Diptera) of the Sundarbans Biosphere Reserve, West Bengal, India. *Occas. Pap. - Rec. Zool. Surv. India* (2009).
18. Klong-klaew, T. *et al.* Observations on morphology of immature Lucilia porphyrina (Diptera: Calliphoridae), a fly species of forensic importance. *Parasitol. Res.* **111**, 1965–1975 (2012).
19. Stevens, J. & Wall, R. The evolution of ectoparasitism in the genus Lucilia (Diptera: Calliphoridae). *Int. J. Parasitol.* **27**, 51–59 (1997).
20. Stevens, J. R., West, H. & Wall, R. Mitochondrial genomes of the sheep blowfly, Lucilia sericata, and the secondary blowfly, Chrysomya megacephala. *Med. Vet. Entomol.* **22**, 89–91 (2008).
21. Junqueira, A. C. M. *et al.* The mitochondrial genome of the blowfly Chrysomya chloropyga (Diptera: Calliphoridae). *Gene* **339**, 7–15 (2004).
22. Chen, J., Qiu, D., Yue, Q., Wang, C. & Li, X. The complete mitochondria genome of *Chrysomya phaonis* (Seguy, 1928) (Diptera: Calliphoridae). *Mitochondrial DNA Part B* **1**, 951–953 (2016).
23. Williams, K. A., Lamb, J. & Villet, M. H. Phylogenetic radiation of the greenbottle flies (Diptera, Calliphoridae, Luciliinae). *Zookeys* (2016) doi:10.3897/zookeys.568.6696.
24. Chen, Y. *et al.* The complete nucleotide sequence of the mitochondrial genome of *Calliphora chinghaiensis* (Diptera: Calliphoridae). *Mitochondrial DNA Part B* **1**, 397–398 (2016).
25. Akbarzadeh, K., Wallman, J. F., Sulakova, H. & Szpila, K. Species identification of Middle Eastern blowflies (Diptera: Calliphoridae) of forensic importance. *Parasitol. Res.* **114**, 1463–1472 (2015).
26. Ren, L., Guo, Q., Yan, W., Guo, Y. & Ding, Y. The complete mitochondria genome of *Calliphora vomitoria* (Diptera: Calliphoridae). *Mitochondrial DNA Part B* **1**, 378–379 (2016).
27. Šuláková, H. & Barták, M. Forensically important Calliphoridae (Diptera) associated with animal and

human decomposition in the Czech Republic: preliminary results. *Cas. slezskeho zemskeho Muz.* **62**, 255–266 (2013).

28. Seo, B. Y., Cho, J., Lee, G.-S., Park, J. & Park, J. The complete mitochondrial genome of Exorista japonica (Townsend, 1909) (Diptera:Tachinidae). *Mitochondrial DNA Part B* **4**, 2244–2245 (2019).
29. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell* (2015) doi:10.1016/j.molcel.2015.05.035.
30. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* **59**, 744–754 (2015).
31. Frumkin, I. *et al.* Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4940–E4949 (2018).
32. Buhr, F. *et al.* Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell* **61**, 341–351 (2016).
33. Zhao, Z. *et al.* The Mitochondrial Genome of Elodia flavipalpis Aldrich (Diptera: Tachinidae) and the Evolutionary Timescale of Tachinid Flies. *PLoS One* **8**, 61814 (2013).
34. Moriyama, E. & Powell, J. R. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. *Nucleic Acids Res.* **26**, 3188–3193 (1998).
35. Moriyama, E. N. *et al. Scientific Correspondence*. *Nucleic Acids Research* vol. 26 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC147868/pdf/264540.pdf (1998).
36. S.A., T., R., D. & A., Y. M. Chrysomya bezziana Infestation. *Arch. Iran. Med.* **5**, 56–58 (2002).
37. Lessinger, A. C. *et al.* The mitochondrial genome of the primary screwworm fly Cochliomyia hominivorax (Diptera: Calliphoridae). *Insect Mol. Biol.* **9**, 521–529 (2000).
38. Evans, K., Edited, K. A. & Richardson, S. J. Evaluating the effects of temperature on larval Calliphora vomitoria (Diptera: Calliphoridae) consumption.
39. Sukontason, K. L. *et al.* Larval Morphology of Chrysomya nigripes (Diptera: Calliphoridae), a Fly Species of Forensic Importance. *J. Med. Entomol.* **42**, 233–240 (2005).
40. Anderson, G. S. & Huitson, N. R. Myiasis in pet animals in British Columbia: the potential of forensic entomology for determining duration of possible neglect. *Can. Vet. J. = La Rev. Vet. Can.* **45**, 993–8 (2004).
41. Keshavarzi, D., Fereidooni, M., Assareh, M., Nasiri, Z. & Keshavarzi,, D. A Checklist of Forensic Important Flies (Insecta: Diptera) Associated with Indoor Rat Carrion in Iran. *J. Entomol. Zool. Stud. JEZS* **3**, 140–142 (2015).
42. Ribbeck, R., Danner, G. & Erices, J. [Wound myiasis in cattle infested by Lucilia caesar (Diptera: Calliphoridae)]. *Angew. Parasitol.* **28**, 229–31 (1987).
43. WEIGL, S. *et al.* The mitochondrial genome of the common cattle grub, Hypoderma lineatum. *Med. Vet. Entomol.* **24**, no-no (2010).
44. Logar, J. & Marinič-Fišer, N. Cutaneous myiasis caused by Hypoderma lineatum. *Wien. Klin. Wochenschr.* **120**, 619 (2008).
45. Pruett, J. H. Immunological control of arthropod ectoparasites—a review. *Int. J. Parasitol.* **29**, 25–32 (1999).
46. ADW: Hypoderma lineatum: CLASSIFICATION. https://animaldiversity.org/accounts/Hypoderma_lineatum/classification/.

TH-2821_146106003

47. Ana Maria Lima de Azeredo-Espin. The complete mitochondrial genome of the human bot fly Dermatobia hominis (Diptera: Oestridae). D0221 https://esa.confex.com/esa/2004/techprogram/paper_16801.htm (2004).
48. Goff, M. L., Campobasso, C. P. & Gherardi, M. Forensic Implications of Myiasis. in *Current Concepts in Forensic Entomology* 313–325 (Springer Netherlands, 2009). doi:10.1007/978-1-4020-9684-6_14.
49. ADW: Dermatobia hominis: CLASSIFICATION. https://animaldiversity.org/accounts/Dermatobia_hominis/classification/.
50. Zhang, D. *et al.* Phylogenetic inference of calyptrates, with the first mitogenomes for Gasterophilinae (Diptera: Oestridae) and Paramacronychiinae (Diptera: Sarcophagidae). *Int. J. Biol. Sci.* **12**, 489–504 (2016).
51. Gao, D.-Z. *et al.* The complete mitochondrial genome of Gasterophilus intestinalis, the first representative of the family Gasterophilidae. *Parasitol. Res.* **115**, 2573–2579 (2016).
52. Roelfstra, L. *et al.* Protein expression profile of Gasterophilus intestinalis larvae causing horse gastric myiasis and characterization of horse immune reaction. *Parasit. Vectors* **2**, 6 (2009).
53. ADW: Gasterophilus intestinalis: INFORMATION. https://animaldiversity.org/accounts/Gasterophilus_intestinalis/.
54. Ravinia pernix - Details - Encyclopedia of Life. http://eol.org/pages/781449/details.
55. Zhang, C., Fu, X., Zhu, Z., Xie, K. & Guo, Y. The complete mitochondrial genome sequence of *Helicophagella melanura* (Diptera: Sarcophagidae). *Mitochondrial DNA Part A* **27**, 3905–3906 (2016).
56. Szpila, K., Mądra, A., Jarmusz, M. & Matuszewski, S. Flesh flies (Diptera: Sarcophagidae) colonising large carcasses in Central Europe. *Parasitol. Res.* **114**, 2341–2348 (2015).
57. Chigusa, Y., Kawai, S., Kirinoki, M., Matsuda, H. & Morita, K. A case of myiasis due to Sarcophaga melanura (Diptera : Sarcophagidae) in a patient suffering from pontine infarction. *Med. Entomol. Zool.* **48**, 141–143 (1997).
58. Diaz, J. H. The Epidemiology, Diagnosis, Management, and Prevention of Ectoparasitic Diseases in Travelers. *J. Travel Med.* **13**, 100–111 (2006).
59. Fu, X., Che, K., Zhu, Z., Liu, J. & Guo, Y. The complete mitochondria genome of Sarcophaga africa (Diptera: Sarcophagidae). *Mitochondrial DNA* 1–2 (2014) doi:10.3109/19401736.2014.982582.
60. Wells, J. D., Pape, T. & Sperling, F. A. H. DNA-Based Identification and Molecular Systematics of Forensically Important Sarcophagidae (Diptera). *J. Forensic Sci.* **46**, 15105J (2001).
61. Shi, J. *et al.* The complete mitochondrial genome of the flesh fly, *Parasarcophaga portschinskyi* (Diptera: Sarcophagidae). *Mitochondrial DNA* 1–2 (2014) doi:10.3109/19401736.2014.971282.
62. Yan, J. *et al.* The complete mitochondria genome of *Parasarcophaga similis* (Diptera: Sarcophagidae). *Mitochondrial DNA* 1–2 (2014) doi:10.3109/19401736.2014.958708.
63. Chigusa, Y. *et al.* Two cases of otomyiasis caused by Sarcophaga peregrina and S. similis (Diptera : Sarcophagidae). *Med. Entomol. Zool.* **45**, 153–157 (1994).
64. Cherix, D., Wyss, C. & Pape, T. Occurrences of flesh flies (Diptera: Sarcophagidae) on human cadavers in Switzerland, and their importance as forensic indicators. *Forensic Sci. Int.* **220**, 158–163 (2012).
65. Zhong, M. *et al.* The complete mitochondrial genome of the flesh fly, *Boettcherisca peregrine* (Diptera: Sarcophagidae). *Mitochondrial DNA* **27**, 106–108 (2016).
66. Ambedkar, B., Fahd Abd Algalil Ph D Research Student, C. M., Abd Algalil, F. M. & Zambare, S. P.

Molecular identification of forensically important flesh flies (Diptera : Sarcophagidae) using COI Gene. *J. Entomol. Zool. Stud. JEZS* **5**, 263–267 (2017).

67. Nelson, L. A., Cameron, S. L. & Yeates, D. K. The complete mitochondrial genome of the flesh fly, *Sarcophaga impatiens* Walker (Diptera: Sarcophagidae). *Mitochondrial DNA* **23**, 42–43 (2012).
68. Ramakodi, M. P., Singh, B., Wells, J. D., Guerrero, F. & Ray, D. A. A 454 sequencing approach to dipteran mitochondrial genome research. *Genomics* (2015) doi:10.1016/j.ygeno.2014.10.014.
69. Giangaspero, A. *et al.* Wound Myiasis Caused by *Sarcophaga (Liopygia) Argyrostoma* (Robineau-Desvoidy) (Diptera: Sarcophagidae): Additional Evidences of the Morphological Identification Dilemma and Molecular Investigation. *Sci. World J.* **2017**, 1–9 (2017).
70. Liao, H., Yang, X., Li, Z., Ding, Y. & Guo, Y. The complete mitochondria genome of *Parasarcophaga albiceps* (Diptera: Sarcophagidae). *Mitochondrial DNA Part A* **27**, 4696–4698 (2016).
71. Shao, Y. jun *et al.* Structure and evolution of the mitochondrial genome of Exorista sorbillans: the Tachinidae (Diptera: Calyptratae) perspective. *Mol. Biol. Rep.* **39**, 11023–11030 (2012).
72. Yang, F., Du, Y., Cao, J. & Huang, F. Analysis of three leafminers' complete mitochondrial genomes. *Gene* **529**, 1–6 (2013).
73. Minkenberg, O. P., & van Lenteren, J. C. The leafminers, Liriomyza bryoniae and L. trifolii (Diptera: Agromyzidae), their parasites and host plants: a review. *Agric. Univ.* **86**, (1986).
74. Spencer, K. A. *Host Specialization in the World Agromyzidae (Diptera)*. (Springer Netherlands, 1990).
75. Yang, F., Du, Y., Wang, L., Cao, J. & Yu, W. The complete mitochondrial genome of the leafminer Liriomyza sativae (Diptera: Agromyzidae): Great difference in the A+T-rich region compared to Liriomyza trifolii. *Gene* **485**, 7–15 (2011).
76. Zhang, B., Nardi, F., Hull-Sanders, H., Wan, X. & Liu, Y. The Complete Nucleotide Sequence of the Mitochondrial Genome of Bactrocera minax (Diptera: Tephritidae). *PLoS One* **9**, e100558 (2014).
77. Hafsi, A. *et al.* Host plant range of a fruit fly community (Diptera: Tephritidae): does fruit composition influence larval performance? *BMC Ecol.* **16**, 40 (2016).
78. Yong, H.-S., Song, S.-L., Lim, P.-E., Eamsobhana, P. & Suana, I. W. Complete Mitochondrial Genome of Three Bactrocera Fruit Flies of Subgenus Bactrocera (Diptera: Tephritidae) and Their Phylogenetic Implications. *PLoS One* **11**, e0148201 (2016).
79. Luo, Q.-C. *et al.* The mitochondrial genomes of Culex tritaeniorhynchus and Culex pipiens pallens (Diptera: Culicidae) and comparison analysis with two other Culex species. *Parasit. Vectors* **9**, 406 (2016).
80. Hua, Y.-Q. *et al.* Sequencing and analysis of the complete mitochondrial genome in *Anopheles culicifacies* species B (Diptera: Culicidae). *Mitochondrial DNA* 1–2 (2015) doi:10.3109/19401736.2015.1060434.
81. Yukuhiro, K., Sezutsu, H., Itoh, M., Shimizu, K. & Banno, Y. Significant Levels of Sequence Divergence and Gene Rearrangements have Occurred Between the Mitochondrial Genomes of the Wild Mulberry Silkmoth, Bombyx mandarina, and its Close Relative, the Domesticated Silkmoth, Bombyx mori. *Mol. Biol. Evol.* **19**, 1385–1389 (2002).
82. Singh, D. *et al.* The mitochondrial genome of Muga silkworm (Antheraea assamensis) and its comparative analysis with other lepidopteran insects. *PLoS One* **12**, e0188077 (2017).
83. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).

TH-2821_146106003

84. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **10**, (2009).
85. Altschup, S. F., Gish, W., Pennsylvania, T. & Park, U. Basic Local Alignment Search Tool 2Department of Computer Science. 403–410 (1990).
86. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
87. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–77 (1999).
88. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods 2012 94* **9**, 357–359 (2012).
89. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Appl. NOTE* **25**, 2078–2079 (2009).
90. Bronstein, O., Kroh, A. & Haring, E. Mind the gap! The mitochondrial control region and its power as a phylogenetic marker in echinoids. *BMC Evol. Biol.* **18**, 80 (2018).
91. Bernt, M. *et al.* MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* **69**, 313–319 (2013).
92. Rombel, I. T., Sykes, K. F., Rayner, S. & Johnston, S. A. ORF-FINDER: a vector for high-throughput gene identification. *Gene* **282**, 33–41 (2002).
93. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
94. Hall, A. T. BioEdit: a user friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98 (1999).
95. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
96. Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. GenBank. *Nucleic Acids Res.* **42 .**, D32–D37 (2014).
97. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
98. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
99. NCBI Sequin. http://www.ncbi.nlm.nih.gov/Sequin.
100. Katoh, K., Kuma, K. I., Toh, H. & Miyata, T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
101. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).
102. Sievers, F. & Higgins, D. G. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. in 105–116 (Humana Press, Totowa, NJ, 2014). doi:10.1007/978-1-62703-646-7_6.
103. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
104. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772–772 (2012).
105. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **61**, 539–542 (2012).

106. Rambaut, Andrew, Marc A. Suchard, D. Xie, and A. J. D. Tracer v1. 6. 2014. (2015).
107. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
108. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
109. Revell, L. J. phytools : an R package for phylogenetic comparative biology (and other things). 217–223 (2012) doi:10.1111/j.2041-210X.2011.00169.x.
110. Xu, B. & Yang, Z. PAMLX: A Graphical User Interface for PAML. *Mol. Biol. Evol.* **30**, 2723–2724 (2013).
111. Xia, X. DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *J. Hered.* **108**, 431–437 (2017).
112. CIMminer. https://discover.nci.nih.gov/cimminer/home.do.
113. Sun, X., Yang, Q. & Xia, X. An Improved Implementation of Effective Number of Codons (Nc). *Mol. Biol. Evol.* **30**, 191–196 (2013).
114. Cutter, A. D., Wasmuth, J. D. & Blaxter, M. L. The evolution of biased codon and amino acid usage in nematode genomes. *Mol. Biol. Evol.* **23**, 2303–2315 (2006).
115. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
116. Jiang, Y., Deng, F., Wang, H. & Hu, Z. An extensive analysis on the global codon usage pattern of baculoviruses. *Arch. Virol.* **153**, 2273–2282 (2008).
117. Wei, L. *et al.* Analysis of codon usage bias of mitochondrial genome in Bombyx mori and its relation to evolution. *BMC Evol. Biol.* **14**, 1–12 (2014).
118. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 2653–2657 (1988).
119. Zhang, W. *et al.* Comparative Analysis of Codon Usage Patterns Among Mitochondrion , Chloroplast and Nuclear Genes in Triticum aestivum L . **49**, 246–254 (2007).
120. Sueoka, N. & Kawanishi, Y. DNA G+C content of the third codon position and codon usage biases of human genes. *Gene* **261**, 53–62 (2000).
121. He, B. *et al.* Analysis of codon usage patterns in Ginkgo biloba reveals codon usage tendency from A/U-ending to G/C-ending. *Sci. Rep.* **6**, 1–11 (2016).
122. Montgomery, Douglas C., Elizabeth A. Peck, and G. G. V. *Introduction to Linear Regression Analysis*. (2021).
123. Bradley, R. A. & Srivastava, S. S. Correlation in polynomial regression. *Am. Stat.* **33**, 10–14 (1979).
124. Wood, S. N. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **62**, 413–428 (2000).
125. Faraway, J. J. *Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models*. *Extending the Linear Model with R* (Chapman and Hall/CRC, 2016). doi:10.1201/9781315382722.
126. Chen, S.-C., Wei, D.-D., Shao, R., Dou, W. & Wang, J.-J. The Complete Mitochondrial Genome of the Booklouse, Liposcelis decolor: Insights into Gene Arrangement and Genome Organization within the Genus Liposcelis. *PLoS One* **9**, e91902 (2014).
127. Zhang, X. *et al.* Comparative Mt Genomics of the Tipuloidea (Diptera: Nematocera: Tipulomorpha) and

its implications for the phylogeny of the Tipulomorpha. *PLoS One* **11**, (2016).

128. Lewis, O. L., Farr, C. L. & Kaguni, L. S. Drosophila melanogaster mitochondrial DNA: completion of the nucleotide sequence and evolutionary comparisons. *Insect Mol. Biol.* **4**, 263–278 (1995).
129. Cameron, S. L., Yoshizawa, K., Mizukoshi, A., Whiting, M. F. & Johnson, K. P. Mitochondrial genome deletions and minicircles are common in lice (Insecta: Phthiraptera). *BMC Genomics* **12**, 394 (2011).
130. Oliveira, M. T. *et al.* Structure and evolution of the mitochondrial genomes of Haematobia irritans and Stomoxys calcitrans: The Muscidae (Diptera: Calyptratae) perspective. *Mol. Phylogenet. Evol.* (2008) doi:10.1016/j.ympev.2008.05.022.
131. Bernt, M. *et al.* A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. *Mol. Phylogenet. Evol.* **69**, 352–364 (2013).
132. Cameron, S. L. How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. *Syst. Entomol.* (2014) doi:10.1111/syen.12071.
133. Boore, J. L. Animal mitochondrial genomes. *Nucleic Acids Res.* **27**, 1767–80 (1999).
134. Zhang, D.-X. & Hewitt, G. M. *Insect Mitochondrial Control Region: A Review of its Structure, Evolution and Usefulness in Evolutionary Studies*. *Biochemical Systematics and Ecology* vol. 25 (1997).
135. Clary, D. O., Goddard, J. M., Martin, S. C., Fauron, C. M. R., & Wolstenholme, D. R. Drosophila mitochondrial DNA: a novel gene order. *Nucleic Acids Res.* **10**, 6619–663 (1982).
136. Lessinger, A. C., Junqueira, A. C. M., Conte, F. F. & Azeredo-Espin, A. M. L. Analysis of a conserved duplicated tRNA gene in the mitochondrial genome of blowflies. *Gene* **339**, 1–6 (2004).
137. Beckenbach, A. T. Mitochondrial genome sequences of nematocera (lower diptera): Evidence of rearrangement following a complete genome duplication in a winter crane fly. *Genome Biol. Evol.* **4**, 89–101 (2012).
138. Negrisolo, E., Babbucci, M. & Patarnello, T. The mitochondrial genome of the ascalaphid owlfly Libelloides macaronius and comparative evolutionary mitochondriomics of neuropterid insects. *BMC Genomics* (2011) doi:10.1186/1471-2164-12-221.
139. Lavrov, D. V, Brown, W. M. & Boore, J. L. A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede Lithobius forficatus. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13738–42 (2000).
140. By, E., Grosjean, H. & Benne, R. Modification and Editing of RNA. **25**, 1945–450 (1998).
141. Liu, Q.-N. *et al.* The first complete mitochondrial genome for the subfamily Limacodidae and implications for the higher phylogeny of Lepidoptera. *Sci. Rep.* **6**, 35878 (2016).
142. Patricio Fernández-Silva *, J. A. E. and J. M. *Replication and transcription of mammalian mitochondrial DNA*. http://megasun.bch.umontreal.ca/gobase/ (2003).
143. Taanman, J.-W. The mitochondrial genome: structure, transcription, translation and replication. *Biochim. Biophys. Acta - Bioenerg.* **1410**, 103–123 (1999).
144. Crochet, P.-A. & Desmarais, E. Slow Rate of Evolution in the Mitochondrial Control Region of Gulls (Aves: Laridae). *Mol. Biol. Evol* **17**, 1797–1806 (2000).
145. Atray, I., Bentur, J. S. & Nair, S. The asian rice gall midge (Orseolia oryzae) mitogenome has evolved novel gene boundaries and tandem repeats that distinguish its biotypes. *PLoS One* (2015) doi:10.1371/journal.pone.0134625.
146. Song, N., Liang, A.-P. & Ma, C. The complete mitochondrial genome sequence of the planthopper, Sivaloka damnosus. *J. Insect Sci.* **10**, 76 (2010).

147. Chen, J.-Y., Chang, Y.-W., Zheng, S.-Z., Lu, M.-X. & Du, Y.-Z. Comparative analysis of the Liriomyza chinensis mitochondrial genome with other Agromyzids reveals conserved genome features. *Sci. Rep.* **8**, 8850 (2018).

148. Duarte, G. T., De Azeredo-Espin, A. M. L. & Junqueira, A. C. M. The Mitochondrial Control Region of Blowflies (Diptera: Calliphoridae): A Hot Spot for Mitochondrial Genome Rearrangements. *J. Med. Entomol.* **45**, 667–676 (2008).

149. Struhl, K. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 8419–23 (1985).

150. Mirkin, E. V, Castro Roa, D., Nudler, E. & Mirkin, S. M. Transcription regulatory elements are punctuation marks for DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7276–81 (2006).

151. Singh, D. *et al.* The mitochondrial genome of Muga silkworm (Antheraea assamensis) and its comparative analysis with other lepidopteran insects. *PLoS One* (2017) doi:10.1371/journal.pone.0188077.

152. Sun, Y. *et al.* Comparative mitochondrial genome analysis of Daphnis nerii and other lepidopteran insects reveals conserved mitochondrial genome organization and phylogenetic relationships. *PLoS One* **12**, e0178773 (2017).

153. Beckenbach, A. T. & Stewart, J. B. Insect mitochondrial genomics 3: the complete mitochondrial genome sequences of representatives from two neuropteroid orders: a dobsonfly (order Megaloptera) and a giant lacewing and an owlfly (order Neuroptera). *Genome* **52**, 31–38 (2009).

154. Song, S.-N., Tang, P., Wei, S.-J. & Chen, X.-X. Comparative and phylogenetic analysis of the mitochondrial genomes in basal hymenopterans. *Sci. Rep.* **6**, 20972 (2016).

155. Salvato, P., Simonato, M., Battisti, A. & Negrisolo, E. The complete mitochondrial genome of the bag-shelter moth Ochrogaster lunifer (Lepidoptera, Notodontidae). *BMC Genomics* **9**, 331 (2008).

156. Roberti, M. *et al.* DmTTF, a novel mitochondrial transcription termination factor that recognises two sequences of Drosophila melanogaster mitochondrial DNA. *Nucleic Acids Res.* **31**, 1597–1604 (2003).

157. Cameron, Stephen L., Christine L. Lambkin, Stephen C. Barker, and M. F. W. A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. *Syst. Entomol.* **32**, 40–59 (2007).

158. Nardi, F., Carapelli, A., Fanciulli, P. P., Dallai, R. & Frati, F. The Complete Mitochondrial DNA Sequence of the Basal Hexapod Tetrodontophora bielanensis: Evidence for Heteroplasmy and tRNA Translocations. *Mol. Biol. Evol.* **18**, 1293–1304 (2001).

159. Li, X. *et al.* The First Mitochondrial Genome of the Sepsid Fly Nemopoda mamaevi Ozerov, 1997 (Diptera: Sciomyzoidea: Sepsidae), with Mitochondrial Genome Phylogeny of Cyclorrhapha. *PLoS One* **10**, e0123594 (2015).

160. Slomovic, S., Laufer, D., Geiger, D. & Schuster, G. Polyadenylation and degradation of human mitochondrial RNA: the prokaryotic past leaves its mark. *Mol. Cell. Biol.* **25**, 6427–35 (2005).

161. Ojala, D., Montoya, J. & Attardi, G. tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**, 470–474 (1981).

162. Cameron, S. L. & Whiting, M. F. The complete mitochondrial genome of the tobacco hornworm, Manduca sexta, (Insecta: Lepidoptera: Sphingidae), and an examination of mitochondrial gene variability within butterflies and moths. *Gene* **408**, 112–123 (2008).

TH-2821_146106003

163. Wei, S.-J. *et al.* New views on strand asymmetry in insect mitochondrial genomes. *PLoS One* **5**, e12708 (2010).

164. Uddin, A., Mazumder, T. H., Choudhury, M. N. & Chakraborty, S. Codon bias and gene expression of mitochondrial ND2 gene in chordates. *Bioinformation* **11**, 407–12 (2015).

165. Zhang, N. X., Yu, G., Li, T. J., He, Q. Y., Zhou, Y., Si, F. L., ... & Chen, B. The Complete Mitochondrial Genome of Delia antiqua and Its Implications in Dipteran Phylogenetics. *PLoS One* **10**, e0139736 (2015).

166. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21 (1981).

167. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34 (1985).

168. Akashi, H., Eyre-Walker, A. & Akashi, H. Translational selection and molecular evolution An interplay among experimental studies of protein synthesis, evolutionary theory, and comparisons of DNA sequence data has shed light on the roles of natural selection and genetic drift in 'silent' DNA evolution. *Curr. Opin. Genet. Dev.* **8**, 688–693 (1998).

169. Akashi, H. Gene expression and molecular evolution. *Current Opinion in Genetics and Development* vol. 11 660–666 (2001).

170. Laurent Duret. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640–649 (2002).

171. Grantham, R., Gautier, C. & Gouy, M. Codon frequencies in 119 individual genes confirm corsistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* **8**, 1893–1912 (1980).

172. Angellotti, M. C., Bhuiyan, S. B., Chen, G. & Wan, X.-F. CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res.* **35**, W132–W136 (2007).

173. Swire, J., Judson, O. P. & Burt, A. Mitochondrial Genetic Codes Evolve to Match Amino Acid Requirements of Proteins. *J. Mol. Evol.* **60**, 128–139 (2005).

174. Chen, H., Sun, S., Norenburg, J. L. & Sundberg, P. Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms (Nemertea). *PLoS One* **9**, e85631 (2014).

175. Hershberg, R. & Petrov, D. A. Selection on Codon Bias. *Annu. Rev. Genet.* (2008) doi:10.1146/annurev.genet.42.110807.091442.

176. Sueoka, N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. U. S. A.* **48**, 582–592 (1962).

177. Zhou, M. & Li, X. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. *Mol. Biol. Rep.* **36**, 2039–2046 (2009).

178. Nie, X. *et al.* Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family. *Plant Mol. Biol. Report.* **32**, 828–840 (2014).

179. Guan, D. L., Qian, Z. Q., Ma, L. Bin, Bai, Y. & Xu, S. Q. Different mitogenomic codon usage patterns between damselflies and dragonflies and nine complete mitogenomes for odonates. *Sci. Rep.* **9**, 1–9 (2019).

180. Eyre-Walker, A. Synonymous codon bias is related to gene length in Escherichia coli: Selection for translational accuracy? *Mol. Biol. Evol.* **13**, 864–872 (1996).

181. Marinho, M. A. T. *et al.* Molecular phylogenetics of Oestroidea (Diptera: Calyptratae) with emphasis on

Calliphoridae: Insights into the inter-familial relationships and additional evidence for paraphyly among blowflies. *Mol. Phylogenet. Evol.* (2012) doi:10.1016/j.ympev.2012.08.007.

182. Kutty, S. N., Pape, T., Wiegmann, B. M. & Meier, R. Molecular phylogeny of the Calyptratae (Diptera: Cyclorrhapha) with an emphasis on the superfamily Oestroidea and the position of Mystacinobiidae and McAlpine's fly. *Syst. Entomol.* **35**, 614–635 (2010).
183. Wiegmann, B. M. *et al.* Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* **108**, 5690–5695 (2011).
184. McAlpine, J.F., 1989. *Phylogeny and classification of the Muscomorpha. Manual of Nearctic Diptera 3*. (1989). doi:doi/10.1086/417000.
185. Nirmala, X., Hyp, V. & Z, M. Molecular phylogeny of Calyptratae (Diptera : Brachycera): the evolution of 18S and 16S ribosomal rDNAs in higher dipterans and their use in phylogenetic inference. **10**, 475–485 (2001).
186. Winkler, I. S. *et al.* Explosive radiation or uninformative genes? Origin and early diversification of tachinid flies (Diptera: Tachinidae). *Mol. Phylogenet. Evol.* **88**, 38–54 (2015).
187. Rognes, K. The Calliphoridae (Blowflies) (Diptera : Oestroidea) are Not a Monophyletic Group 1. **68**, (1997).
188. Mesnil, L. P. Larvaevorinae (Tachininae). In: Lindner, E. (Ed.), Die Fliegen der palaearktischen Region 10 (Lieferung 263). 881–928 (1966).
189. Cerretti, P. *et al.* Signal through the noise? Phylogeny of the Tachinidae (Diptera) as inferred from morphological evidence. *Syst. Entomol.* **39**, 335–353 (2014).
190. Abel, O. Das biologische Trägheitsgesetz. *Palaeontol. Zeitschrift* **11**, 7–17 (1929).
191. Blomberg, S. P. & Garland, T. Tempo and mode in evolution: Phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology* vol. 15 899–910 (2002).
192. Edwards, S. V. & Naeem, S. The phylogenetic component of cooperative breeding in perching birds. *Am. Nat.* **141**, 754–789 (1993).
193. Mckitrick, M. C. *PHYLOGENETIC CONSTRAINT IN EVOLUTIONARY THEORY: Has It Any Explanatory Power? Annu. Rev. Ecol. Syst* vol. 24 www.annualreviews.org (1993).
194. Bacigalupe, L. D., Nespolo, R. F., Opazo, J. C. & Bozinovic, F. Phenotypic flexibility in a novel thermal environment: Phylogenetic inertia in thermogenic capacity and evolutionary adaptation in organ size. *Physiol. Biochem. Zool.* **77**, 805–815 (2004).
195. Shen, Y.-Y. *et al.* Adaptive evolution of energy metabolism genes and the origin of flight in bats. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8666–8671 (2010).
196. Simpson, G. L. Modelling palaeoecological time series using generalised additive models. *Front. Ecol. Evol.* **6**, 149 (2018).
197. Runge, C. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Math. und Phys.* **46**, 224–243 (1901).
198. Epperson, J. F. On the Runge Example. *Am. Math. Mon.* **94**, 329–341 (1987).
199. Brandis, G. & Hughes, D. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLOS Genet.* **12**, e1005926 (2016).
200. Zhou, Z. *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci.* (2016) doi:10.1073/pnas.1606724113.

201. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115 (2013).
202. Shen, Y. Y., Shi, P., Sun, Y. B. & Zhang, Y. P. Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. *Genome Res.* **19**, 1760–1765 (2009).
203. Tessa E.F. Quax,1,2,4 Nico J. Claassens,1,4 Dieter So¨ ll, 3 and John van der Oost1 & *. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **59**, 149-161. (2015).
204. Angov, E. Codon usage: Nature's roadmap to expression and folding of proteins. *Biotechnology Journal* (2011) doi:10.1002/biot.201000332.
205. Rimer, J., Cohen, I. R. & Friedman, N. Do all creatures possess an acquired immune system of some sort? *BioEssays* **36**, 273–281 (2014).
206. Mcdonagh, L. M. & Stevens, J. R. The molecular systematics of blowflies and screwworm flies (Diptera: Calliphoridae) using 28S rRNA, COX1 and EF-1α: insights into the evolution of dipteran parasitism. *Parasitology* **138**, 1760–1777 (2011).
207. Kumar, B. Biocontrol of Insect Pests. *Ecofriendly Pest Manag. Food Secur.* 25–61 (2016) doi:doi.org/10.1016/b978-0-12-803265-7.00002-6.
208. Janzen, D. H. The caterpillars and their parasitoids of a tropical dry forest. *Tachinid Time* **8**, 1–3 (1995).
209. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3645 (2010).
210. Chakraborty, S. *et al.* Codon usage and expression level of human mitochondrial 13 protein coding genes across six continents. *Mitochondrion* (2017) doi:10.1016/J.MITO.2017.11.006.
211. Zhou, Z., Dang, Y., Zhou, M., Yuan, H. & Liu, Y. Codon usage biases co-evolve with transcription termination machinery to suppress 1 premature cleavage and polyadenylation 2 3.

TH-2821_146106003

CHAPTER 3

Diptera Phylogeny with Larger Taxa

> "Normalcy is the antithesis of evolution."
>
> — Siddhartha Mukherjee

TH-2821_146106003

CHAPTER 3

Reconstruction of Dipteran Phylogeny with larger taxa

Abstract:

With the advent of next-generation sequencing, the number of loci available for phylogenetic analysis has increased unprecedently. Concatenated molecular sequence alignments of multiple loci of large taxon set may contain a variety of different signals, one of which is the historical signal that we often try to recover by phylogenetic analysis. Other signals, such as those arising due to codon usage bias, compositional heterogeneities, among-lineage and among-site rate heterogeneities, invariant sites, substitution saturation, homoplasy or heterotachy may interfere adversely with the recovery of the historical signal. Subsequently, misinterpretation of phylogenetic information leading to difficulty in determining branching patterns of diversification and resulting in dissimilar resolution of phylogenetic trees from different approaches. We demonstrate how non-phylogenetic signals in mitochondrial genomic data from Diptera (True flies) have an influence on phylogeny. In this study such discordance of signal was captured by measuring and profiling phylogenetic informativeness (PI) of different homogeneous phylogenetic trees. The homogeneous analysis using concatenated raw alignment reveals that the outgroup species have a close relationship with certain of the in-group species, as well as long-branching of those specific in-group species. Assessment with different heterogeneity test within dataset suggests that different taxa and families of Diptera lineages possess unusual base composition and usage of synonymous codon variants. In

addition, substitution saturation predominant in the dataset, 3rd codon position is mostly saturated by transition than transversion and heterogeneous sequence divergence shows 3rd codon positions are more rate-heterogeneous than other two codon positions. The use of coalescent-based species trees to examine gene-tree discordance revealed mitochondrial gene tree discordance in the backbone of species tree. We inspected datasets for evidence of other historical processes, such as recombination, or reticulate evolution and their outcome invoked the idea of inferring unique features from the data without limiting our focus to a single binary tree. The network represents obvious ambiguity in the signal might observed due to rapid and reticulate diversifications may explain conflicts among genes and the challenges for resolving evolutionary relationships.

3.1 Introduction:

Diptera (True flies), one of the most diverse insect orders (150,000 species reported) evolved from early aquatic ancestors, first learnt to fly with two wings in the Permian period, went through three episodes of adaptive radiation, and the current lower Diptera rapidly radiated during the Triassic period [1–3]. Beyond its sheer numbers and extensive history, they present a remarkable variety of morphological and behavioral diversity in numerous ecological niches [2,4]. However, resolving Diptera phylogeny using the molecular data and consequently, enumeration of diversification dating and molecular evolution of major lineages is still controversial. Here we reveal the nature of phylogenetic approaches to shed light on key issues.

Phylogenetics is the study of the evolutionary history of a group of taxa by detecting homologous characters in a given set of data and estimating the evolutionary history of species by comparing those characters using tree reconstruction techniques [5,6]. However, establishing the phylogenetic relationship of any group of species is fraught with difficulties, such as the selection of appropriate taxa, genes, character-set partitioning, and the substitution model for assessing the integrity of the resultant relationship, which evolves at an appropriate rate to

resolve unknown ancestral branching [7–10]. In addition, the execution of a decision based on an imprecise knowledge of the genes used in prior study, as well as inadequate taxon sampling due to the lack of genes for certain taxa in a database, leads to incorrect phylogenetic inference [11–13]. Therefore, genes chosen for phylogenetic reconstruction should have the correct information to provide the best phylogenetic tree resolution on given taxa. Generally, single-gene phylogenies are not well suited for proper species tree resolution due to a lower number of informative sites and the presence of stochastic noise [14]. Large data sets, on the other hand, may reduce stochasticity while increasing systematic errors like base compositional heterogeneity[15–18], among-site rate variation[19–21], heterotachy[22,23], those can lead to inaccurate phylogenetic relationships regardless of the inference method used.

The whole-tree calculation is often unable to detect the presence of heterogeneity of information in different parts of the phylogenetic tree[10]. Although, support for specific branches or clade of a tree is estimated by bootstrap values, posterior Bayesian probabilities, and these tests are key to the reliability of an inferred tree[24,25]. Earlier several indicators were anticipated to determine the amount of signal present in the data, such as the tree length distribution skewness, consistency index[26,27]. However, these parameters could not provide the power of the characters to reveal any true internodes that define clades in a specific depth of the tree[10]. The measure Phylogenetic Informativeness (PI) suggested by Townsend (2007) offers an empirical estimation of the character set evolution rate at a certain state space, which generates the historical profile of the expected real parsimony of informativeness based on canonical polytomy[10,28]. Although PI can detect historical signals present on the gene but unable to explain other signals, such as compositional heterogeneity, rate heterogeneity across lineages, invariant sites, substitutional saturation, homoplasy (inconsistent site), or other biological phenomenon which have adverse effect in the inference of the proper historical signal [29–33].

The selection of an appropriate model is a crucial step in phylogeny construction. But, in conventional molecular phylogenetics, it is frequently assumed that sequences evolve in all lineages under stationary, reversible, and homogeneous conditions[15,16,34]. If the extant sequence has evolved from its ancestors with different base composition or same site has evolved several times in distantly related taxa, the assumption may be violated and such taxa may cluster irrespective of their true phylogenetic relationship[18,35]. Historical signals are therefore weakened to the extent that they are no more effective than the conflicting signals, and the techniques used to infer evolutionary distances and topologies are likely to produce biased results[33]. The sequence sites with a high substitution rate in distinct lineages are likely to exhibit homoplastic character states either as a consequence of a rapid molecular evolution process or as a result of long evolutionary periods[12]. The existence of homoplasy in an alignment is doubtful as high frequencies of convergent nucleotide states will distort branch length and ancestral relationships between nucleotide sequences estimation in phylogenetic analysis, as the hidden substitution rate is often underestimated or overestimated[12,36]. Thus, fast-evolving sites are especially difficult for phylogenetic reconstruction because they are likely to have undergone many changes, diluted the actual signal and confounded phylogenetic inference[5,37,38]. The substitution saturation is another the most frequently discussed cause of phylogenetic artifacts [39]. If multiple substitutions occur frequently at the same position of an alignment, the apparent pairwise genetic distances underestimate the real distances and the alignment is said to be saturated [40]. In a saturated data matrix, synapomorphies may be erased by additional mutations, which can weaken branch support, and potentially misled phylogenetic inference [41]. Although mitochondrial genes are extensively used as a phylogenetic marker for a diverse group of organisms without assessing the presence of phylogenetic signals and noise within it. Even potential reasons for signal distortion from widely used mitochondrial genes as a marker of phylogeny are not clear whether heterogeneity in composition, rate,

divergence, or saturation among sites of phylogenetic inference affects the deep relationship of mitochondrial insect genomes. Therefore, there is a serious need to measure phylogenetic signal in this context as well as various influencing factors which reduce the historical signal for proper phylogenetic tree resolution.

Mitochondrial genomes have been widely used to infer phylogenetic relationships among insects but deep relationships are found to ineffective and contentious in different studies [42–45]. Thus, focus has been placed on several aspects, to minimize the effect of nonphylogenetic signals such as (i) sampling more species to correctly infer multiple substitutions at a site [40,46,47]. (ii) gene removal, reduced encoding and use of RY-coding of protein-coding genes to minimize saturation and compositional bias [48–52]. (iii) the phylogenetic utility of gene re-arrangement, and the alignment of RNAs based on secondary structures[53–56]. (iv) the detection and exclusion of fast-evolving sites, such as the slow-fast (SF) process [38,57] and the observed variability (OV) sorting method [58,59].

Despite these advancements some of the issues have not properly addressed. Increasing taxon sampling leads to variation in evolutionary rates among lineages as a result of that multiple substitution at the same site (homoplasy) causing a tree to suffer from a problem known as the Long-branch attraction (LBA)[14,60,61]. In these situations, maximum parsimony is misled by model misspecification, LBA may occur even if the evolutionary model adequately fits the data, and there is growing evidence that maximum-likelihood and distance methods can also affected by rate heterogeneity among lineages when inappropriate substitution models are used [62–64]. The utilization of larger taxa or characters that are still debatable in the design of phylogenetic studies as larger character sets contribute both to phylogenetic signal and noise [41,65,66]. On the other side, the informativeness of increasing taxonomic sampling is crucially dependent on the chronology of the ancestral lineage of the taxa added to the data set [26,67,68]. The compositional and mutational biases violate substitution models, which

accommodate among-site rate variation by using a GTR (general-time-reversible) model with a uniform gamma distribution to satisfy differences in rates (fast or slow sites), but ignore variation in other parameters that vary across characters[23,69] . Further, evolutionary models that do not account for compositional heterogeneity can lead to incorrect unions of unrelated species with similar base compositions [70,71]. The site-heterogeneous mixture models can address this issue by adding several rate categories and diversifying site-specific exchange abilities, reducing the detrimental impacts of compositional and mutational bias by relaxing the assumption of homogeneity across sites in conventional models [69,72–74]. The use of the most accurate models can extract genuine phylogenetic signals and lower the risk of systematic errors, but it cannot explain all inconsistencies, particularly when the actual phylogenetic signal is poor (e.g., for ancient phylogenetic relationships) or the nonphylogenetic signal predominates[5,40,41]. A simplified model may not properly consider different aspects of the evolutionary process, while a complex model may overfit the data, both of which would compromise phylogenetic inferences [75]. In addition, data partitioning allows for the integration of heterogeneity into molecular evolutionary process models, freeing parameter values from collective estimates across all data in a given data set [76]. As a result, data partitioning is commonly used in phylogenetics to account for differences in site-wide substitution patterns, even if the partitioning method chosen has an impact on tree topology, branch lengths, and posterior or bootstrap support for following branches [8,77,78]. Since the selection of a partitioning strategy and a molecular evolution model may have an impact on the outcome of phylogenetic analysis, the model selection process is therefore relevant. The emergence of next-generation sequencing technologies has produced a deluge of molecular data and provides more space for marker selection and stimulates the possibility of a more effective experimental strategy. But, the shortage of an ideal approach for utilizing sequencing data to guide more clade-specific molecular phylogenetic research remains understated.

Most of these studies aim to reconstruct phylogenetic trees of group of species, in which the accumulation of signal from many genes provide enough information to resolve phylogenetic noise and uncertainty in resolving relationship [79–81]. However, very less work has been done on the distribution of topological conflict and concordance among individual gene tree histories. Rather, the conflict between trees built using various methods (e.g., concatenation and coalescence) and subsets of a larger dataset is generally investigated [82,83]. Alternative approaches have become more common as phylogenomics has evolved over the previous decade to address the large quantities of data and inherent gene-tree–species-tree conflicts. These methods comprise the internode certainty (IC) and tree certainty (TC) scores[84–86] Bayesian concordance factors[87], and other concordance measures[88] evaluate gene tree concordance broadly referring to a phylogeny from any subsampled genomic region. We must continue to investigate competing signals within gene trees, not just to properly comprehend species trees, but also because such conflict may give an entrance into a unique way genome's molecular evolution.

The potential sources of conflicting signals include not only technical glitches like among-site rate variation, substitutional saturation, or codon usage bias, but also a wide range of biological phenomena such as horizontal gene transfer (HGT), incomplete lineage sorting (ILS), and hybridization or homoplasy [89,90]. Hybridization can contribute in evolutionary dynamics and adaptive variation transmission, resulting in adaptation to new environments and emergence of novel phenotypes[91–93]. The incomplete lineage sorting (ILS) arise when a huge number of speciation events occur in a short period of time, gene genealogies are unlikely to be fully sorted along evolutionary lineages [94]. These type of reticulation event cannot be modeled by traditional species tree or concatenation methods, hence other algorithms that are computationally intensive are required to investigate if organism relationships are more complicated than bifurcation [95,96].

Given all issues related to phylogenetic analysis, mitochondrial protein coding genes (mPCGs) of 112 Diptera (True Flies) and 4 outgroup species were chosen for this study. Maternally inherited mitochondrial DNA, vulnerable to mutation, low homologous recombination, preserved gene content, and abundant in genetic polymorphism, allowing it to be a popular choice for phylogeography, phylogenetics, and molecular dating studies[97,98,99]. Our preliminary outcome of Dipteran phylogeny shows the short radiation of Scizophora as the crown lineage in recent history as well as the long branching of the distant stem lineages. However, the discrepancy in proper placement of taxon, appearance of long-branch attraction symptom in the phylogenetic tree insisted to look at the deep evolutionary performance of character set used to build phylogenetic tree and the probable reason behind the dispute. However, how effectively a character set of mitochondrial genes deduce phylogenetic resolution in a certain epoch of tree depth is not yet rigorously studied in insect evolution.

Here, in this study First, we begin with the reconstruction of dipteran phylogeny using concatenated all mitochondrial protein coding genes. We implemented codon-site wise partitioning strategies of the dataset in maximum likelihood, maximum parsimony, and Bayesian interface to build a phylogenetic tree. However, none of them could produce complete resolution of the fly tree, we removed the gaps from the alignment and repeated the tree construction techniques. Then we estimated phylogenetic informativeness of different genes and characters from the previously yielded different phylogenetic trees. The phylogenetic informativeness profile provides a spectral view of the informativeness of different genes and their characters across the tree depth and loss of signal towards the root. We investigated different factors behind the phylogenetic signal loss including base compositional heterogeneity, among-lineage codon usage bias, base substitution saturation, among-site rate variation, and divergence rate heterogeneity, all of which are likely to violate phylogenetic assumptions in nucleotide datasets. Further, we analyzed internode certainty and gene tree

discordance in Diptera, as well as possible incidences of mitochondrial introgression and inferred potential hybridization events to examine reticulation patterns.

3.2 Materials and Method:

3.2.1 Sequence alignment, data partitioning and gene evolutionary rate analysis:

All the dipteran mitochondrial protein coding genes (mPCG) sequence data used in this study was obtained from Genbank. Amino acid guided alignment for each mPCGs generated using MAFFT algorithm embedded in TranslatorX server [100,101].The GBlock strict filtering was employed to trim poorly aligned sites from the protein alignment not allowing many contiguous non-conserved positions [102]. Stop codons were removed from the genes by comparing through PAL2NAL webserver [103]. The ribosomal RNA genes were also aligned using Clustal omega [104]. Multiple reports have shown that in phylogenetic analysis, gene rate heterogeneity can be significant, and splitting data onto matrices by set of rate would result in different topologies [105,106]. In order to measure the rate heterogeneity of mitochondrial genes, we calculated average pairwise (p) distance of each gene by MEGA7 [107]. The pairwise (p) distances of each gene were estimated through bootstrap method using 10,000 replicates assuming gamma distributed with invariant sites (G+I) and homogeneous pattern among lineages. Individual aligned mPCGs were concatenated using nexus module of biopython programme.

Former research shows that data partitioning strategies may influence topology and nodal support estimates [76,108]. PartitionFinder was used for assessing optimal partition strategies and the fit of potential nucleotide substitution models by comparing Bayesian Information Criterion (BIC) scores [109]. Individual aligned genes were concatenated using the nexus module of Bio-python programme. First one is raw nucleotide sequence alignment (WG dataset) and second one is the eliminated gap and unaligned nucleotide sequence alignment (GB dataset).

3.2.2 Likelihood Mapping:

To assess the potential influence of taxon sampling, four-cluster likelihood mapping analyses were performed using PUZZLE 5.3 [110]. The Likelihood frequencies were mapped on a triangle partitioned into three possible topologies. The analyses were performed on two dataset raw mitochondrial alignment and GBlock gap trimmed alignment using HKY model and 4 discrete gamma categories and sampled for 10,000 randomly quartets.

3.2.3 Phylogeny reconstruction Homogeneous model:

***Maximum likelihood (ML)*:** All maximum likelihood (ML) analyses were conducted using raxmlHPC-SSE3 version 8.2 [111]. We set ML analyses each dataset using bootstrap replicates 10000 for attaining optimal tree. Primarily the codon partitioned (3X13) concatenated gene set was tested under the general time-reversible nucleotide model with gamma-distributed rate heterogeneity and invariant sites (GTRGAMMAI) assuming compositional homogeneity among the taxa [39,112]. Each dataset we ran two ways first unconstraint topology second constraint topology (restricting outgroup and some taxa into their desired family).

Bayesian inference (BI): All Bayesian phylogenetic analyses were conducted using MrBayes version 3.2.7 [113]. The concatenated 13 protein coding genes partitioned codon positions of all genes, using partition finder of 116 species. We performed single analyses of 200 million generations each (depending on dataset), consisting of two independent runs with 4 chains per run and the first 25% discarded as burn-in. The convergence of the runs was assessed by checking the potential scale reduction factor (PSRF) values of each parameter in MrBayes and the Effective Sample Size (ESS) values of each parameter in Tracer 1.6 [114]. Values of PSRF close to 1.00 and ESS above 200 were considered as good indicator of convergence. Nodes recovered with posterior probabilities (PP) $\geq$ 0.95 are considered to be strongly supported, standard deviation of split frequencies (< ~0.05), and potential scale reduction factors (~1.0).

Maximum Parsimony: Parsimony analyses were conducted using PAUP 4.0b10 [115], with a heuristic search with 100 random-addition-sequence replicates. Each heuristic search was done with tree-bisection-reconnection (TBR) branch swapping algorithm, and holding a single tree per replicate. Clade support was assessed using nonparametric bootstrapping, with 1000 pseudoreplicates. Each pseudoreplicate consisted of 20 random-addition-sequence replicates (with TBR branch swapping), holding, and saving only one tree per replicate (although a given pseudoreplicate may ultimately save multiple equally parsimonious trees). However, this constraint was not possible in the bootstrap analyses.

Partition models that allow genes to evolve with different substitution models have been adopted to exhibit different evolutionary scenarios [116]. Three types of partition models (a) Edge-Linked-equal: model with equal branch lengths, all partition share same set of branch length; (b) Edge-Linked-Proportional: each partition has its own partition-specific rate, which adjusts all its branch lengths, offering variable evolution rates across partitions and c) Edge-Unlinked: most parameter-rich partition model, every partition has its own set of branch lengths[117]. Here in IQTree we executed the Edge-Linked-proportional and the Edge-Unlinked model using CIPRES Gateway (http://www.phylo.org/). Empirical base-frequency counts derived directly from the alignment by the GTR+F method and the Γ-distributed rates were calculated using a four-category approximation using the + G4 method. Tree branches were tested using an SH-like aLTR, a local BP check of 1000 replicates and aLTR, aBays parametric test.

3.2.4 Phylogenetic informativeness profiling:

Phylogenetic informativeness (PI) is an approach for measuring the information quality of characters in terms of its evolutionary rate[7]. The informativeness is obtained by calculating the probability that a character would exhibit a single change on the internal branch of a symmetric four-taxon tree but no change on the four external branches. This approach analyses a data set's

ability to resolve the phylogeny at different timelines and taxonomic levels by integrating information from all the characters in a data set or all the base pairs in a gene[7]. Here, net and site-specific PI were calculated for each dataset (raw and gap removed) using an ML technique performed on the PhyDesign webserver with the HyPhy software [118,119]. HyPhy provides invariable positions with a zero rate, which is required for a conservative estimation of the PI profiles at ancient timescales[7]. For the initial evaluation of the phylogenetic informativeness of each gene and codon position, we used different trees formed by maximum likelihood, bayesian inference and maximum parsimony methods. Since it has been demonstrated that PI profiles may be employed efficiently even when the divergence period is unknown, we adopt a relative timeframe by assigning a value of 1 to the whole tree length from the ingroup's basal node to the tips[7,120].

3.2.5 Evaluation of substitutional saturation, codon usage bias, sequence composition and divergence heterogeneity:

Phylogenomic data is the concatenation of a set of distinct loci, each specified by the evolutionary constraints that shape its substitution pattern, which may lead to homoplasious substitutions that may mislead phylogenetic inferences [39]. Processes that cause homoplasy in nucleotide data can involve substitutional saturation among lineage, compositional heterogeneity or codon usage bias [40,41,121–123]. Substitution saturation was assessed by plotting raw number of transversions and transitions against F84 and GTR distances and noting whether a plateau is attained using DAMBE software [124]. The saturation level is calculated by measuring the slope of the regression line in the graph, the greater the slope of the regression line in the graph, the greater the slope, the higher the degree of saturation. Saturation is calculated for two subsets of the concatenated data: the combined first and second positions and the third codon positions separately to evaluate the impact of saturation on phylogenetic reconstructions. Additionally, we calculated index of substitution saturation (*Iss*) and critical *Iss.c* using Xia's

method as implemented in DAMBE [32]. If *Iss* is not smaller than *Iss.c*, then we can conclude that the sequences have experienced severe substitution saturation and should not be used for phylogenetic reconstruction.

The basic composition of each gene and each codon position was obtained from the alignment produced by the TranslatorX program. The nucleotide percentage of each species was plotted to demonstrate the compositional heterogeneity among lineages. To evaluate the influence of base compositional heterogeneity we used two separate methods. First, we measured the pairwise disparity index (I_D) for all 13 protein coding genes together, I_D calculates the observed substitution pattern differences for pairs of sequences while indirectly estimating the level of base composition heterogeneity. With 1000 replications deployed in MEGAX, we assessed the substitution pattern homogeneity (I_D -test) using the Monte Carlo Method [125]. We measured the probability of rejecting the null hypothesis that sequences have evolved with the same pattern of substitution at alpha < 0.001. However, it argued that this method should be used with caution, because site homology is not properly considered when calculating I_D [17]. Thus, another approach, known as the matched-pair test of symmetry, as implemented in SeqVis, was used to analyze the similarity of specific sites and tests against the null hypothesis that a pair of sequences formed under the same conditions[126,127].

The sequence divergence heterogeneity was assessed by AlIGROOVE with the default sliding window size, providing an approximate predictor of a specific sequence or the heterogeneity of a clade in terms of the entire dataset [128,129]. Indels in nucleotide datasets are considered as ambiguity and the BLOSUM62 matrix was used as the default amino acid substitution matrix. The metric establishes pairwise sequence distances between specific terminals or subclades with terminals outside the focal category. The distances are then compared with the distances across the entire data matrix, and the metric values range from -1 (distances differ from the

average for the entire data matrix) to + 1 (distances equivalent to the average for the whole matrix) [129].

To evaluate synonymous codon bias among different dipteran lineage, the Relative Synonymous Codon usage (RSCU) values were estimated using CodonW software (version 1.4.2, http://codonw.sourceforge.net/). Effective number of codon (ENc) designates the degree of codon bias for genes; where it computes departures from uniform codon usage without any prior dependency over the sequence length or specific information of preferred codons, although it might have influenced by base composition[130]. Over a range of values from 20 to 61 lower values indicate greater codon bias[131]. In General occasion, ENc values lesser than 36 signifies strong codon bias. Here in this study, we have followed calculation of Nc from the study of Sun, X. et al. at 2012 and estimated through DAMBE 6 software [124,132].

3.2.6 Phylogenetic analysis through different Heterogeneous model:

Irregular nucleotide compositions are an essential and possible source for misleading tree-reconstruction methods, even though strong bootstrap values are obtained. We explored another method to deal with base compositional heterogeneity namely LogDet transformation that is compatible with sequences of varying nucleotide composition that emerged under basic but asymmetric stochastic evolutionary models [15]. We performed Neighbor-joining and further heuristic search applying 1000 bootstrap replicates via LogDet or paralinear distance using the concatenated mitochondrial genes deployed in PAUP 4b10 [115].

Recent phylogenomics research has demonstrated the potential of site-based heterogeneous models (e.g. CAT-based models) to mitigate artefacts caused by mutational saturation and irregular substitution patterns, which are major challenges in the analysis of genomic data and ancient events [133]. To test for among-site compositional heterogeneity in the sample dataset we performed MCMC analysis using CAT+Poisson and CAT+GTR model implemented in

Phylobayes after systematic removal of constant sites from the alignment[72]. Minimum number of cycles 20,000; minimum effective size 1000; initial 500 cycles excluded from convergence checks.

To determine the impact of nucleotide substitution variations influencing the observed codon-usage bias, degenerate codon sites of redundant nucleotide sequence were reduced to a single triplet in synonymous codons. Hence, a separate three-state general time reversible model (GTR3) was employed, where C and T are fused into a single merged pyrimidine state Y and the hidden transition between the two pyrimidines C and T is exempted from this calculation[18]. We have also applied a two-state model (GTR2) for nucleotides where T with C is combined into one common state T and A with G into another common state A, in certain situations causing the base composition to appear more stationary. This model is sensitive to transversions only and the two frequency parameters are expected in the order {T, A}. There are no relative rate parameters, the corresponding parameter list is {}. Both of this model was applied in Treefinder tool and concatenated 13 PCGs was used to infer phylogenetic tree [134].

Supermatrices for phylogenetic inferences are unable to cope with heterogeneous evolution due to heterotachy, since evolutionary rates vary across sites and lineages over time, resulting in possible contradictions from complex evolutionary models in multi-gene data sets [23,135]. To overcome this matter, we used the General Heterogeneous evolution On a Single Topology (GHOST) model [136]. GHOST is a non-data partitioned edge-unlinked mixture model that accounts for heterotachous evolution by combining several site classes on the same tree topology, each with its own set of model parameters and edge lengths. Considering the heterogeneous evolution, GHOST model was applied in both supermatrices with nucleotides of raw alignment (WG) and trimmed alignment (GB). For each dataset we invoked analysis using GTR+FO*H4 model in linked tree topology and unlinked branch lengths, substitution rates and inferred base frequencies with 1000 bootstrap replicates.

3.2.7 Estimation of homoplasious site and Supernetwork and Neighbour-Net analysis:

Phylogenetic networks simplify phylogenetic trees because they permit the representation of conflicting signal or alternative phylogenetic histories [137]. Since recombination, hybridization, gene transfer, and gene flow which result in histories that are not properly modelled by a single tree, there is a clear need to use networks instead of branching trees when the evolutionary history is not treelike; even if the underlying history is treelike, parallel evolution, model heterogeneity, and sampling error make defining a tree difficult [138]. Networks may provide a powerful method in such situations to reflect uncertainty or to visualize a space of feasible trees.

A homoplasy is a character shared across clades in a phylogeny that may not have a common ancestor, is an indicator of inconsistency between the phylogenetic tree and the sequences used to construct it. Throughout the generation and analysis of sequencing data, homoplasy on a phylogeny may be formed spontaneously, through convergent evolution or recombination, or unnaturally[36]. Before the phylogenetic relationships in a tree are interpreted, it is crucial that any inconsistencies between the sequencing data and the tree are identified. Here we used HomoplasyFinder to identify homoplasious site present in the sequence [36].

Single-gene phylogenies usually give rise to uncertainties in phylogenetic analysis which may be attributable to i) stochastic errors due to inadequate data or ii) violation of orthological assumptions caused by processes like incomplete lineage sorting [41,139]. However, given that mitochondrial genomes maternally inherited, no lineage sorting problem is expected. To evaluate the consistency among single genes, Maximum-likelihood (ML) analysis were performed for each protein coding gene alignment deduced from TranslatorX using RaxML under GTRGAMMAI model using 5000 pseudo replicates for bootstrap support was estimated. To visualize conflicts among genes, we constructed a Supernetwork for 13 genes using SplitsTree5[140]. As single gene trees usually lack statistically significant support, to reduce the

risk of overestimating the conflicts among single gene trees. When calculating the supernetwork, we used the Z-closure option, mean edge weights, set splits transformation as equal angle and left all other parameter as default settings[141].

Neighbor-Networks works in a unique way: it creates a set of weighted splits first, and uses a split graph to represent such splits. Neighbor-Net is a distance-based strategy equivalent to Pyramid Clustering and Split Decomposition, and it has the advantage of assisting in the development of far more settled networks than those generated by split decomposition [142]. Neighbor-Net is available as part of the SplitsTree 5.0 software package and here in this study we used concatenated 13 mitochondrial protein coding genes of 116 taxa [140]. The large number of sequences and small number of locations in such data contribute to sampling error, resulting in significant homoplasy between the sequences, as well as the data's relative lack of detail. This is an ideal situation for a network study so we can deduce systematic features without focusing on a single tree [138].

3.2.8 Detection and Visualization mitochondrial Gene Tree Discordance:

We first calculated the internode certainty all (ICA), which quantifies the degree of disagreement on each node of a target tree (i.e. species tree estimates) given individual gene trees in order to investigate discordance between gene tree and species tree estimates [143]. We identified the number of conflicting and concordant bipartitions on the species trees. ICA values ~1 indicate strong concordance in the bipartition of interest, while ICA values close to 0 indicate equal support for one or more conflicting bipartitions and negative ICA values imply the bipartitions of interest conflict with one or more bipartitions that have a higher frequency, and ICA values close to -1 indicate the absence of concordance for the bipartition of interest [143]. We calculated ICA and the number of conflicting/concordant bipartitions with PhyParts [144], using the estimated ASTRAL species trees as the map tree and the individual gene trees. The gene trees were build using dnaML program of Phylip package. Additionally, in order to

distinguish strong conflict from weakly supported branches, we carried out Quartet Sampling (QS) with 100 replicates[145]. QS subsamples quartets from the input tree and alignment and assesses the confidence, consistency, and informativeness of each internal branch by the relative frequency of the three possible quartet topologies[145]. The ICA and QS scores both allow for an alternative branch support that indicates underlying gene tree conflict and is unaffected by the excessively high levels of bootstrap support found in phylogenomic data[146]. We then explored individual gene tree resolution by calculating the Tree Certainty (TC) score in RAxML using the majority rule (majority rule (MR), extended majority rule (MRE) and majority rule bipartitions with ≥ 75% support) consensus tree across the 100 bootstrap replicates[84,143].

3.2.9 Species Network Analysis with the Reduced Data Set:

We inferred the species networks for modeling of ILS and gene flow using three distinct approaches in PhyloNet v.3.6.9 namely maximum parsimony (MP), maximum pseudo-likelihood (MPL) and maximum likelihood (ML) [96,147,148]. Due to computational restrictions, and given our focus to identify potential reticulating events among major suspected clades from previous neighbour net analysis we reduced our taxon sampling to 10 ingroup taxa. The 13 gene trees were build using dnaML program of Phylip package and then drop.tip function from ape package was used to reduce the tree size (https://cran.r-project.org/web/packages/ape/index.html). The MP and MPL method were invoked through InferNetwork_MP, InferNetwork_MPL specifying maximum 3 reticulation performing 20 runs by examining 1000 networks during the search. The ML analysis was reduced due to its high computation expense and invoked by InferNetwork_ML using 3 reticulations performing 5 runs by examining 100 networks during the search.

3.2.10 Hypothesis Testing and Detecting Conflict Using Four-Taxon Data Sets:

Given the signal of multiple clades potentially involved in hybridization events detected by PhyloNet (see Results), we next conducted quartet analyses to explore a single event at a time. First, we further reduced the 10-taxon(net) data set to three taxa additionally we included one outgroup (*B. mori*). We created twenty such incidences with four-taxon combination using drop.tip function of ape package in R using gene trees produced by dnaML program of phylip package. Again, ASTRAL was used to estimate species trees from the four-taxon gene trees and gene tree conflict explored with PhyParts[144]. We then explored resolution of individual four-taxon species trees using the gene trees by calculating the Tree Certainty (TC) score in RAxML using the majority rule consensus tree across the 200 bootstrap replicates[143]. We carried out 20 constraint searches for each of three topologies in RAxML with the GTRGAMMA model, then calculated site-wise log-likelihood scores for the three constraint topologies in RAxML using GTRGAMMA and carried out the AU test using Consel v.1.20[149].

3.2.11 Test of introgression:

The advent of genomic data has revealed the exchange of genetic material between numerous species [150]. In conjunction with the rapid growth of genomic resources, scientists have developed many statistical tests to detect introgression, such as the "ABBA/BABA" test [151,152]. This method was first used to calculate the extent of genetic exchange across Neanderthals and homo sapiens [151,153]. The concept behind this test is that it considers ancestral ('A') and derived ('B') alleles across the genomes of four taxa. In the absence of introgression, the allelic patterns ‘ABBA' and ‘BABA' should appear equally. Gene flow between two taxa is shown by an abundance of either ABBA or BABA, resulting in a *D*-statistic that is greatly different from zero. Introgression between P2 and P3 is indicated by a positive *D*-statistic (i.e. excess of ABBA), while introgression between P1 and P3 is indicated by a negative *D*-statistic (i.e.

excess of BABA) (Fig. 3.1A,B). *D*-statistics are based on comparing the proportions of ABBA and BABA sites patterns observed in the data:

$$D = \frac{\#\ \text{ABBA sites} - \#\ \text{BABA sites}}{\#\ \text{ABBA sites} + \#\ \text{BABA sites}}$$

A Z-score has used to determine the significance of a *D*-statistic, with a Z-score greater than 3 or less than -3 indicating a significant result. We created 10 scenarios of 4 taxa assuming *E. sorbilans, P. steinenii, P. vanderplanki, S. grandicornis* as hybrid species and *B. mori* as outgroup species. The *D*-statistic and associated z score for the null hypothesis of no introgression (D=0) was conducted on GBlock concatenated alignment through evobir package in R 4.0.2 by CalcD function using 1000 bootstrap replicates for estimating variance[154].

We also estimated *D*-statistic and f4-ratio statistic through ADMIXTOOLS software using admixr package in R 4.0.2 [155,156]. ADMIXTOOLS software needs data in EIGENSTRAT format that was converted using "convertf" utility of EIGENSOFT. The *f*4-ratio statistic generally used to detect the proportion of ancestry shared by any potential admixed species. For instance, the phylogeny of Figure 3.1C where, the population X is an admixture of populations B' and C' (possibly with subsequent drift) and we have genetic data from populations A, B, X, C, O. Thus, we would normally compute a *f*4-ratio statistic. Here, the estimates in both numerator and denominator are obtained by summing over many SNPs.

$$f4 - \text{ratio} = \frac{f4(\text{A}, \text{O}; \text{X}, \text{C})}{f4(\text{A}, \text{O}; \text{B}, \text{C})}$$

Additionally, the *f*-branch statistic was calculated in the reduced ten species dataset to determine the number of possible gene flow between donor-recipient combinations. Many high *D*-statistics and *f*4-ratio values can be attributed to a single gene flow event. At the same time, correlations, mainly *f*4-ratio scores, can provide insight into the timing of introgression events and specific donor-recipient combinations. The *f*-branch or *f*b metric was introduced in

Malinsky et al. (2018) to unravel correlated *f*4-ratio results and assign gene flow evidence to specific, possibly internal, branches on a phylogeny by building upon the logic developed by

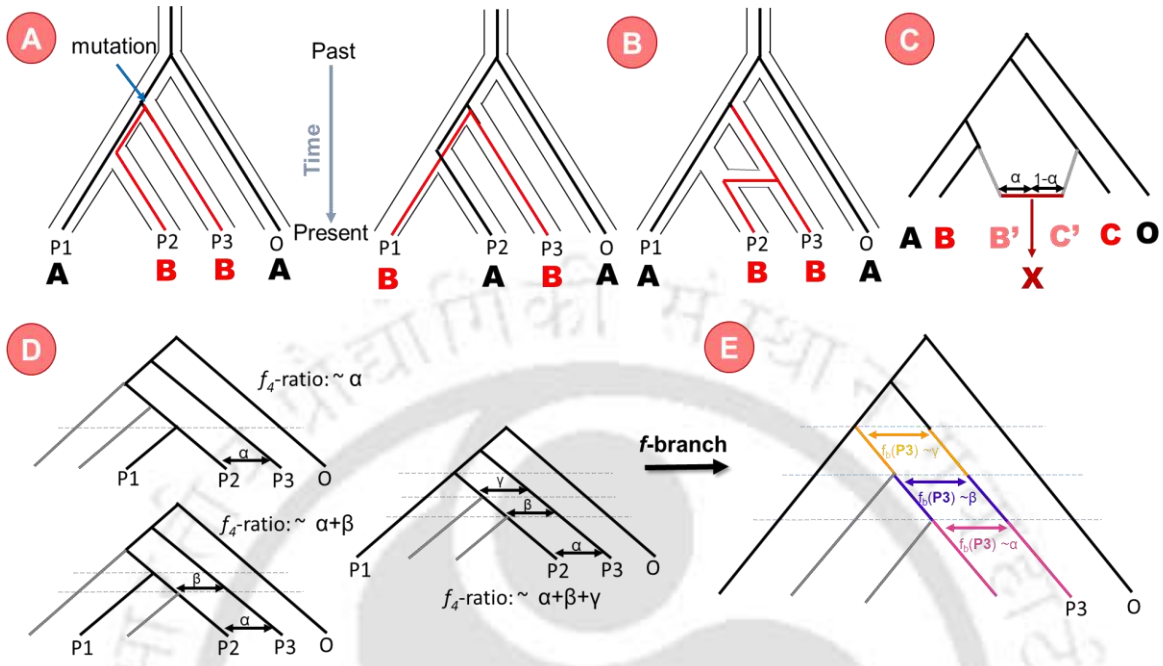


Figure 3.1: Basic principles behind the *D* and *f*-branch statistics. (A) Example genealogies illustrating the sharing of derived alleles, labeled as 'B' across populations P2 and P3 (the ABBA pattern) and P1 and P3 (the BABA pattern) because of incomplete lineage sorting. Both patterns are thought to be equally possible in the absence of gene flow (B) Gene flow between P2 and P3 provides additional loci with ABBA patterns, which would resulting in a positive *D* statistic. (C) A basic phylogeny for explaining *f*4-ratio estimation. (D) Interdependences between distinct *f*4-ratio scores are illustrated in this example, which can provide information on the time of introgression. Different choices for the P1 population in this case offer constraints on when gene flow may have occurred. (E) The *f*-branch, or *f*b statistic, distinguishes between admixture at different time periods by allocating signals to distinct (potentially internal) branches in the population/species tree, based on relations between the *f*4-ratio results from different four taxon tests. (This figure is followed from Malinsky, M. et. al 2021)

Martin et al. (2013), as shown in Figure 3.1D, E [157,158]. The *f*b(P3) statistic represents excess allele sharing between the population or species P3 and the descendants of the branch designated b, compared to allele sharing between P3 and the descendants of the sister branch of b, given a certain tree (with known or hypothesized relationships). Formally:

$$f\mathrm{b}\,(\mathrm{P3}) = \mathrm{median}_{\mathrm{A}}\,[\min_{\mathrm{B}}\,[f4\mathrm{ratio}\,(\mathrm{A},\mathrm{B};\mathrm{P3},\mathrm{O})\,]]$$

Where, B denotes populations or taxa descended from branch b, and A denotes descendants from branch b's sister branch. All positive *f*4-ratio values with A in P1 and B in P2 are included in the computation.

3.3 Result and Discussion:

3.3.1 Likelihood mapping:

We calculated quartet weights using the Tree-Puzzle tool, which computes a posterior likelihood for each quartet based on the Hasegawa–Kishino–Yano substitution model[159,160]. Figure 3.2 shows the corresponding "likelihood mapping"—a depiction of the phylogenetic content of a sequence alignment in which each point represents a weighted quartet[160]. The likelihood mapping reveals that most quartets are mapped at the triangle's three vertices, implying tree-like behavior with a positive phylogenetic signal since the sum of unresolved and partially resolved regions is less than 10%.

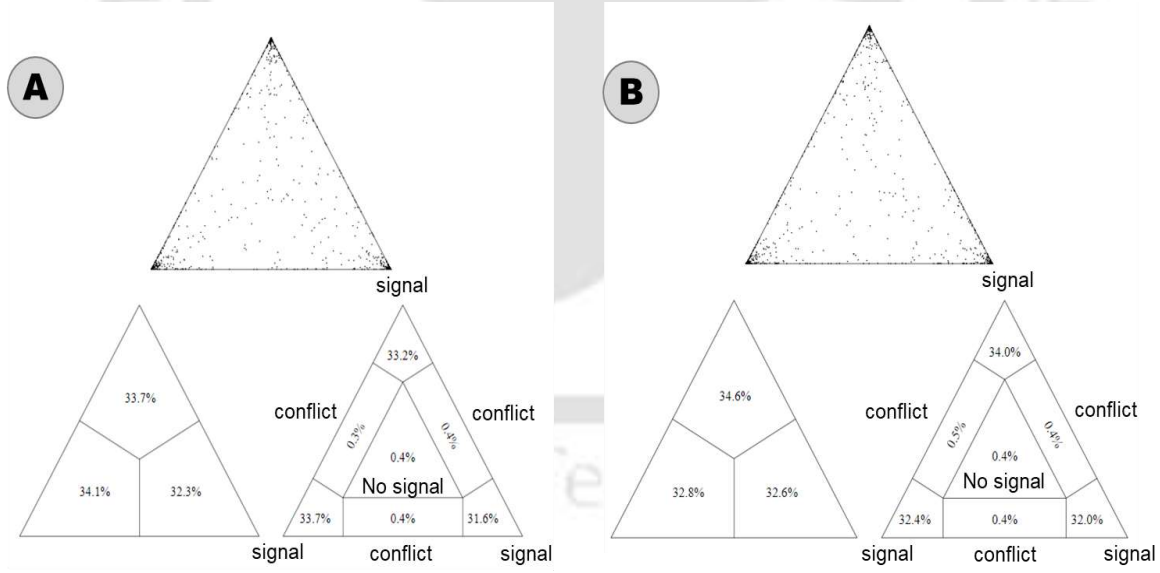


Figure 3.2: The Likelihood mapping of sequence alignment of mitochondrial protein coding genes, A) Raw Alignment, B) Gblock Alignment. The concentration of points at the ends of the triangles on the top, as well as the sum of the percentages (>90%) at the ends of the triangles on the bottom-right, indicate the good quality of the phylogenetic signal of the investigated genes.

3.3.2 Phylogenetic results under homogeneous models:

Here we present complete Diptera phylogeny using mitochondrial protein coding genes of 112 flies and 4 out group species from different orders. Initially, we found phylogenetic relationships using raw12PCG123 datasets through RaxML 8.2.x and MrBayes 3.2.7, as detailed here (Fig. 3.3).

Homogeneous analysis of nucleotide data of protein coding genes produced non-monophyly of Diptera flies by both Bayesian Inference (BI) and Maximum Likelihood (ML) methods as Cecidomyiidae family clustered with outgroups. Although, linking node of remaining Diptera have strong support from ML and BI analysis (ML/BI: 97/1.00). Whereas, its major groups were found to be monophyletic, with Brachycera (100/1.00) receiving total support from ML and BI and Schizophora (66/1.00) obtaining only partial support from ML but complete support from BI. In Calyptratae, families including Calliphoridae (100/1.00) and Sarcophagidae (100/1.00) of the Oestroidea superfamily were found to be monophyletic in this study. This analysis, however, was unable to recover the Tachinidae and Oestridae families as monophyletic. One of the most captivating observations from this study is that *Exorista sorbillans*, which is traditionally assigned to the Tachinidae family of Calyptratae, is always found to be a sister group of acalyptrate Drosophilidae family with strong support from ML and BI both (100/1.00). Whereas, Oestridae family established paraphyletic relation and located as sister clade of other Oestroidea superfamily in BI analysis and in ML two species (*H. lineatum*, *D. hominis*) grouped with Tachinidae with relatively low bootstrap support (81). Presence of Muscoidea superfamily within the Oestroidea superfamily with poor bootstrap support (39) from ML analysis strong posterior probability (1.00) from BI is evident from this study. Muscoidea situated at paraphyletic position with both Calliphoridae and Sarcophagidae family and the Muscidae family recovered as monophyletic in both ML or BI analyses (73/1.00). In Acalyptratae, Drosophilidae (100/1.00) and Tephritidae (100/1.00) each family

of fruit fly show monophyletic relation. Opomyzoidea superfamily also appears monophyly with low bootstrap support (ML:56) and its family Fergusoninidae exist as sister lineage of Agromyzidae. But in BI, Fergusoninidae grouped with Syrphidae family and located as sister to Schizophora.

In this analysis Nematocera found as the sister group in the stem lineage with paraphyletic relation with remaining Diptera flies, although they are unable to maintain traditional taxonomic hierarchy at the superfamily level. *Arachnocampa flava* conventionally belongs to Sciaroidea superfamily but here in this study it clustered with *Cramptonomyia spenceri* of Pachyneuroidea superfamily in ML analysis (92) and *Sylvicola fenestralis* from Trichoceroidea as sister to that group (60). Whereas, in BI *A. flava* grouped with Culicoides arakawae of Chironomoidea with low posterior probability 0.59 and located as sister to majority of Diptera. Other two species of Trichoceroidea superfamily namely *Paracladura trichopteran* and *Trichocera bimacula* grouped with Tipuloidea superfamily (87/0.72). The *Orseolia oryzae* and *Rhopalomyia pomum* of Cecidomyiidae family of Sciaroidea superfamily shows prominent long branching attraction (LBA) effect and grouped with the outgroup species. Families such as Culicidae (100/1.00), Simuliidae (100/1.00), and Psychodidae (100/1.00) exhibit monophyly within the suborder Nematocera, whereas families such as Chironomidae do not. *P. steinenii* is traditionally classified as a member of the Chironomidae family, but based on ML and BI analysis, this Chironomid fly is grouped with the Simuliidae family (65/1.00) where both of this family belong to Chironomoidea superfamily. Whereas, other two members of Chironomidae family (*C. tepperi* and *P. vanderplanki*) grouped together and located as sister to rest of Diptera along with *C. arakawae*, a member of Ceratopogonidae and both families belong to Chironomoidea superfamily. These two assemblages of the Chironomoidea superfamily form a polyphyletic relationship with each other even after being in different

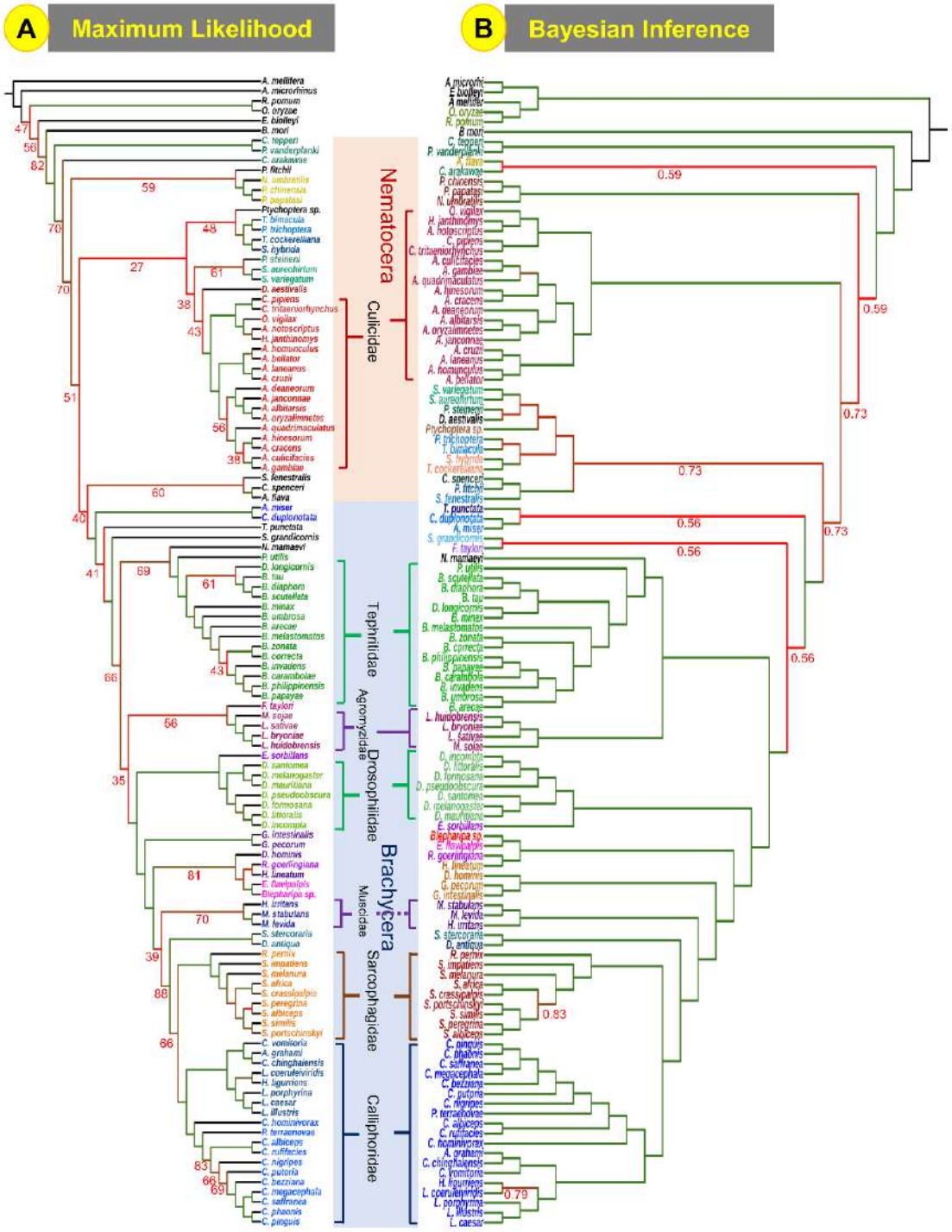


Figure 3.3: Complete phylogeny of Diptera flies inferred from A) Maximum Likelihood method implemented in RaxML v. 8. and B) Bayesian method implemented in MrBayes 3.2.7 using mitochondrial PCGs of 112 Diptera species and 4 outgroup species. Green branches are showing maximum bootstrap support (bs) or posterior probability (pp) (100/1.00); Red branches are minimum bs and pp. The monophyletic families are mentioned in the figure.

positions of phylogeny in both ML and BI. The topological resolution of this phylogeny, however, is not without congruence, most notably the proximity of a tachinid fly *Exorista sorbilans* to the Drosophilidae family, the altered position of *P. steinenii* to other Chironomidae flies, and the closeness of the cecidomyiidae family to the out-group.

In addition, bootstrap support was not adequate at backbone nodes of some major clades of Diptera. After the removal of gaps (GB dataset) from the alignment as well as analysis with other methods also show their incompetency to extract the proper phylogenetic resolution from these two datasets. Further we specified constrained mainly in two families one from Brachycera (Tachinidae) and another from Nematocera (Chironomidae) and repeated the analysis using same partitioning strategy. In ML analysis it improved the bootstrap support for multiple clades in Nematocera suborder which was previously supported loosely.

3.3.3 Profiling of phylogenetic informativeness (PI):

We estimated Phylogenetic informativeness (PI) to reflect the resolution of signal detected by the datasets' phylogenetic tree building techniques. The PI profiles of the 13 protein coding genes and their partitions are shown in Figure 3.4. The graphical profiles of the phylogenetic informativeness of 13 genes and 39 codon partitions with different phylogenetic trees provided by homogenous models of ML and BI analysis revealed a diversity of informativeness. The nucleotide sequences of 13 genes demonstrate differential power for resolution of ancestral branching order. We found that there were obvious qualitative differences between the PI profiles of 3 codon positions between different genes reflecting their different evolutionary constraints imposed on them. While different nucleotide positions show variation in informative power and 2nd codon positions of all PCGs show relatively stable informativeness throughout the tree depth. It is apparent that the 3rd codon positions exhibit high informative power at nearer to the tip, as the profile of the 3rd codon sites have a firm peak in the recent past (Fig. 3.4).

The sites residing at 3rd position at each codon which must withstand substitution without changing the amino acid sequence yield phylogenetic informativeness for recent divergences, while 1st and 2nd codon positions generate only PI for ancient divergences. The PI profile of unconstrained BI analysis using WG dataset shows that all sites in the codon have greater informativeness closer to the tip, and it decreases abruptly. A similar profile also observed in IQ edge proportional trees. It implies that these two methods with WG dataset could not provide substantial phylogenetic signal towards the root of the tree. The Diptera phylogenetic tree build by ML method shows relatively better phylogenetic informativeness. Different trees with GBlock alignment have more stable informativeness across the tree depth. We also noticed that after constraining the tree in various tree building approaches, the informativeness improved. The profile generated from BI and IQ tree presented in Figure show bimodal distribution of rates could correspond to synonymous changes of any gene[10].

3.3.4 Assessment of Different Heterogeneity within dataset:

Rate Heterogeneity: The amino acid guided aligned matrix consists of 116 taxa and 13 mitochondrial protein coding genes for the 112 Diptera species and 4 Outgroup have 11835 bp characters and 9018 bp charterers after the removal of gaps and missing data. We investigated the distribution of Transitions and Transversions variation across the 13 coding sequences as well as combined dataset after removal of stop codons. To determine the sequence variability of the mitochondrial genes from samples p-distances of the sequences in the nucleotide levels were calculated. The average p-distances of the 13 individual mitochondrial gene alignment ranged from 0.173%, 0.313%; and eliminating outgroup the p-distance varies 0.166% to 0.302% and *atp8* shows most variation and *cox1* is the least. Again, based on p-distances no clear gene rate category could be identified. Therefore, the dataset was not partitioned or split based on the rate category of genes in the subsequent analysis.

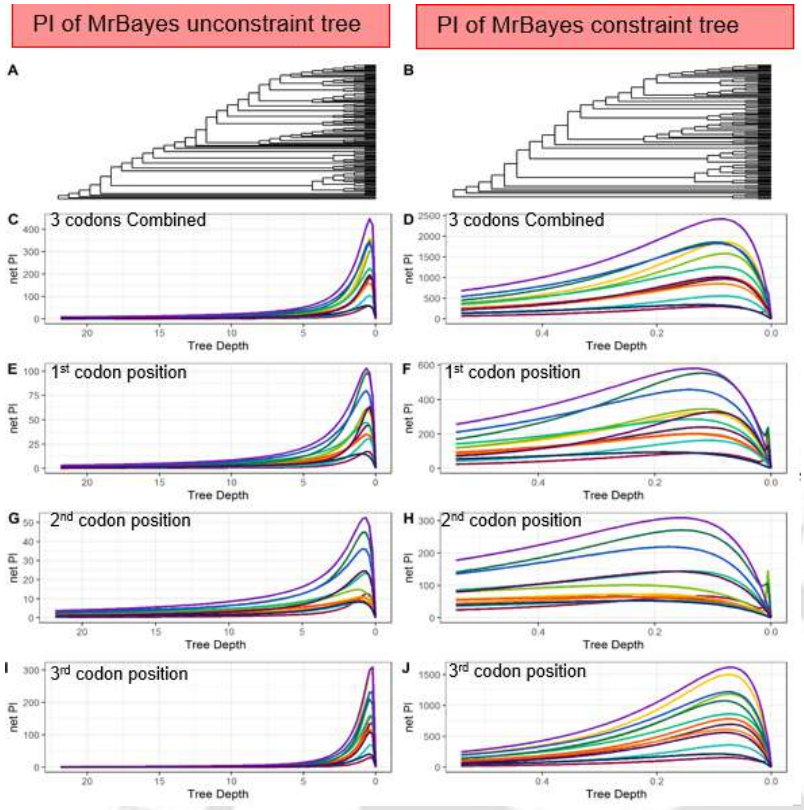


Figure 3.4 i: (A) PI of MrBayes unconstrained tree for different codon positions. (B) PI of MrBayes constrained tree for different codon positions. We restrict us to continue further MrBayes analysis with GB dataset as the MrBayes analysis is computationally very expensive and unable to provide desirable result. (200 million generations took almost 19 days).

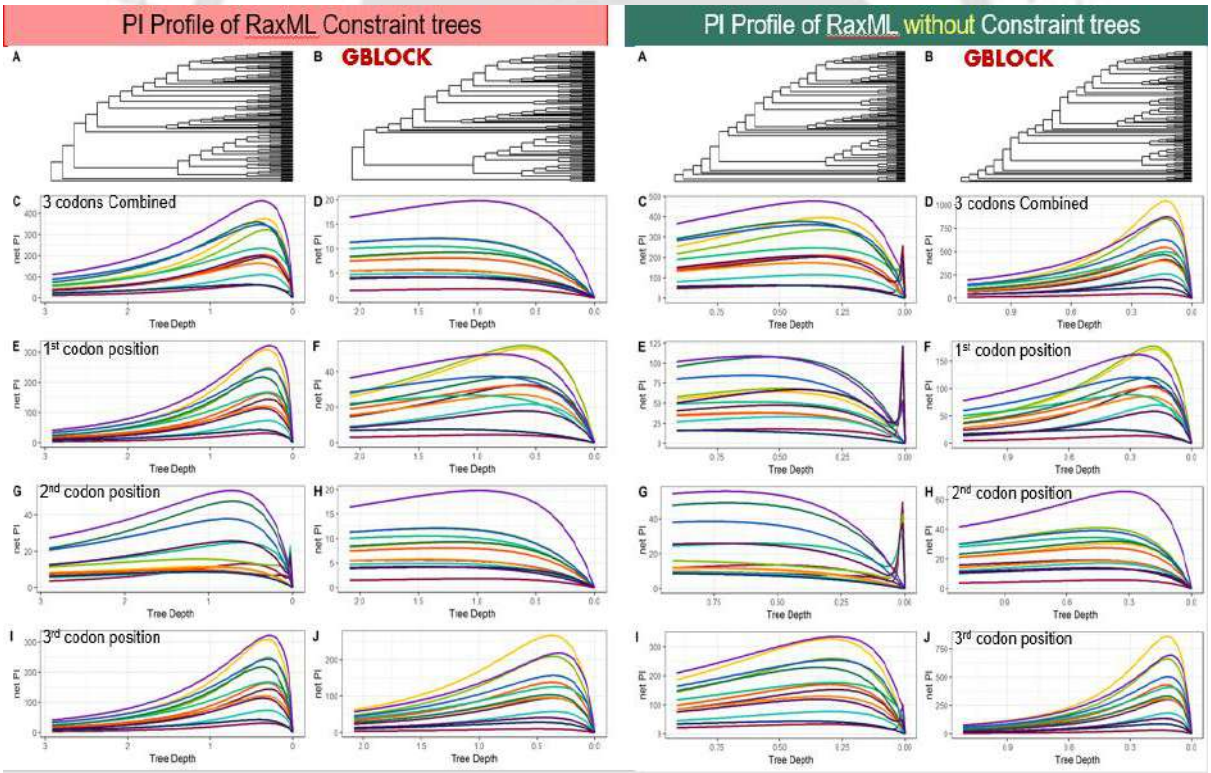


Figure 3.4 ii: PI profile of RaxML constraint tree (A) WG dataset (B) GB dataset

Figure 3.4 iii: PI profile of RaxML unconstraint tree (A) WG dataset (B) GB dataset

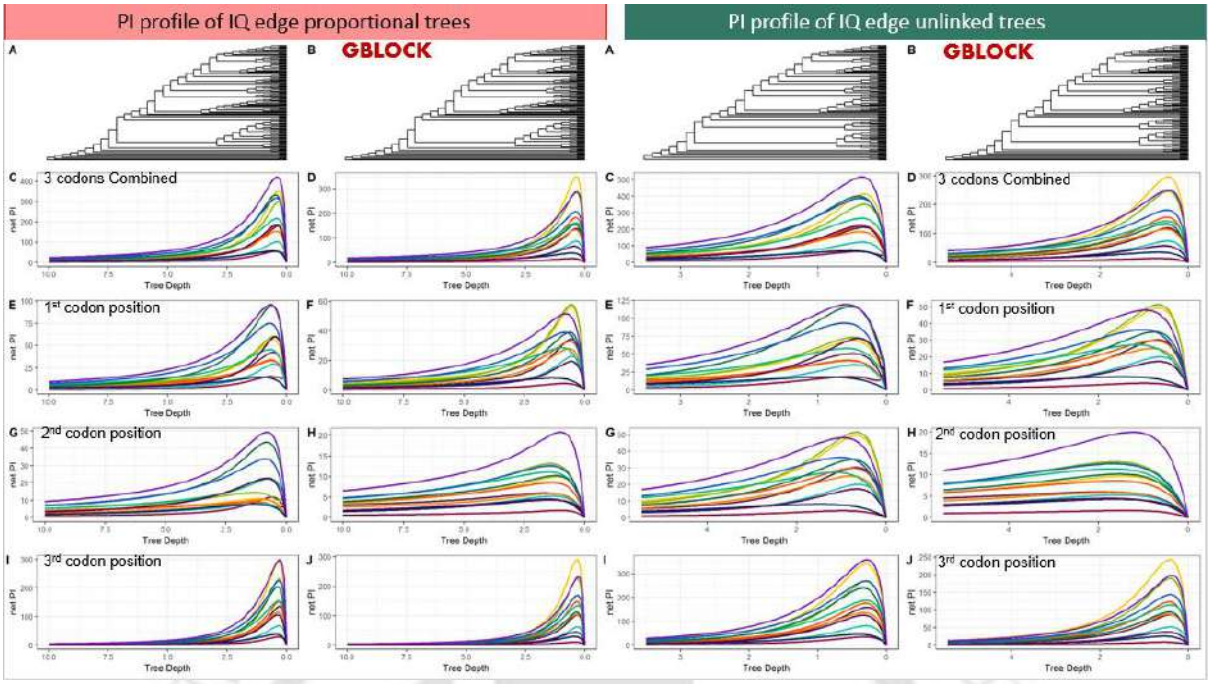


Figure 3.4 iv: PI profile of IQedge proportional tree (A) WG dataset (B) GB dataset

Figure 3.4 v: PI profile of IQedge unlinked tree (A) WG dataset (B) GB dataset

Pairwise comparison of the disparity index (I_D) among dipteran species presented that closely related taxa generally had a lower I_D than distantly related taxa. For instance, I_D between two Culicidae species *A. cracens* and *A. albitarsis* is 0, whereas that between *A. cracens* and distantly related oestroid species *G. pecorum* was 33.989. Within Diptera *B. minax* have the most divergent base composition (mean I_D= 72.002), followed by *O. oryzae* (52.894), *R. pomum* (47.128), *B. umbrosa* (38.772) and 76 species among 112 exhibited I_D value more than 6 and none of them have lower than 5. Indicating that overall dataset has high level of base compositional heterogeneity. The I_D calculated from individual codon position suggested that level of base compositional heterogeneity lowest in 2nd position, ascended by 1st and 3rd position. Further the matched-pair tests of symmetry based on the nucleotide dataset and the individual codon position suggested that the assumption of stationarity, reversibility and homogeneity were not fulfilled for either data set (Table 3.1). Based on this test the ranked order of the datasets from best to worst was: 2nd position > 1st position > 3rd position > 123

position. The matched-pair test suggests that evolution at all three codon positions is not necessarily stationary, and thus cannot be both reversible and homogeneous, which could explain why phylogenetic artefacts exist[127].

Table 3.1: Matched-paired test of Symmetry; Null Hypothesis: A pair of sequence has evolved under same conditions

p values interval	1st codon position		2nd codon position		3rd codon position		123 position	
	number	proportion	number	proportion	number	proportion	number	proportion
0.00001	3214	0.47362	1198	0.17654	4196	0.61833	4586	0.6758
0.00005	3490	0.51429	1369	0.20174	4392	0.64721	4794	0.70645
0.0001	3618	0.53316	1444	0.21279	4485	0.66092	4876	0.71854
0.0005	3958	0.58326	1693	0.24948	4723	0.69599	5117	0.75405
0.001	4118	0.60684	1814	0.26732	4848	0.71441	5230	0.7707
0.005	4538	0.66873	2263	0.33348	5199	0.76614	5487	0.80858
0.01	4760	0.70144	2509	0.36973	5362	0.79016	5639	0.83098
0.05	5300	0.78102	3392	0.49985	5769	0.85013	5969	0.87961

Base compositional Heterogeneity: The nucleotide composition among different lineages of Diptera varies across all mitochondrial coding region or individual coding positions (Fig. 3.5). Average GC content of 112 Diptera shows 25.22% whereas the 2nd codon position has higher in GC content 33.48% with lower standard deviation 1.48 and 3rd codon position shows lowest GC content 10.67% and higher standard deviation 5.14. The Brachycera, lower brachycera, Nematocera have about the same average GC percentage but after separating it into different families it provides fascinating results. The Tephritidae family or oriental fruit flies have highest overall GC content (29.30%) and on average 11.19% higher than that of Sciaroidea superfamily which have lowest GC content. Within Nematocera, Chironomoidea have the highest overall (27.30 %) and 3rd codon (15.61%) GC content, followed by Tipuloidea yet they have higher 1st (32.78%) and 2nd (33.77%) codon position GC content than Chironomoidea. Sciaroidea superfamily which possess lowest overall GC content as well as 3rd codon position

(4.89%) belongs to Nematocera. Within Brachycera, Tephritidae family have overall higher GC content (29.30%) and in 3rd codon position (18.01%) followed by Oestridae (28.14%, 17.04%). Tachinidae flies have lowest overall GC content (22.47%) as well as in 3rd codon position (6.22%). Tachinidae and Oestridae flies are conventionally not very different and belongs to same superfamily yet they persist significant variation in GC content specially at 3rd codon position. This appears to suggests mitochondrial PCGs (mPCGs) are heavily biased

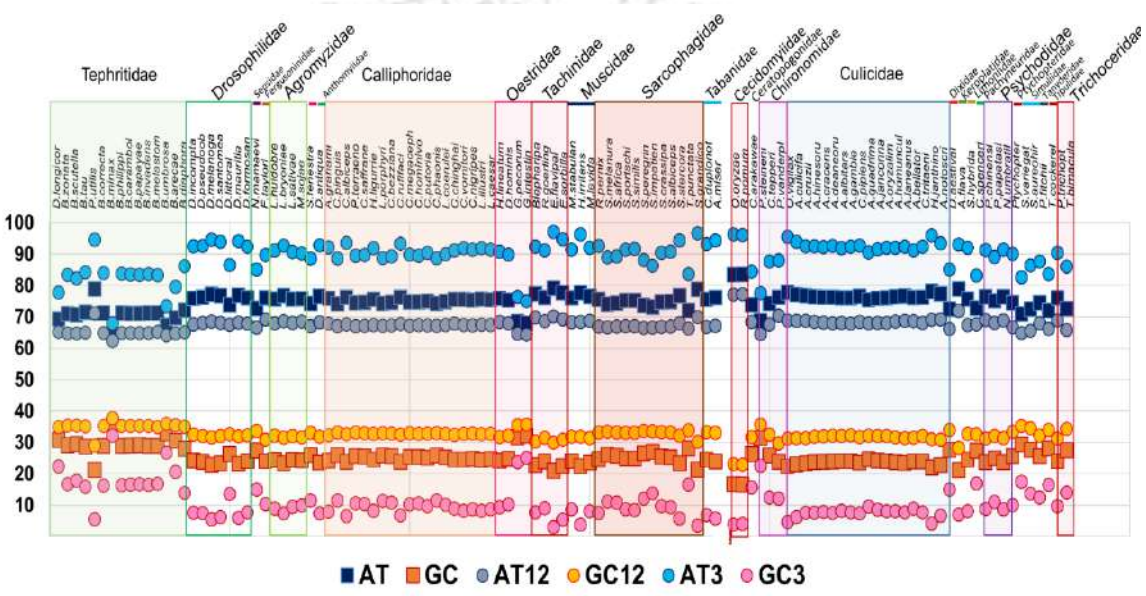


Figure 3.5: Nucleotide contents percentage of each species for the entire set of genes, and first, second, or third codon positions only.

Table 3.2: Variability of Each Base at each codon position Across the 112 species studied

Codon Position	%A Range	%C Range	%G Range	%T Range
First	12	10	12	14
Second	5	8	6	9
Third	14	19	12	24

towards AT content with potentially more preference towards T-ended codons (46.41%), followed by A-ended codons (42.91%) and the mean GC ending codons were accounting for only 10.67%. At 3rd codon position, by arranging the species by increasing percentage of T, a very large variation in composition from 33% to 54% is observed. Also, as the T% increases across the species, a significant corresponding decrease in the C% can be evident (r= -0.92801, p= 3.4432E-185). The G% at 3rd position is consistently low (11.8% to 1.3%) and A% is always high above 34%. It is also apparent that with the increasing of A% at 3rd codon position the G% decreases (r= -0.91339, p= 1.9417E-204) significantly. The robust changes in nucleotide composition involving C and T as well as A and G are broad and significant enough to allow us to look for similar changes within the dipteran mitogenome in other site classes. The relationship between C and A (r= -0.88313, p= 4.5353E-180) or T and G (r= -0.88102, p=1.7089E-207) also show prominent substitution. The composition of C and T at 1st and 2nd codon position does not vary as much as at 3rd codon position, but the correlation between them is more significant (1st codon position: r= -0.9506, p= 3.5428E-209; 2nd codon position: r= -0.93207, p= 4.6694E-248). This does not signify that at 3rd position the relationship between C and T is weaker than the rest, but it implies that increased variation arising from neutral substitutions in 3rd codon position introduces noise to the data[18]. Table 3.2 shows that the composition of the 2nd codon position is incredibly rigid, with very little deviation in composition across the species occurred. Therefore, vast heterogeneity in nucleotide content persists in the dipteran coding regions and this finding allows an extensive study of codon usage, which could have an insightful impact on synonymous codon bias and substitutional saturation.

Codon usage bias and substitutional saturation: Analysis of RSCU shows distinctive trends of codon usage across various linages, and it is observed that AT ending codons are most favoured by dipteran mitochondrial protein coding genes[161]. According to this analysis the most

prevalent codon in mPCGs is UCU (mean RSCU: 2.81) and CGA (mean RSCU: 2.76). Our comparative RSCU values for total sense codons indicate some variation in the RSCU values of 112 dipteran mPCGs, but the overall trend is close and a clear distinction between A / U (Red) and G / C (Green) end codons is obvious in the cluster analysis (Fig. 3.6). When the RSCU value increases for any codon the RSCU of other synonymous codons decreases indicating greater bias incidence. It is also apparent from the Cluster Figure 3.6 that rise in the GC3 content might lead to reduction of codon usage bias.

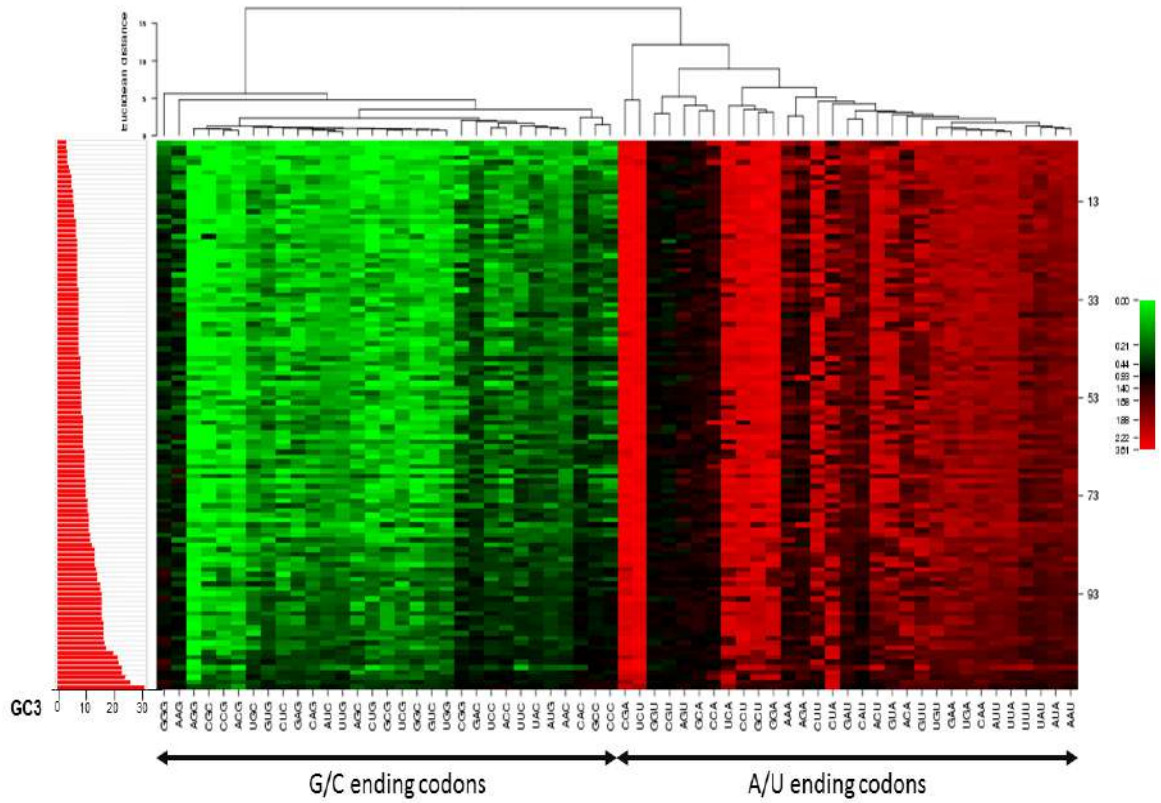


Figure 3.6: Heat map of RSCU values in Diptera mitogenome. The heat-map was drawn with CIMminer, using the quantile binning method. Bigger RSCU values, suggesting more frequent codon usage, are represented with brighter shades of red.

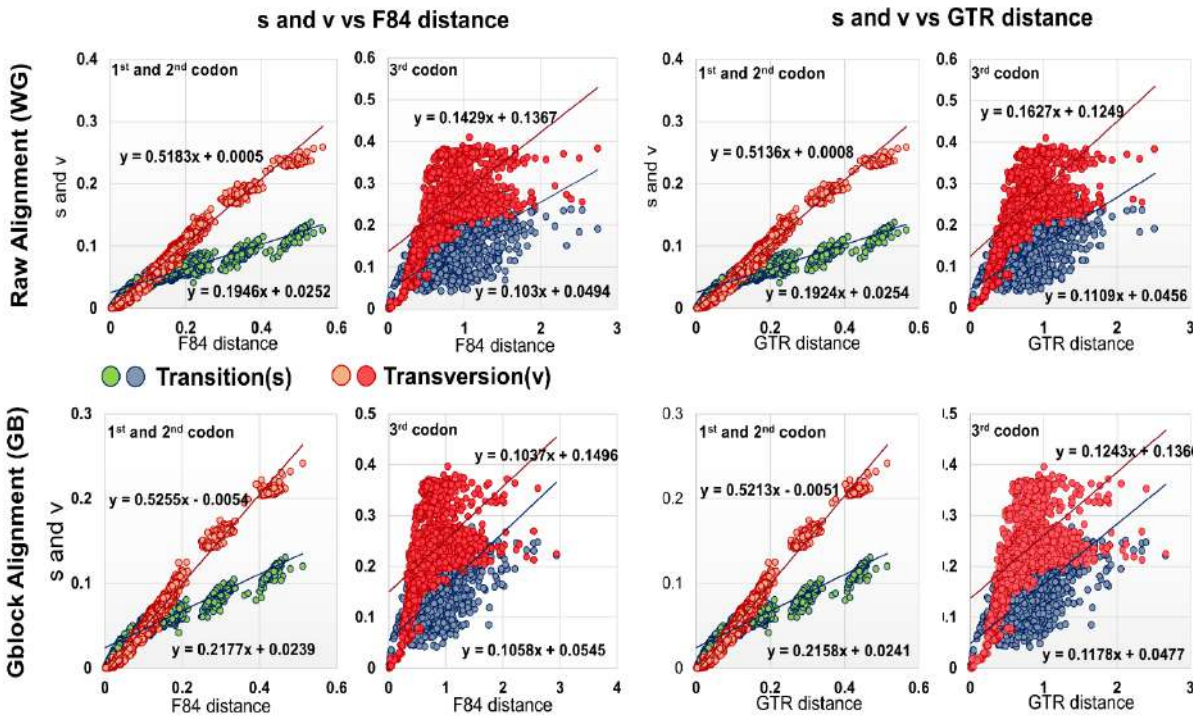


Figure 3.7: Substitution saturation plot of Transition and Transversion vs F84 and GTR distance of two datasets (WG and GB). Substitution saturation define by slope of the regression line, lower the slope higher the substitution saturation.

Table 3.3: Index of substitution saturation (Iss) measurement

		1-2-3 codon position		1-2 codon position		3 codon position	
	NumOTU	Iss	Iss.c	Iss	Iss.c	Iss	Iss.c
WG data	4	0.411	0.858	0.315	0.854	0.684	0.849
	8	0.449	0.845	0.342	0.846	0.748	0.843
	16	0.478	0.851	0.382	0.843	0.784	0.828
	32	0.514	0.818	0.407	0.815	0.821	0.809
GB data	4	0.296	0.856	0.166	0.852	0.584	0.848
	8	0.298	0.845	0.176	0.846	0.621	0.836
	16	0.295	0.846	0.179	0.836	0.614	0.822
	32	0.302	0.816	0.189	0.813	0.627	0.805

The degree of substitutional saturation was estimated at 3rd and combined 1st and 2nd codon positions. The transition (s) and transversion (v) vs divergence distance (F84 and GTR) show that transition is more substitutionally saturated than transversion. Nucleotide position wise analysis presents that third positions (F84: v slope= 0.1429, s slope= 0.103; GTR: v slope= 0.1627, s slope= 0.1109) are more substitutionally saturated the 1st and 2nd codon positions (F84: v slope= 0.5183, s slope= 0.1946; GTR: v slope= 0.5136, s slope= 0.1924) (Fig. 3.7).

Further, substitution saturation analysis using Xia's method estimated *Iss* and *Iss.c* through randomly selecting sample subsets of 4, 8, 16 and 32 OTUs (Operational taxonomic unit) multiple times and perform the test for each subset to diagnose if substitution saturation exists for these subsets of sequences. Here we use six different datasets combined 13 PCGs, combined 1st + 2nd codon positions and 3rd codon positions and another similar set removing the gaps. The outcome based on 32-taxon

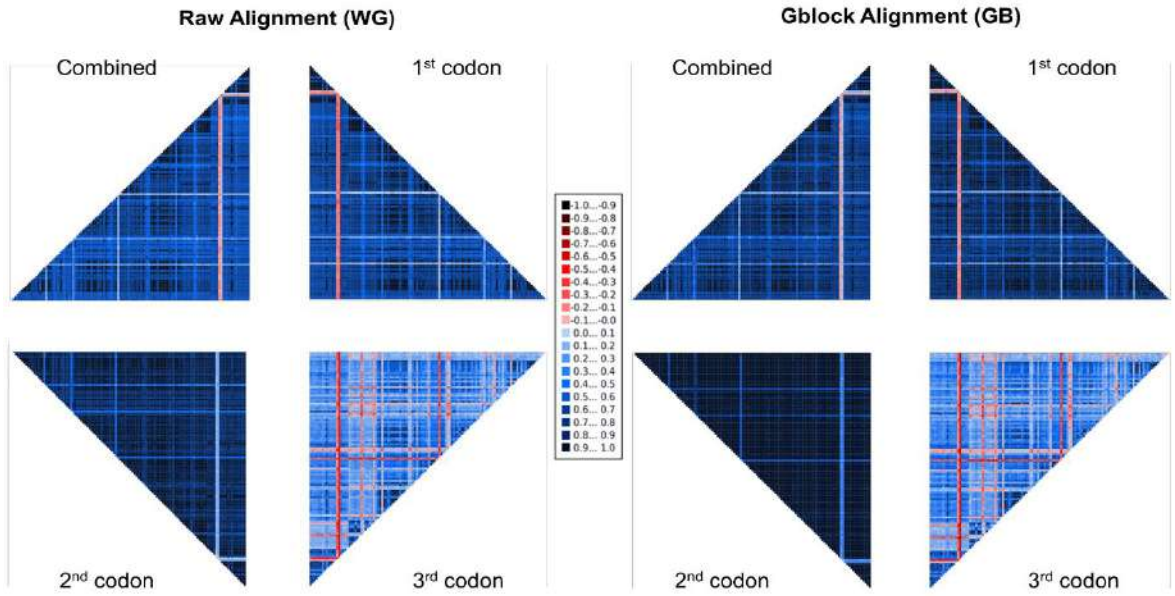


Figure 3.8: AliGROOVE analysis for the data sets. The mean similarity score between sequences is represented by a coloured square, based on AliGROOVE scores from -1, indicating great difference in rates from the remainder of the data set, that is, heterogeneity (red colouring), to +1, indicating that rates match all other comparisons (blue colouring).

simulations presents that the dataset of 3rd codon position with gap presents was much more saturation (*Iss*= 0.821, *Iss.cSym* = 0.809, *Iss.cAsym* = 0.557) than other dataset and combined 1st + 2nd codon positions without gap shows least saturation (*Iss*= 0.189, *Iss.cSym* = 0.813, *Iss.cAsym* = 0.569) (Table 3.3**,** the *Iss.cSym* and *Iss.cAsym* not presented here).

Heterogeneous sequence divergence: The AliGROOVE technique offers a measure of the heterogeneity of sequence divergence by comparing nucleotide divergences in pairs for each terminal or group of terminals identified by an internal node in multiple sequence alignment with all other sequences[128]. The analysis with different dataset of mitochondrial coding sequences shows substantial heterogeneity in sequence divergence for multiple subsets of taxa (Fig. 3.8). The heterogeneity was strongest for data sets Combined_wg, Combined_wg_3rd_codon, Combined_gb and Combined_gb_3rd_codon that include all nucleotide positions, and only 3rd codon position compared with the Combined_wg_1st_codon, Combined_wg_2nd_codon, Combined_gb_1st_codon and Combined_gb_2nd_codon data sets, indicating that third codon positions are greatly more rate-heterogeneous than the first and second codon positions and consistently scored low in the AliGROOVE pairwise comparisons. This analysis further indicates that the dataset from which the gap and missing data were omitted improved the effect of random sequence similarities with more positive similarity scores, resulting in fewer heterogeneity of the dataset. On each dataset, the outgroup taxa show lower similarity scores to the Diptera members. In the case of group members, few taxa from Tephritidae, Cecidomyiidae, Oestridae family and superfamily such as Pachyneuroidea, Chironomoidea obtained lower similarity scores than pairwise comparisons with other sequences.

3.3.5 Phylogenetic outcome from Heterogeneous models:

The new model is the general time-reversible three-state model (GTR3), which combines C and T into a single aggregate pyrimidine state, Y. This model is effective for explaining DNA with highly varied C and T compositions across genes and species[18]. This model was used to generate a phylogenetic tree using both raw concatenated (WG) and Gblock trimmed (GB) alignments of 13 mPCGs (Fig. 3.9). Similarly, to make base composition look steadier in some circumstances, we implemented the general time-reversible two-state model (GTR2), in which T and C are merged into one common state T and A and G into another common state A (Fig. 3.10). The long branch attraction issue may have been alleviated because of these analyses, but the grouping of Cecidomyiidae family species with out-group species remains in the phylogeny. Furthermore, when using the GTR2 and GTR3 with a discrete Gamma model, the LBA effect reappears. The *E. sorbilans*, which was previously classed with the Drosophilidae

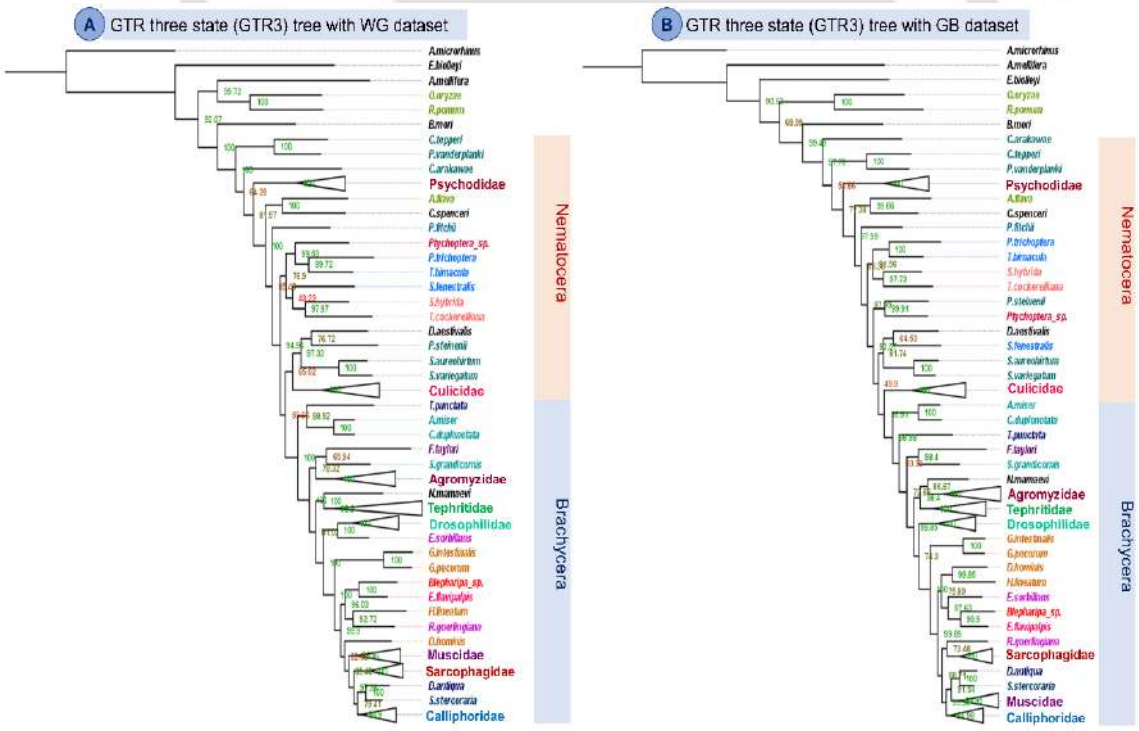


Figure 3.9: Diptera Phylogeny inferred by three state GTR model (GTR3) from A) raw dataset or without Gblock (WG) dataset and B) Gblock (GB) dataset using mitochondrial PCGs of 112 Diptera species and 4 outgroup species. The maximum bootstrap support (bs) is shown in green and minimum bs is shown in red colour. The collapsed nodes are denoting monophyletic family with more than two species.

family in homogeneous analyses, was now grouped within the Tachinidae family (GTR3/GTR2: 99.9/97.63) with strong support only in GB dataset by both two and three state models. But another Tachinid fly *Rutilia goerlingiana* is now placed as sister to Sarcophagidae family in the same superfamily. On the other hand, in Nematocera, the phylogenetic position of *P. steinenii* became more irregular after implementing the recoding method.

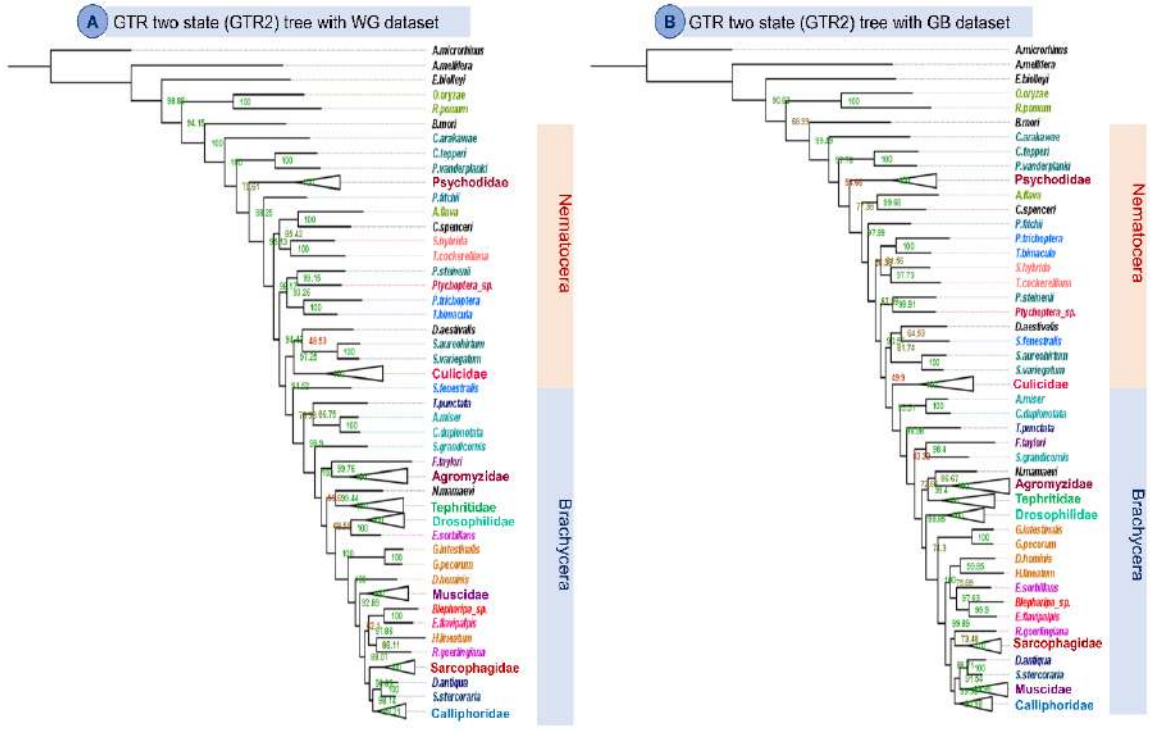


Figure 3.10: Diptera Phylogeny inferred by two state GTR model (GTR2) from A) raw dataset or without Gblock (WG) dataset and B) Gblock (GB) dataset using mitochondrial PCGs of 112 Diptera species and 4 outgroup species. The maximum bootstrap support (bs) is shown in green and minimum bs is shown in red colour. The collapsed nodes are denoting monophyletic family with more than two species.

Regular tree-building approaches using nucleotide sequences can become ineffective due to variation in the base composition of the sequences around the taxa[29,161,162]. LogDet transformation[163] commutatively eases this complexity by using the determinants of the matrices, and integrating them as real numbers[15]. The neighbor-joining analysis after LogDet transformation analysis from two datasets of 11835 bp (WG dataset) and 9018 bp (GB dataset) nucleotide depicts distinction in phylogenetic output. The Calliphoridae and the Sarcophagidae

family, like previous ML and BI analyzes, maintain their monophyletic relationship and Tachinidae and Oestridae were unable to. Again, *E. sorbilans*, a Tachindae fly clusters with Drosophilidae family which is like earlier ML and BI analysis. *P. utilis* of Tephritidae, *F. taylori* of Opomyzoidea superfamily deviates from their former location in phylogenetic tree. Members of Chironomoidea superfamily unable to claim monophyly and even Culicidea superfamily slips monophyly as Dixidae family diverged their relation and grouped with different family of Nematocera using both datasets. Similar to previous analysis flies from Sciaroidea superfamily (*R. pomum* and *O. oryzae*) once again came together with the outgroup taxa. Further applying 1000 bootstrap replicates with NJ tree heuristic search on both datasets using LogDet distance measurement provides two incongruent phylogenetic trees (Fig. 3.11).

However, this treatment release *E. sorbilans* and placed it within Oestroidea superfamily by the dataset that was eliminated from the gaps and ambiguity. The Tephritidae family failed to retain monophyly, as *Procecidochares utilis* slipped from its position after the LogDet treatment. Although, this approach yielded extensive polytomy in numerous nodes of the tree through both the datasets. Tipuloidea superfamily and Psychodidae family of Lower Diptera retains their monophyly. The uncertain positions of the members of Muscoidea flies in between the Oestroidea superfamily is always ascertain from all these analyses. Therefore, neither of this approach recover a topology indistinguishable to the conventional topology.

Since the rate of evolution is known to vary across sites[164,165] and across lineages[135,166], time-homogeneous models of sequence evolution have long been acknowledged as inappropriate. This phenomenon of sequence evolution is known as heterotachy, and it may be addressed using the General Heterogeneous Evolution On a Single Topology (GHOST) model developed by Crotty, Stephen M., et al. and published in 2020[136]. GHOST is an edge-unlinked mixture model with many site classes on the same tree topology, each with their own set of model

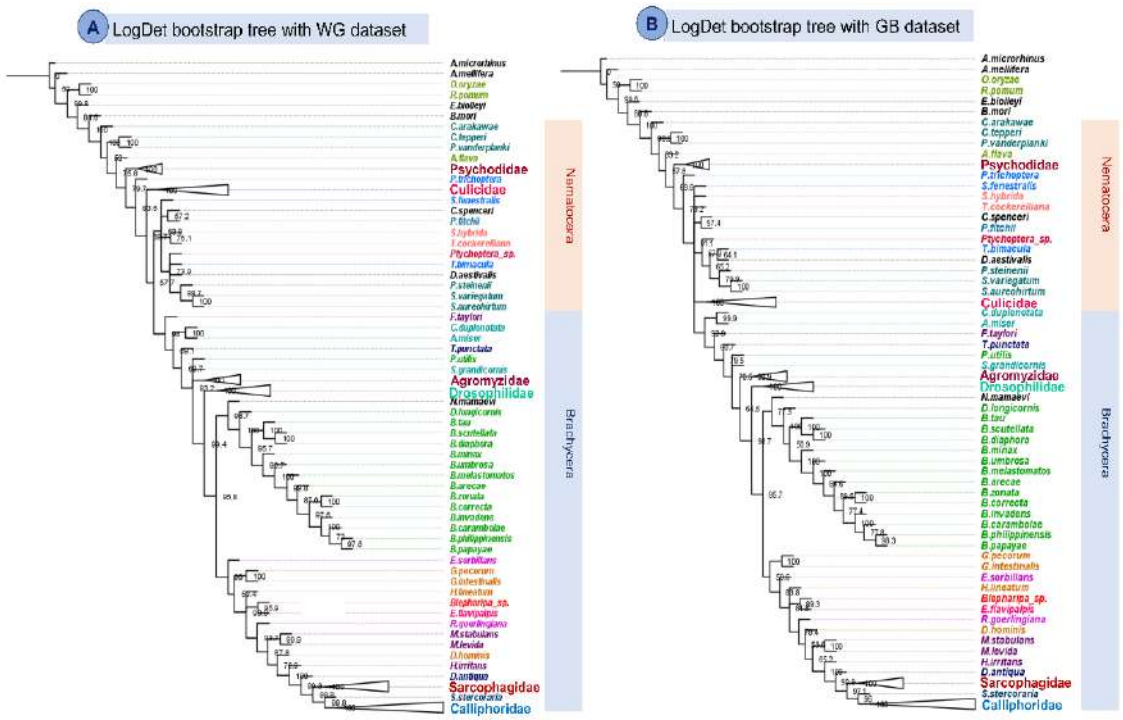


Figure 3.11: Diptera Phylogeny inferred by 1000 bootstrap analysis of LogDet transformation of A) raw dataset or without Gblock (WG) dataset and B) Gblock (GB) dataset using mitochondrial PCGs of 112 Diptera species and 4 outgroup species. The bootstrap support (bs) is shown in each node. The collapsed nodes are denoting monophyletic family with more than two species.

parameters and edge lengths. Thus, GHOST is a right fit for heterotachous evolution. Unlike an edge-unlinked partition model, the GHOST model does not need a priori data partitioning, which might lead to model misspecification. Herein, GHOST model was implemented for both WG and GB datasets and the outcomes from these analyses suggest that the tachinid fly *E. sorbilans* retains its position with Drosophilidae (Fig. 3.12). *P. steinenii*, a Chironomidae fly, remains in the Simuliidae family, while *Dixella aestivalis*, a Culicoidea superfamily fly, has also joined the Simuliidae family after GHOST treatment. After implanting this model Oestridae family recovered as monophyletic clade by GB dataset. Although the LBA problem in homogenous phylogenetic trees appears to reduce after using the GHOST model, the proximity of the cecidomyiidae family to the out-group remains despite these analyses with both datasets.

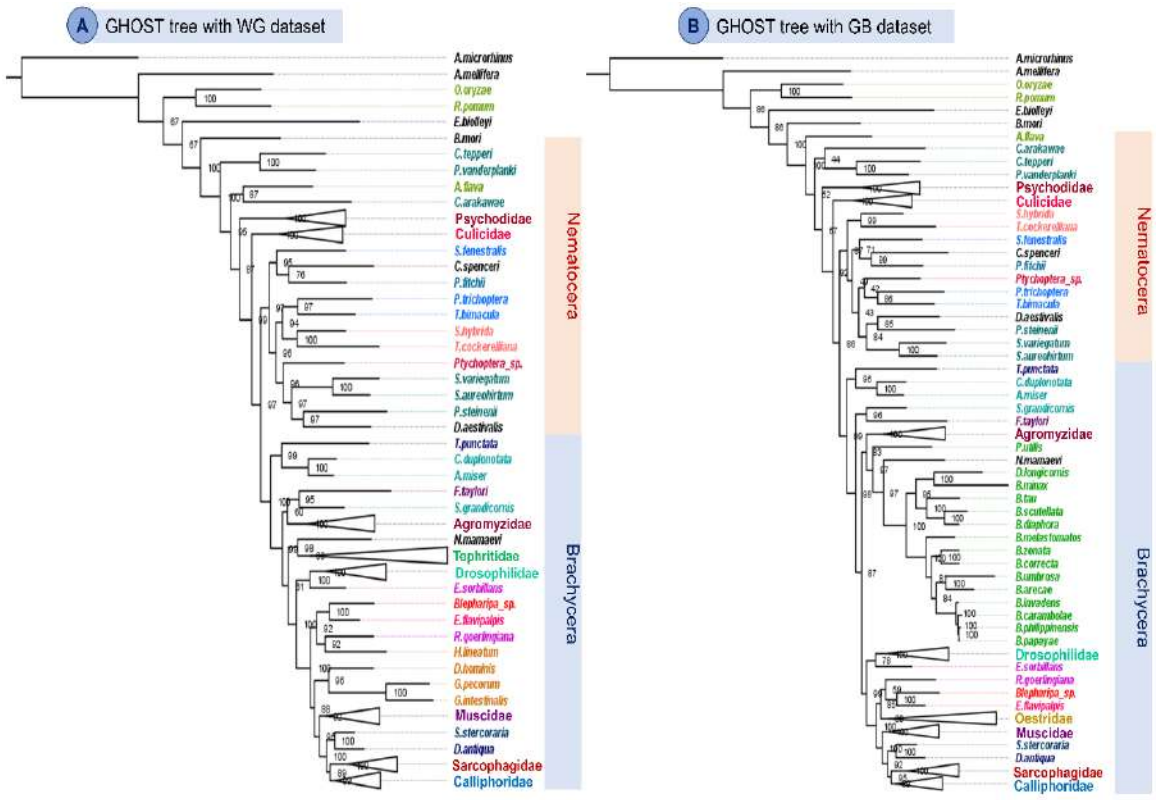


Figure 3.12: Diptera Phylogeny inferred by General Heterogeneous Evolution On a Single Topology (GHOST) model from A) raw dataset or without Gblock (WG) dataset and B) Gblock (GB) dataset using mitochondrial PCGs of 112 Diptera species and 4 outgroup species. The bootstrap support (bs) is shown in each node. The collapsed nodes are denoting monophyletic family with more than two species.

In PhyloBayes, the CAT-GTR model is a Dirichlet process mixture of equilibrium frequency profiles and general exchange rates. The CAT model fully accounts for positional heterogeneity in the substitution process, allowing for a more accurate detection of multiple substitutions at a single site[167]. The CAT-GTR model was used to analyze two datasets: the WG dataset and the GB dataset after the removal of the 3rd codon position. For both datasets, the results demonstrated that the analysis with this model was unable to resolve polytomy in multiple nodes and unable to separate Brachycera and Nematocera suborder appropriately (Fig. 3.13). As CAT-GTR in PhyloBayes is very computationally costly to conduct, this appears to suggest that increasing the number of cycles to converge the Markov chains or using an alternative ML

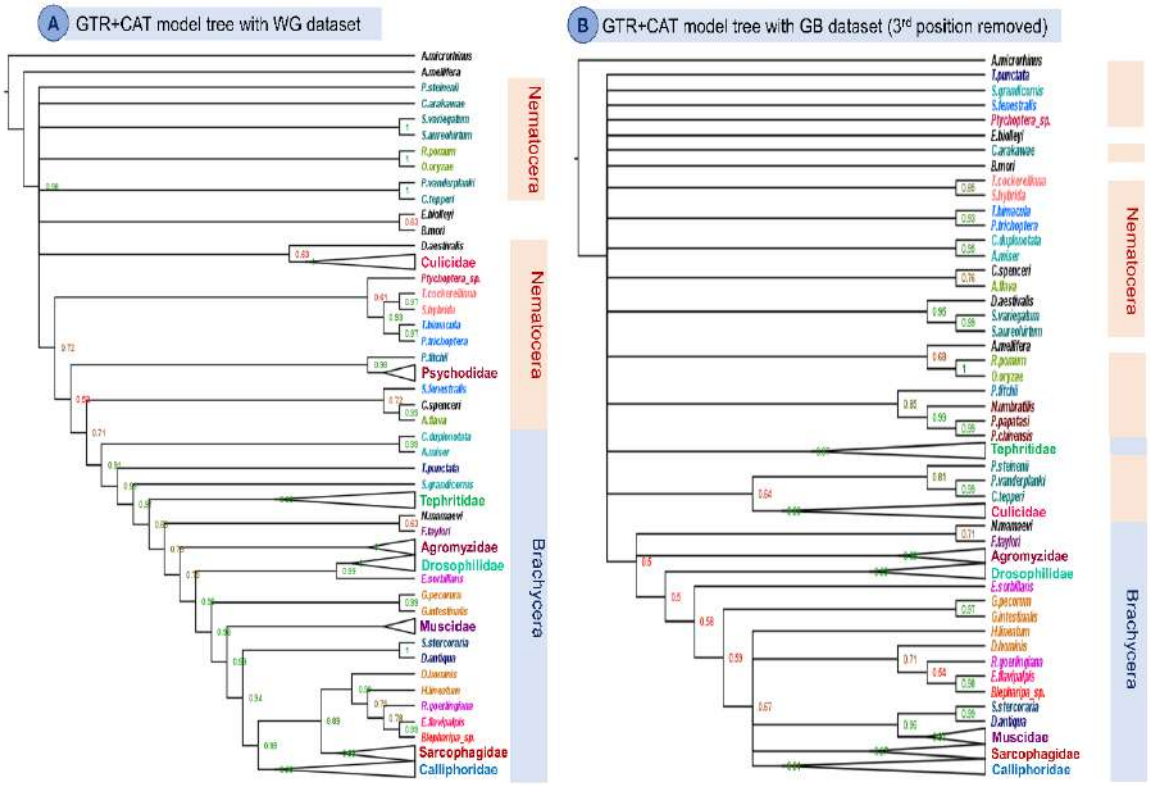


Figure 3.13: Diptera Phylogeny inferred by CAT+GTR model from A) raw dataset or without Gblock (WG) dataset and B) Gblock (GB) dataset without 3rd codons, using mitochondrial PCGs of 112 Diptera species and 4 outgroup species. The posterior probability is shown in each node. The collapsed nodes are denoting monophyletic family with more than two species.

method may be essential to infer acceptable phylogenetic tree resolution[72,168]. However, the second dataset's phylogenetic tree was able to distinguish *E. sorbilans* from Acalyptratae and place it in a paraphyletic relationship with all Calyptratae. Furthermore, the phylogenetic tree from second dataset suggests that the Chironomid fly, *P. steinenii*, has been grouped with two other members of the same family.

3.3.6 Internode Certainty Analyses:

Our analysis of concordance reveals that a significant number of bipartitions on individual gene trees are not well supported. For this analysis gene trees are build using dnaML software, species tree created by ASTRAL and ICA was calculated by PhyParts software. Although the previously constructed RaxML tree is not completely identical to the ASTRAL tree.

This analysis not supporting the clustering of *O. oryzae* and *R. pomum* with the out-group *A. melifera* as the ICA score is negative (-0.035) but *B. mori* placed as sister clade of other Diptera with ICA value 0.012, 6 supporting gene trees out of 13. The backbone of the Diptera was characterized by high level of gene tree discordance, and many nodes supported alternate topologies and the ICA value varies between 0.002 and -0.006 (Fig. 3.14). The major diversification node of Brachycera and Nematocera shows no concordant genes to supports the main topology and the ICA value -0.002 suggesting that most gene trees supporting alternate topologies. Similarly, diversification from lower Brachycera to higher Brachycera does not support by any gene trees allowing alternate topologies. While the ICA scores in many of the nested clades of some major families, such as Calliphoridae, Tephritidae, Agromyzidae, Culicidae, and Psychodidae are comparatively higher. In Calliphoridae clade out of 18 nodes only one node has negative ICA, two nodes show complete concordant gene trees (ICA=1) and 10 nodes show more than 50% concordant gene trees. Out of 14 nodes in the Tephritidae clade, 7 had more than 50% concordant gene trees and ICA values ranging from 0.609 to 0.042. The three nodes in the Agromyzidae clade have ICA scores of 0.201,0.354, and 0.609, respectively, and more than 70% of the gene trees are concordant with the species tree. Whereas, families like Sarcophagidae, Oestridae, and some lower Diptera show negative ICA value and thus support for alternate topologies. In Sarcophagidae 6 nodes out of 8 nodes show very low and negative ICA score. This analysis not supported the Muscidae fly *H. irritans* placed as sister taxa of Sarcophagidae. While other two Muscidae flies show strong concordant by the gene trees. This analysis could not distinguish monophyly of Muscidae, Tachinidae and Oestridae family and no support from gene trees and negative ICA score suggest provision for alternate topologies. The placement of *E. sorbilans* as paraphyletic position is weakly supported by the gene trees, only 5 out of 13 gene trees supporting with ICA score .028. The position of Sepsidae fly *N. mamaevi* as sister taxa of Tephritidae is not supported by gene trees and negative ICA

value (-.014) suggest alternate topology. The position of *S. grandicornis* is not supported by any gene trees and ICA score -0.006 suggest provision for alternate topology, also placement

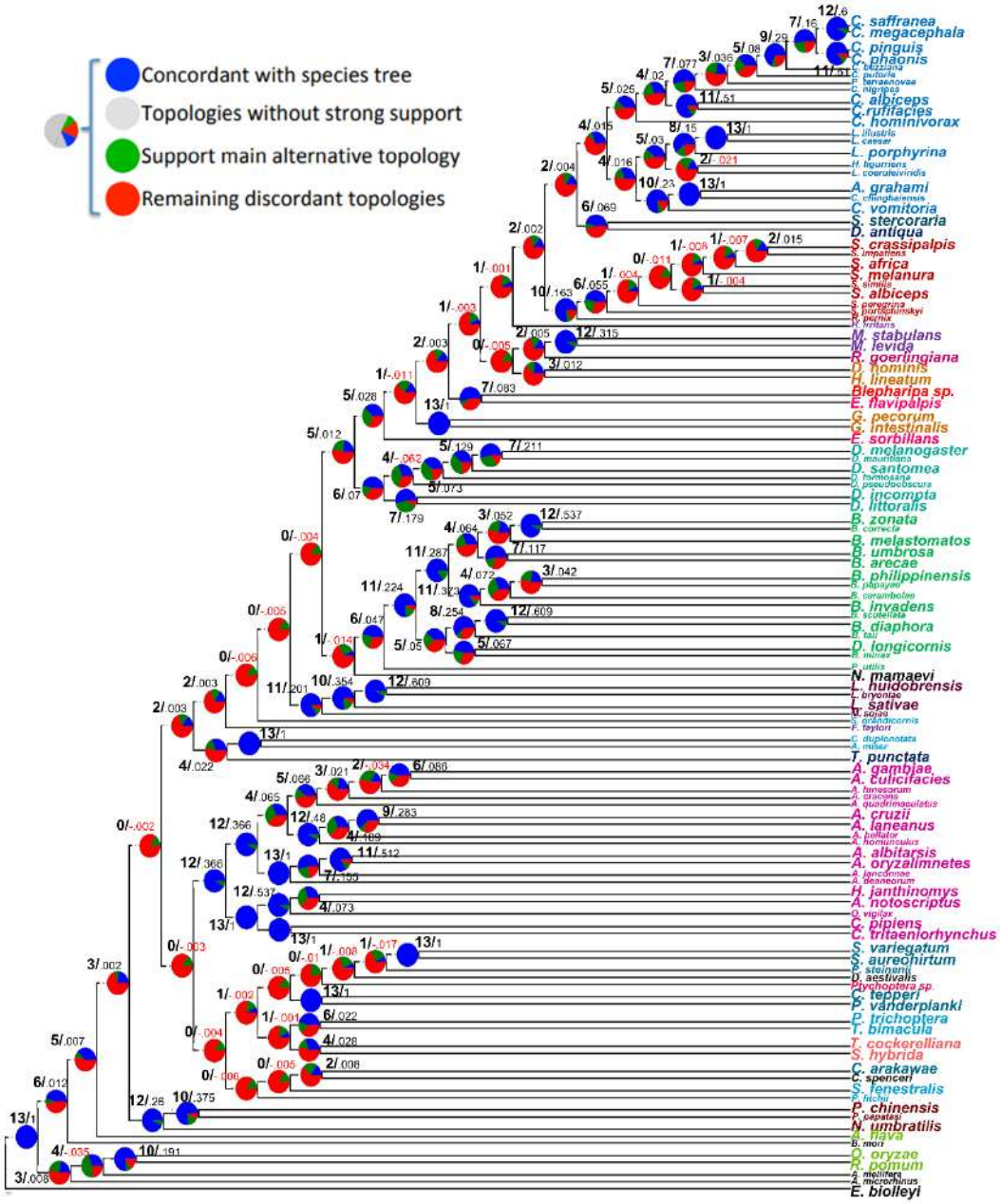


Figure 3.14: Species trees of the COMPLETE dataset inferred with ASTRAL (GB dataset). Maximum likelihood. Pie charts next to the nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support)

of *F. tylori* outside of schizophora is very weakly supported by only 2 gene trees. In Nematocera, Culicidae family have highly concordant nodes with one exception *A. hinesorum* (ICA=-0.034), ten nodes have more than 50% concordant gene trees including three nodes with full concordant (ICA=1). The two nodes of Psychodidae family show strong support by more than 70% gene trees concordant with species tree. The position of Chironomidae fly *P. steinenii* with Simuliidae family is not supported by gene trees, the negative ICA score (-0.017) suggests alternate topology, similarly *D. aestivalis* and *Ptychoptera sp.* shows negative ICA score supporting alternate topology. The connecting node of Trichoceridae and Tipuloidea clade show negative ICA (-0.001), also paraphyletic *P. fitchii , S. fenestralis* with *C. arakawae,* and *C. spenceri* show negative ICA suggesting alternate topology instead of main reference toplogy. Therefore, these analyses imply that low phylogenetic information in the sampled loci. Analysis with both the dataset show discordant on backbone of Diptera species tree persist with the other mitochondrial gene trees (Fig. 3.14).

To quantify phylogenetic uncertainty and discriminate between branches with little information and those with well-supported but mutually exclusive evolutionary histories, the Quartet Sampling method was performed (Fig. 3.15) [145]. The coalescent based ASTRAL tree was used here to assess QC (Quartet Concordance)/QD (Quartet Differential) /QI (Quartet Informativeness) score of each node in that tree. The outcome of this analysis displays those twenty-five nodes have QC = 1, suggesting that all quartet trees are concordant with the focal branch. Negative QC score observed in twenty-eight nodes implying counter-support for an alternative branch like in the ICA scores [143]. The QI score more than 0.5 found in 107 nodes suggesting the quartets are well informed, whereas three nodes have QI score less than 0.1 indicating uncertainty of the quartets. In this analysis we found that QD = 0 for twenty-seven nodes indicating complete skew towards one of two discordant alternative relationships[145]. In this analysis some major family clades showed strong support whereas, we found low support

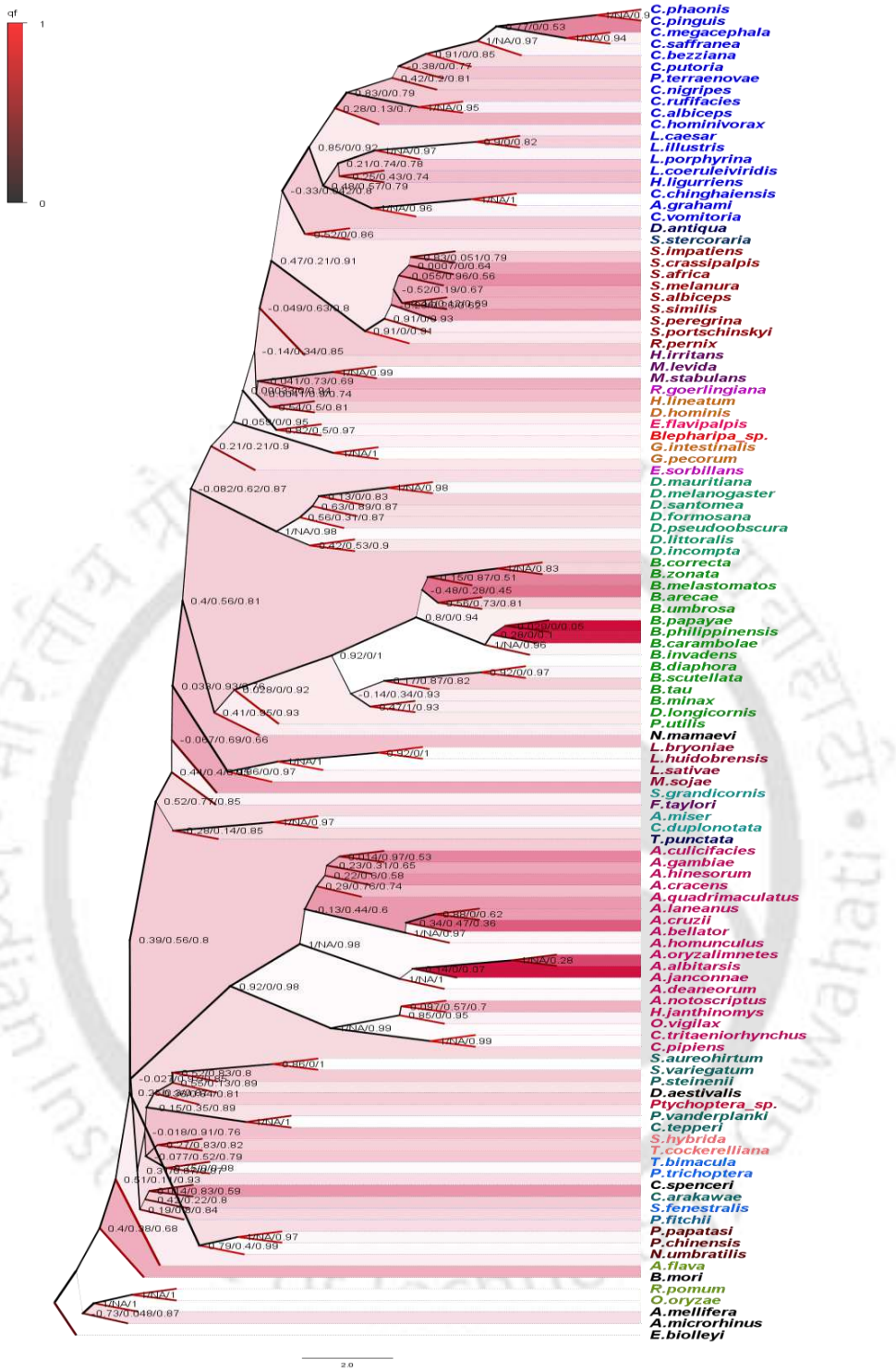


Figure 3.15: Quartet Sampling phylogeny with ASTRAL tree, QC (Quartet Concordance)/ QD (Quartet Differential)/ QI (Quartet Informativeness) scores (100 replicates of full GB alignment) for Diptera.

along the "backbone" relating these clades. Full support (1/-/1) by QS method provided to Cecidomyiidae (*R. pomum* and *O. oryzae*) flies when clustering with outgroup *A. melifera* and

relatively satisfactory support to *B. mori* (0.4/0.38/0.68) and skewed discordance with its position.

The node where main Diptera lineages started diversifying, the parphylatic *A. flava* shows strong support with skewed frequency (0.51/0.11/0.93) indicating presence of secondary evolutionary history. The QS score (0.79/0.4/0.99) of Psychodidae clade indicate strong quartet support with satisfactory QD value. The node that separates the Culicidae family from the other Nematocera showed counter support (-0.027/0.97/0.8) since the QC score is negative but the frequency of alternate potential topologies is similar (QD ≈ 1). All lower Diptera nodes have a QS score (0.31/0.87/0.87) reflecting satisfactory quartet support and a modest skew in discordant frequencies indicating no alternate history preferred. The diverging node of two Chironomidae flies *P. vanderplanki* and *C. tepperi* showed QS score (-0.15/0.35/0.89) suggesting strong counter support and sampled quartets supported alternative topologies.While another Chironomidae fly, *P. steinenii* clustered with Simuliidae family displayed QS score (0.52/0.83/0.8) indicating strong quartet support. The paraphylic relationship of *D. aestivalis* with the same clade has a QS score of (0.55/0.13/0.89), suggesting high quartet support with discordant skew and the presence of a supported secondary evolutionary history. The connecting node of Tipuloidea and Trichoceridea showed QS score (-0.077/0.52/0.79) inferring counter support. Whereas another fly of Trichoceridea superfamily, *S. fenestralis* paraphyletic relation with *C. spenceri* and *C. arakawae* cluster displyed QS score (-0.42/0.22/0.8) indicating a strong majority of quartets support one of the alternative discordant quartet arrangement history. *P. fitchii* in paraphyletic relation with same clade showed QS score (0.19/0.8/0.84) suggesting weak majority of quartets support the focal branch. The QS score of Brachycera diverging node (0.52/0.77/0.85) showed strong support and low skew suggesting lower possibility of alternate history. The node postion of *F. tylori* QS score (0.44/0.4/0.95) relatively satisfactory quartet support and skewed discordance, whereas node of *S.*

grandicornis displayed counter support (-0.067/0.69/0.66) indicates presence of secondary evolutionary history. The node of Agromyzidae origination showed QS score (0.033/0.93/0.76) very weak quartet support and almost similar possibility of two discordant topologies. The Tephritidae origination node have QS score (0.4/0.56/0.81) indicates acceptable quartet support whereas monophyletic Drosophilidae QS score (-0.082/0.62/0.87) suggest counter support with low skew. The tachind fly *E. sorbilans* paraphyletic position showed QS score (0.21/0.21/0.9) implied weak quartet support with a skewed frequency for an alternative placement. The separate cluster of other two tachind fly *Blepharipa sp.* and *E. flavipalpis* displayed QS score (0.00083/0/0.94) suggest very waek quartet support and complete skew towards one discordant toplogy signify almost perfect conflict. Another tachind fly *R. goerlingiana* clusterd with Muscidae family and QS score (-0.041/0.73/0.69) suggest counter support with low skew. Another Muscidae fly *H. irritans* in counter supported in separate node QS score (-0.049/0.63/0.8). The QS score (0.059/0/0.95) of clustered Oestridae flies *G. pecorum* and *G. intestinalis* indicate very waek quartet support and complete skew towards one discordant toplogy signify almost perfect conflict. Other two Oestridae flies *H. lineatum* and *D. hominis* clustered separately and QS score (-0.0041/0.9/0.74) of their diverging node suggest counter support with low skew. Therefore the species from Muscidae, Oestridae, and Tachinidae families showed high incongruence in the position. The diverging node of Sarcophagidae showed QS score (0.47/0.21/0.91) suggest that acceptable quartet support with skewed frequency. The QS score (-0.33/0.042/0.8) of *S. stercoraria* and *D. antiqua* diverging node showed a strong counter support with skewed frequency signifying a strong majority of quartets support one of the alternative discordant quartet arrangement history. The diverging node of monophyletic Calliphoridae family displayed QS score (0.85/0/0.92) indicating strong quartet support and complete discordant skew.

3.3.7 Neighbour-net network an alternative relation:

To explore the relationships in greater detail, we used a neighbor-net network analysis on the multiple alignment of 116 Diptera mitochondrial sequences. The neighbor-net method[138], which is based on the neighbor-joining algorithm[169], builds circular splits and employs a circular network algorithm[170] to generate planar networks. A split is a division of a set of data (sequences) into two groups; the set can be partitioned by many splits, and then a network can be made from these splits. Each split will define an edge that connects the two divisions, resulting in a splits graph. Splits can be compatible or incompatible with one another. Compatible splits relate to phylogenetic tree branches, hence the splits graph for a compatible collection of splits is a tree. Whereas, an incompatible split separates nodes that are not linked by a branch. In order to build a network, incompatible sets of splits must be permitted. "Weakly compatible" splits are used by Neighbor-net[138]. When splits are incompatible (form contradictory groupings), a box (cycle) is included to denote those alternate splits exist. So,

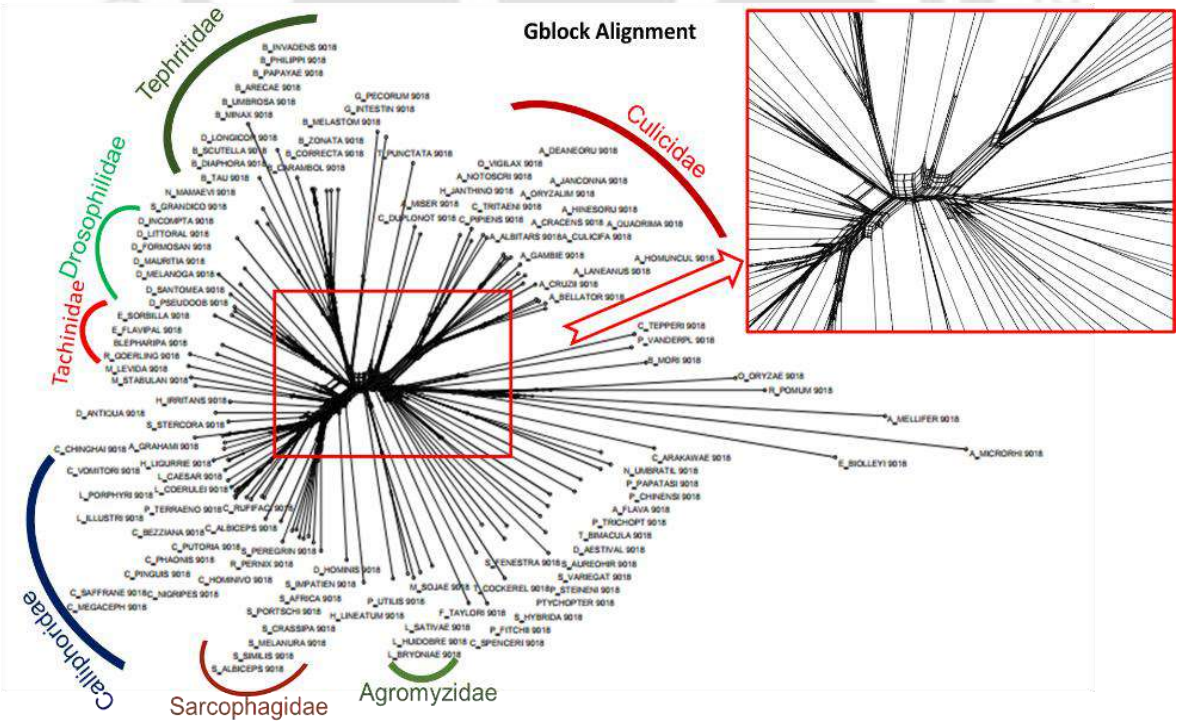


Figure 3.16: Neighbour-net analysis on GBlock alignment with uncorrected-p distance showing network relationship among different clades. The parallelogram indicates alternative relation of different taxa.

boxes in the splits graph could be utilized to find reticulations. A pair of nodes in a splits graph may be connected by a single edge (tree-like section) or a set of parallel edges indicating alternate evolutionary possibilities (reticulate section). The SplitsTree network visualization clearly shows that regions around the Drosophilidae–Tachinidae groups contain boxes in our analysis (Fig. 3.16). Another region for observation is the section of the network around the connecting point of the long branch leading to the Cecidomyiidae (*O. oryzae*, *R. pomum*) family. These boxes reflect uncertainty in the location of the Cecidomyiidae lineages ambiguity already noted in our phylogenetic analysis. Thus, Neighbor-Net is prone to long branch attraction, as evidenced by earlier phylogenetic analyses[138]. Networks, unlike trees, may reflect both the signal supplied by long branch attraction and the signal of the underlying phylogeny. The Neighbor-Net did more than just reflect ambiguity and complexity. We now have an overview of the data's structure, which is not limited to a single bifurcating tree. As a result, the Neighbor-Net serves as an indicator of alternate inferences, and these conclusions are achieved without relying on a single tree.

3.3.8 Sign of reticulate evolution in Diptera:

The presence of potential reticulation on the Diptera inferred through three different methods. Reduced tree of 10 taxa was used for the analysis due to its computation cost. The outcome of maximum parsimony and maximum pseudo-likelihood visualized in dendroscope while maximum likelihood inference of reticulation was more complex and visualized in icytree (Fig. 3.17). The study yielded five most desired networks of reticulation, with a maximum of three reticulation events permitted. The parsimony analysis was scored ‘minimizing deep coalescence,’ or MDC, criterion and likelihood analysis was scored using log-likelihood. The taxa like *P. vanderplanki, S. grandicornis, E. sorbilans, P. steinenii* shows maximum number of reticulations in different analysis. However, the analysis needs to validate with different statistical method.

The species tree with the smallest possible number of extra lineages over all reconciliations of all gene trees in the input is found by minimizing deep coalescences. This criterion only takes into account gene tree topologies[171]. The inability to estimate parameter values outside the network's topology is one restriction of inference based on the MDC criteria. Another restriction is that such inference is not statistically consistent for species, which suggests difficulties in phylogenetic network inference based on the criterion[172]. The ML estimates of phylogenetic networks based on multispecies network coalescence can alleviate the flaws of MDC[96]. It should be noted that estimating the likelihood of a phylogenetic network is a high computation burden in all of the statistical inference techniques provided by PhyloNet. To address this issue, the InferNetwork MPL command in PhyloNet allows for the inference of

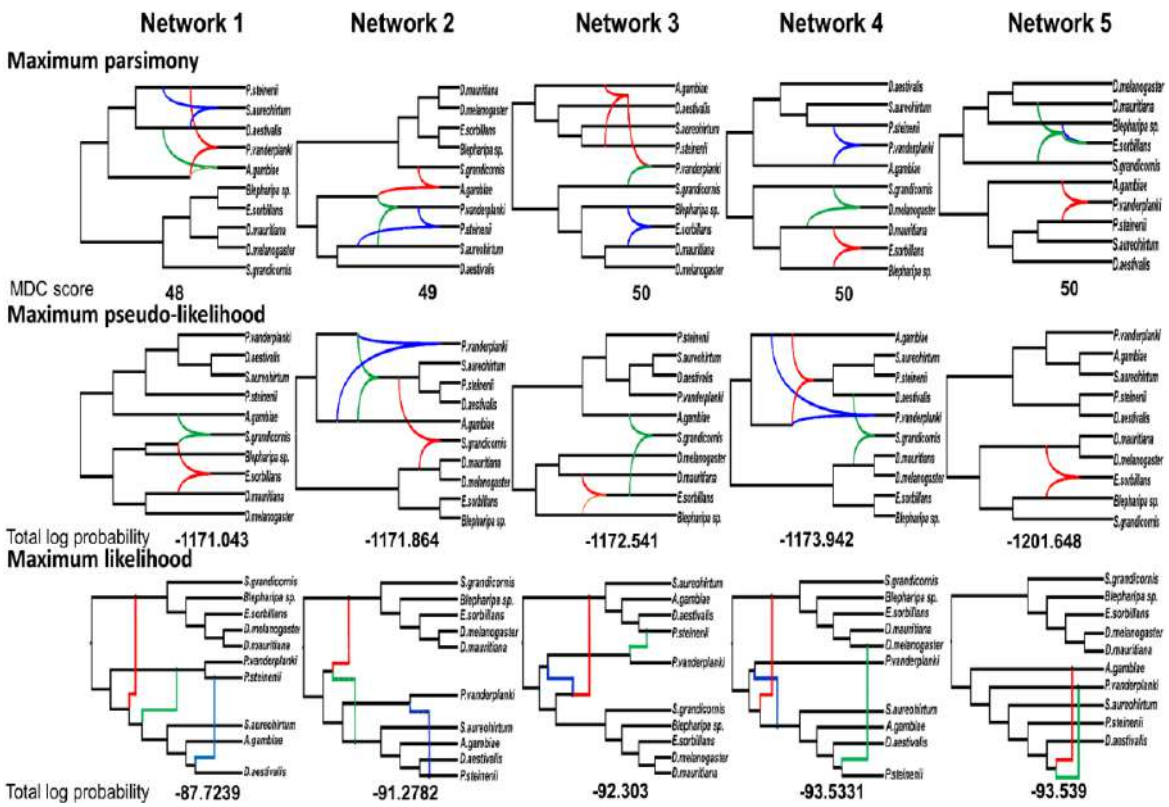


Figure 3.17: PhyloNet network showing five most likely reticulation scenarios in three different methods (Maximum parsimony, Maximum pseudo-likelihood, and Maximum likelihood).

phylogenetic networks based on a "pseudolikelihood" metric[148]. However, using this approach, the input could only be gene tree topologies (branch lengths are not allowed).

Under all the aforementioned criteria, inferring phylogenetic networks is computationally very difficult. Consequently, the PhyloNet inference techniques traverse the space of phylogenetic networks while assessing the applied criterion (e.g., MDC, likelihood, or pseudolikelihood) on the proposed networks. The space of phylogenetic networks is substantially larger than the space of trees based on the same number of taxa. Thus, there is not much that can be done with present cutting-edge phylogenetic network search algorithms other than allowing the search or sampling enough time to converge on a reasonable estimate. It is worth noting that the PhyloNet search heuristics do not fix a tree before looking for reticulations to add to it. Instead, they define moves on phylogenetic networks and navigate across network space without specifying a species tree or backbone tree[171].

3.3.9 Four-Taxon Analyses:

To test for hybridization events one at a time, we further reduced the 10-taxon(net) data set to 20 four- taxon combinations that each included one outgroup (*B. mori*). This analysis displayed 14 out of the 20 bifurcating quartet species trees (H0 and more frequent gene tree) were compatible with the ASTRAL species tree inferred from the complete 116-taxon GB data set. The remaining 4 quartets corresponded to the third most frequent gene tree topology in the 116-taxon ASTRAL species tree and 2 quartets corresponded to the second frequent gene tree topology. The AU test highly supported 12 of the 14 log-likelihood supported H0 quartets; four H2 quartets with three quartets strongly supported by the AU test; and two H1 quartets with one quartet strongly supported by the AU test. This AU test is results also in agreement with other Tree topology tests (Table 3.4). Notably, the tachinid fly *E. sorbilans* grouped with *D. melanogaster* rather than other *Blepharipa sp.* and *E. flavipalpis* in 1X1 and 8X1 quartets that

Table 3.4: Result of quartet-based tree topology test.

code	Hypothesis	quartet trees	rank	item	obs	au	np	bp	pp	kh	sh	wkh	wsh
1X1	H1	(((E_sorbilla,D_melanoga),Blepharipa),B_mori);	1	2	-24.5	**0.989**	0.99	0.991	1	0.989	0.998	0.989	0.993
	H0	(((Blepharipa,E_sorbilla),D_melanoga),B_mori);	2	1	24.5	0.011	0.01	0.009	2.00E-11	0.011	0.011	0.011	0.018
	H2	(((Blepharipa,D_melanoga),E_sorbilla),B_mori);	3	3	31.9	6.00E-77	1.00E-20	0	1.00E-14	2.00E-04	2.00E-04	2.00E-04	0.001
1X2	H2	((Blepharipa,(P_steineni,D_melanoga)),B_mori);	1	3	-8.7	0.865	0.861	0.86	1	0.86	0.933	0.86	0.911
	H0	((P_steineni,(Blepharipa,D_melanoga)),B_mori);	2	1	8.7	0.146	0.138	0.139	2.00E-04	0.14	0.148	0.14	0.221
	H1	((D_melanoga,(Blepharipa,P_steineni)),B_mori);	3	2	14.8	0.006	0.001	0.001	4.00E-07	0.013	0.014	0.013	0.025
2X1	H0	((P_vanderpl,(Blepharipa,E_sorbilla)),B_mori);	1	1	-40.4	1	1	1	1	1	1	1	1
	H2	((E_sorbilla,(Blepharipa,P_vanderpl)),B_mori);	2	3	40.4	1.00E-04	4.00E-05	2.00E-06	3.00E-18	0	0	0	0
	H1	((Blepharipa,(P_vanderpl,E_sorbilla)),B_mori);	3	2	41.3	2.00E-05	7.00E-06	0	1.00E-18	0	0	0	0
2X2	H0	(((P_vanderpl,P_steineni),Blepharipa),B_mori);	1	1	-19.6	0.987	0.988	0.989	1	0.988	0.997	0.988	0.991
	H2	(((Blepharipa,P_steineni),P_vanderpl),B_mori);	2	3	19.6	0.014	0.012	0.011	3.00E-09	0.012	0.012	0.012	0.021
	H1	(((P_vanderpl,Blepharipa),P_steineni),B_mori);	3	2	23.3	0.001	1.00E-04	0	7.00E-11	0.001	0.001	0.001	0.001
3X1	H0	(((S_variegat,P_vanderpl),E_sorbilla),B_mori);	1	1	-3.1	0.608	0.621	0.619	0.957	0.621	0.737	0.621	0.707
	H1	(((S_variegat,E_sorbilla),P_vanderpl),B_mori);	2	2	3.1	0.392	0.379	0.381	0.043	0.379	0.441	0.379	0.481
	H2	(((E_sorbilla,P_vanderpl),S_variegat),B_mori);	3	3	18.6	0.002	3.00E-06	0	8.00E-09	0.002	0.003	0.002	0.004
3X2	H2	((P_vanderpl,(S_variegat,P_steineni)),B_mori);	1	3	-54.7	1	1	1	1	1	1	1	1
	H0	((S_variegat,(P_vanderpl,P_steineni)),B_mori);	2	1	54.7	4.00E-06	3.00E-06	0	2.00E-24	0	0	0	0
	H1	((P_steineni,(P_vanderpl,S_variegat)),B_mori);	3	2	56.4	7.00E-05	1.00E-05	0	3.00E-25	0	0	0	0
4X1	H0	(((Blepharipa,E_flavipal),E_sorbilla),B_mori);	1	1	-8.9	0.9	0.897	0.9	1	0.897	0.95	0.897	0.933
	H2	(((E_sorbilla,E_flavipal),Blepharipa),B_mori);	2	3	8.9	0.109	0.102	0.099	1.00E-04	0.103	0.104	0.103	0.166
	H1	(((Blepharipa,E_sorbilla),E_flavipal),B_mori);	3	2	12.9	0.011	0.002	5.00E-04	3.00E-06	0.015	0.015	0.015	0.031
4X2	H0	((P_steineni,(Blepharipa,E_flavipal)),B_mori);	1	1	-39.9	1	1	1	1	1	1	1	1
	H1	((E_flavipal,(Blepharipa,P_steineni)),B_mori);	2	2	39.9	7.00E-05	1.00E-04	2.00E-05	5.00E-18	0	0	0	2.00E-04
	H2	((Blepharipa,(P_steineni,E_flavipal)),B_mori);	3	3	42.1	2.00E-19	3.00E-10	0	5.00E-19	0	0	0	0
5X1	H0	((E_sorbilla,(D_melanoga,D_mauritia)),B_mori);	1	1	-50.5	1	1	1	1	1	1	1	1
	H1	((D_mauritia,(D_melanoga,E_sorbilla)),B_mori);	2	2	50.5	1.00E-04	4.00E-05	0	1.00E-22	0	0	0	0
	H2	((D_melanoga,(E_sorbilla,D_mauritia)),B_mori);	3	3	50.5	1.00E-04	4.00E-05	0	1.00E-22	0	0	0	0
5X2	H0	((P_steineni,(D_melanoga,D_mauritia)),B_mori);	1	1	-47.6	1	1	1	1	1	1	1	1
	H1	((D_melanoga,(P_steineni,D_mauritia)),B_mori);	2	2	47.6	2.00E-05	8.00E-06	0	2.00E-21	0	0	0	0
	H2	((D_mauritia,(D_melanoga,P_steineni)),B_mori);	3	3	48.7	3.00E-07	9.00E-07	0	7.00E-22	0	0	0	0
6X1	H0	(((C_tepperi,P_vanderpl),E_sorbilla),B_mori);	1	1	-183	1	1	1	1	1	1	1	1
	H2	(((E_sorbilla,P_vanderpl),C_tepperi),B_mori);	2	3	183	2.00E-06	7.00E-07	0	2.00E-80	0	0	0	0
	H1	(((C_tepperi,E_sorbilla),P_vanderpl),B_mori);	3	2	183	2.00E-06	7.00E-07	0	2.00E-80	0	0	0	0
6X2	H0	(((C_tepperi,P_vanderpl),P_steineni),B_mori);	1	1	-57.2	1	1	1	1	1	1	1	1
	H1	(((C_tepperi,P_steineni),P_vanderpl),B_mori);	2	2	57.2	4.00E-07	6.00E-06	0	2.00E-25	2.00E-05	2.00E-05	2.00E-05	2.00E-05
	H2	(((P_steineni,P_vanderpl),C_tepperi),B_mori);	3	3	64.9	3.00E-44	7.00E-16	0	7.00E-29	0	0	0	0
7X1	H0	(((S_variegat,S_aureohir),E_sorbilla),B_mori);	1	1	-147	1	1	1	1	1	1	1	1
	H2	(((S_variegat,E_sorbilla),S_aureohir),B_mori);	2	3	147	3.00E-05	2.00E-06	0	9.00E-65	0	0	0	0
	H1	(((E_sorbilla,S_aureohir),S_variegat),B_mori);	3	2	147	3.00E-05	2.00E-06	0	9.00E-65	0	0	0	0
7X2	H0	(((S_variegat,S_aureohir),P_steineni),B_mori);	1	1	-29.8	0.999	0.998	0.998	1	0.998	1	0.998	0.999
	H1	(((S_variegat,P_steineni),S_aureohir),B_mori);	2	2	29.8	0.001	0.002	0.002	1.00E-13	0.002	0.002	0.002	0.004
	H2	(((P_steineni,S_aureohir),S_variegat),B_mori);	3	3	37.6	1.00E-10	6.00E-08	0	5.00E-17	0	0	0	0
8X1	H2	(((D_melanoga,E_sorbilla),E_flavipal),B_mori);	1	3	-19.4	0.975	0.973	0.973	1	0.969	0.992	0.969	0.979
	H0	(((E_flavipal,E_sorbilla),D_melanoga),B_mori);	2	1	19.4	0.025	0.027	0.027	4.00E-09	0.031	0.031	0.031	0.048
	H1	(((E_flavipal,D_melanoga),E_sorbilla),B_mori);	3	2	27.8	5.00E-51	1.00E-16	0	8.00E-13	3.00E-04	3.00E-04	3.00E-04	0.001
8X2	H1	((E_flavipal,(P_steineni,D_melanoga)),B_mori);	1	2	-3.4	0.647	0.652	0.654	0.969	0.651	0.77	0.651	0.736
	H0	((P_steineni,(E_flavipal,D_melanoga)),B_mori);	2	1	3.4	0.353	0.348	0.346	0.031	0.349	0.401	0.349	0.451
	H2	((D_melanoga,(E_flavipal,P_steineni)),B_mori]);	3	3	15.7	8.00E-16	3.00E-09	0	1.00E-07	0.006	0.007	0.006	0.01
9X3	H2	((P_vanderpl,(S_variegat,A_gambie)),B_mori);	1	3	-15.6	0.946	0.945	0.945	1	0.943	0.982	0.943	0.963
	H0	((A_gambie,(S_variegat,P_vanderpl)),B_mori);	2	1	15.6	0.055	0.055	0.055	2.00E-07	0.057	0.057	0.057	0.091
	H1	((S_variegat,(A_gambie,P_vanderpl)),B_mori);	3	2	24	5.00E-05	6.00E-06	0	4.00E-11	0.001	0.001	0.001	0.002
9X4	H0	((S_grandico,(A_gambie,S_variegat)),B_mori);	1	1	-17.5	0.969	0.968	0.969	1.00E+00	0.964	0.988	0.964	0.979
	H1	((A_gambie,(S_grandico,S_variegat)),B_mori);	2	2	17.5	0.032	0.032	0.031	2.00E-08	0.036	0.036	0.036	0.062
	H2	((S_variegat,(S_grandico,A_gambie)),B_mori);	3	3	23.8	5.00E-04	4.00E-05	2.00E-04	5.00E-11	0.002	0.002	0.002	0.004
10X3	H0	((A_gambie,(S_variegat,P_steineni)),B_mori);	1	1	-44.6	1	1	1	1	1	1	1	1
	H2	((S_variegat,(A_gambie,P_steineni)),B_mori);	2	3	44.6	3.00E-04	2.00E-04	0	4.00E-20	5.00E-05	5.00E-05	5.00E-05	5.00E-05
	H1	((P_steineni,(S_variegat,A_gambie)),B_mori);	3	2	44.8	4.00E-05	1.00E-05	0	4.00E-20	5.00E-05	5.00E-05	5.00E-05	4.00E-05
10X4	H0	((S_grandico,(S_variegat,P_steineni)),B_mori);	1	1	-72	1	1	1	1	1	1	1	1
	H1	((S_variegat,(S_grandico,P_steineni)),B_mori);	2	2	72	1.00E-05	2.00E-06	0	6.00E-32	0	0	0	0
	H2	((P_steineni,(S_variegat,S_grandico)),B_mori);	3	3	72	1.00E-05	2.00E-06	0	6.00E-32	0	0	0	0

rank: the order of the tree, item: the label for the tree. obs: the observed log-likelihood difference, au: the p-value of the approximately unbiased test calculated from the multiscale bootstrap, np: the bootstrap probability calculated from the multiscale bootstrap, bp: the bootstrap probability calculated in the usual manner, kh: the Kishino-Hasegawa test, sh: the Shimodaira-Hasegawa test, wkh: the weighted Kishino-Hasegawa test, wsh: the weighted Shimodaira-Hasegawa test.

supported H1 and H2, respectively. H2 and H1 were supported by 1X2 and 8X2 quartets in which Brachycera fly *D. melanogaster* clustered with Nematocera fly *P. steinenii* instead of *Blepharipa sp.* and *E. flavipalpis*. The quartet, 3X2 supported H2 where a Chironomidae fly *P. steinenii* grouped with a Simuliidae fly, *S. variegatum* instead of another Chironomidae fly *P. vanderplanki*, although these three species belong to same superfamily. The quartet, 9X3 supported H2 in which *S. variegatum*, a taxon from Chironomoidea superfamily clustered with *A. gambiae* which belong to Culicoidea superfamily rather than *P. vanderplanki* of Chironomoidea superfamily. Across all 20 quartets, we found that the majority of genes had very low TC values (the highest TC value for any single node is 1) (data not shown here). It

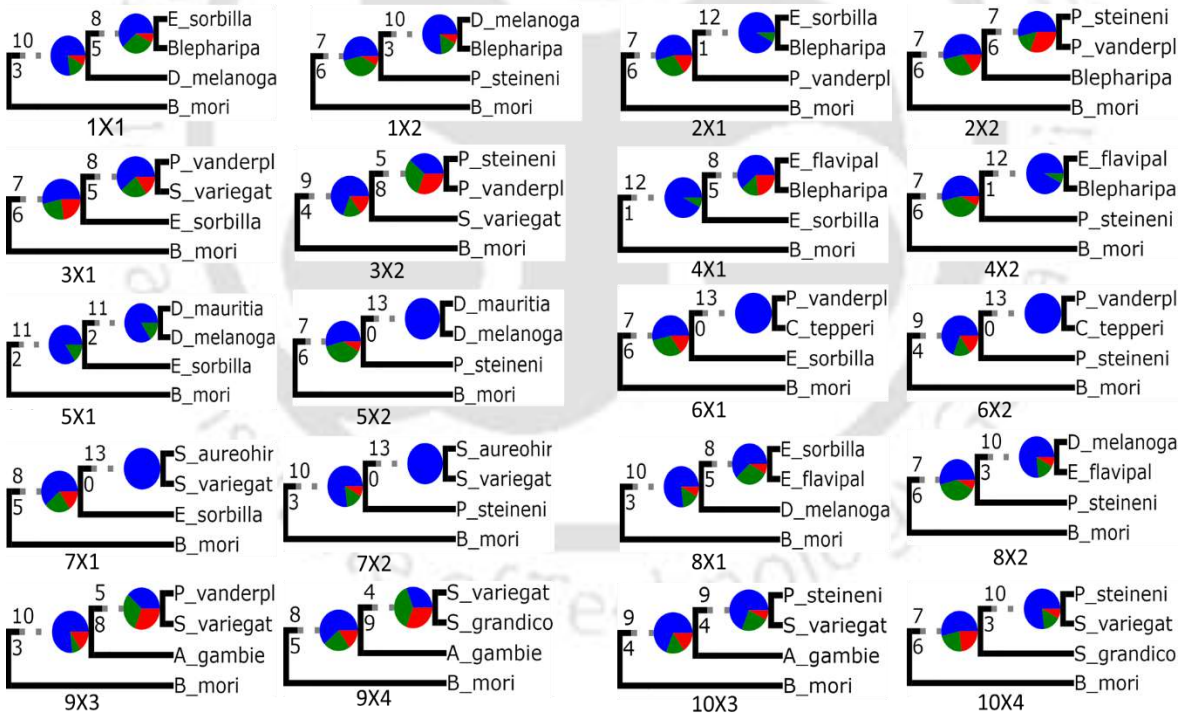


Figure 3.18: Quartet based Internode Certainty Analyses (ICA); Pie charts next to the nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support)

suggests that individual gene trees have high levels of disagreement across bootstrap replicates, indicating uninformative genes and correlating with the AU topological test results[173].

The quartet-based internode certainty analyses also suggest that most of the gene trees failed to entirely support the ASTRAL species tree topology (Fig. 3.18). As evidenced in the quartets 1X1 and 8X1, the connecting node between *Blepharipa sp.* and *E. sorbilans* contains 8 concordant homologs to support the ASTRAL species tree, but the linking node with *D. melanogaster* has 10 homologs to support the ASTRAL species tree. The quartets 1X2 and 8X2 show that the node connecting *D. melanogaster* to *Blepharipa sp.* and *E. flavipalpis* contains 10 concordant homologs, but the node connecting that group to *P. steinenii* has seven. The quartet 3X1 show that node between *P. vanderplanki* and *S. variegatum* has 8 concordant homologs, while quartet 3X2 shows that node between *P. vanderplanki* and *P. steinenii* has 5 concordant homologs. The quartet 9X3 displays 5 concordant homologs between *P. vanderplanki* and *S. variegatum*, while the node linking that group to *A. gambiae* has 10 concordant homologs.

3.3.10 Analysis of Introgression:

Definition of introgression states that transfer of genetic information from one species to another as a result of hybridization between them and repeated backcrossing. Here we created 20 different quartet scenario acquiring knowledge from previous reticulation result to analyze is there any introgression happen or not in different Diptera taxa Z-scores larger than 3 are considered strong evidence of introgression, and we identified seven such cases with Z-scores >3 using evobir analysis (Table 3.5). Result shows *E. sorbilans* has significant introgression with *D. melanogaster*. Whereas, *P. vanderplanki* and *P. steinenii* have significant introgression with *S. variegatum.* Further, to decipher introgression event, estimation of *D*-statistics and *f*4-ratio through ADMIXTOOLS indicates that *E. sorbilans* has significant similarity with *D. melanogaster* and *E. sorbilans* shares significant proportion of ancestry with Drosophila flies

Table 3.5: Sharing of derived alleles between different Diptera species; D = (nABBA - nBABA)/(nABBA + nBABA); Null Hypothesis: D=0, No introgression

	H1	H2	X	ABBA	BABA	D raw statistic	Z-score	SD D statistic	P-value (that D=0)	
1X1	*Blepharipa*	*D. melanogaster*	*E. sorbilans*	282	169	0.250554	5.337766	0.04694	9.41E-08	introgression bw and H2
1X2	*Blepharipa*	*D. melanogaster*	*P. steinenii*	223	179	0.109453	2.18022	0.050203	0.029241	
2X1	*Blepharipa*	*P. vanderplanki*	*E. sorbilans*	168	364	-0.36842	9.404846	0.039174	0	introgression bw and H1
2X2	*Blepharipa*	*P. vanderplanki*	*P. steinenii*	259	249	0.019685	0.460088	0.042785	0.645453	
3X1	*S. variegatum*	*P. vanderplanki*	*E. sorbilans*	178	284	-0.22944	5.215167	0.043994	1.84E-07	introgression bw and H1
3X2	*S. variegatum*	*P. vanderplanki*	*P. steinenii*	202	380	-0.30584	7.880125	0.038812	3.33E-15	introgression bw and H1
4X1	*Blepharipa*	*E. flavipalpis*	*E. sorbilans*	184	142	0.128834	2.326146	0.055385	0.020011	
4X2	*Blepharipa*	*E. flavipalpis*	*P. Steinenii*	140	147	-0.02439	0.408542	0.059701	0.682876	
5X1	*D. melanogaster*	*D.mauritiana*	*E. sorbilans*	46	46	0	0	0.106416	1	No introgression
5X2	*D. melanogaster*	*D.mauritiana*	*P. steinenii*	60	39	0.212121	2.177341	0.097422	0.029455	
6X1	*P. vanderplanki*	*C. tepperi*	*E. sorbilans*	204	167	0.09973	1.969931	0.050626	0.048846	
6X2	*P. vanderplanki*	*C. Tepperi*	*P. steinenii*	245	153	0.231156	4.785789	0.0483	1.70E-06	introgression bw and H2
7X1	*S. variegatum*	*S. aureohirtum*	*E. sorbilans*	122	131	-0.03557	0.572551	0.062131	0.566948	
7X2	*S. variegatum*	*S. aureohirtum*	*P. Steinenii*	147	180	-0.10092	1.897605	0.053181	0.057748	
8X1	*D. melanogaster*	*E. flavipalpis*	*E. sorbilans*	180	256	-0.17431	3.658511	0.047646	0.000254	introgression bw and H1
8X2	*D. melanogaster*	*E. flavipalpis*	*P.steineni*	152	202	-0.14124	2.654947	0.0532	0.007932	
9X3	*S. Variegatum*	*A.gambie*	*P. vanderplanki*	203	228	-0.058	1.224432	0.047373	0.220789	
9X4	*S. Variegatum*	*A.gambie*	*S. grandicornis*	222	219	0.006803	0.143012	0.047568	0.886281	
10X3	*S. Variegatum*	*P. steinenii*	*P. vanderplanki*	192	282	-0.18987	4.210616	0.045094	2.55E-05	introgression bw and H1
10X4	*S. Variegatum*	*P. steinenii*	*S. grandicornis*	164	191	-0.07606	1.473335	0.051622	0.140661	

(Fig 3.19). In another analysis we calculated f-branch statistics with D-suite software and this result also in line with the *D*-statistics and *f*4-ratio result. The f-branch matrix shown in Figure 3.20 suggested that the two cells highlighted by the black arrow refers to excess allele sharing between species S3 (*D. melanogaster*) and S4 (*D. mauritiana*) and the branch leading to species S2 (*E. sorbilans*) (Fig 3.19). Therefore, all the investigations for testing introgression directed to one thing that sharing of genetic element (either allele or gene) did occur between *E. sorbilans* and Drosophila flies.

3.4 Conclusion:

Throughout this chapter, we have discussed how different types of heterogeneity in nucleotide sequences might make it difficult to acquire an appropriate phylogenetic resolution for a

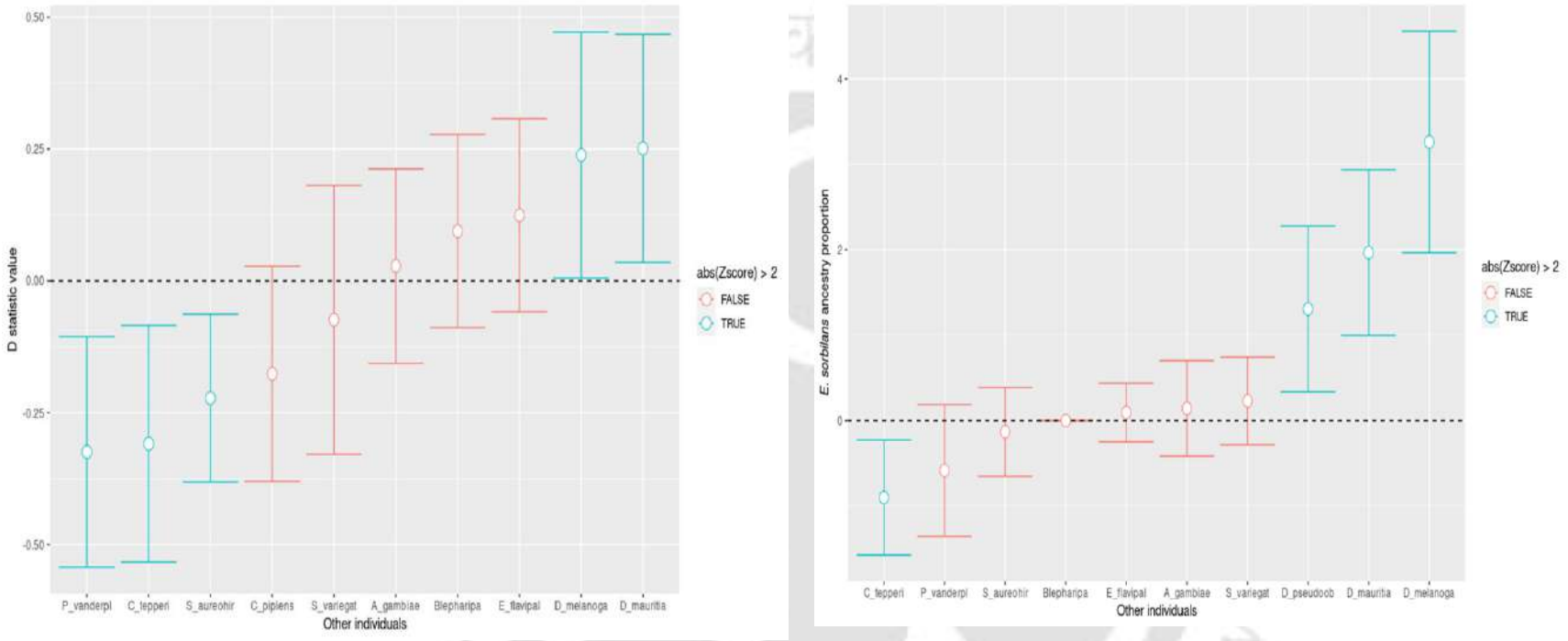


Figure 3.19: Test of introgression through *D-statistics* (Left) and *f*4-ratio statistics in selected species. The left graph shows *E. sorbilans* significantly similar with *D. melanogaster*; the right graph *E. sorbilans* shares significant proportion of ancestry with Drosophila flies.

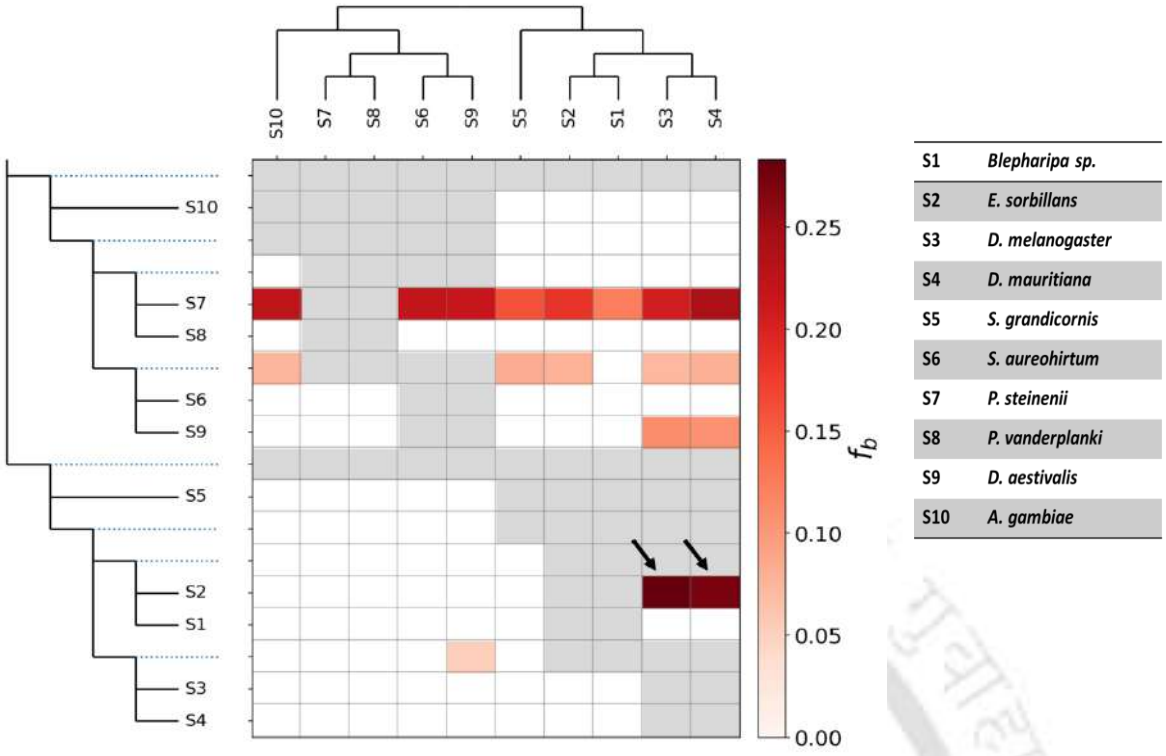


Figure 3.20: Results of Fbranch matrix for 10 selected taxa of Diptera. The tree is shown in an 'expanded' form along the y axis, so that each branch, including internal branches, points to a corresponding row in the matrix with inferred f-branch statistics. The values in the matrix thus refer to excess allele sharing between the branch b identified on the expanded tree on the y axis (relative to its sister branch) and the species P3 identified on the x-axis.

specific taxon. Despite its relevance, there has been an insufficient study on analytical techniques for experimental design in phylogenetics. Only a few approaches, such as the empirical saturation plot, likelihood mapping, and Treeness triangles, have been devised to deal with the issue of experimental design in the context of topological uncertainty[160,174,175]. These graphical approaches have a number of drawbacks, including being difficult to interpret and impractical for large-scale surveys[12,174]. As from the homogeneous phylogenetic analysis through maximum likelihood and Bayesian analysis the clade support from bootstrap or posterior probability provides substantial confidence. However, it does not address whether site patterns reflect historical signal for deep divergences, because it cannot account for systematic errors such as mutational bias[174,176]. In this work, we calculated the phylogenetic informativeness of genes over the entire period, from time zero to the root. We noticed that the

order of genes in terms of informativeness varies over time due to rate differences among sites. The comparison of the informativeness profiles for our datasets against their different chronograms revealed the varying ability of genes and sites for signal throughout their evolution. In this context, slower evolving substitution sites provide some power for deep polytomy resolution, whereas fast evolving synonymous sites evolve too rapidly for proper resolution of the relatively recent polytomy. Thus, set of character underlying profile of BI and IQ tree would be poor choice for the resolution of obscure branching events. Several loci displayed very sharp peaks nearer to tips which are artefacts that arise when the function used to measure substitution levels is unable to produce reliable predictions for locations with indels or ambiguities[177]. In other words, because the nucleotide sequence of protein-coding genes has a consistent, recognizable fast-evolving site class, such as degenerate third codon position sites, polymorphism at silent coding sites can lead to a high informativeness profiles for genes[12]. Furthermore, greater homoplasy in nucleotide sequence may have resulted in inferior predictions because signal is accounted for in the Townsend's informativeness but the possibility for misleading homoplasy is not[10,12].

Multiple factors could have contributed to signal loss and variation owing to homoplasy in accurately inferring Diptera phylogenetic relationships with the dataset of mitochondrial protein coding genes used in our study[5,39]. Different types of heterogeneity in nucleotide data might generate phylogenetic artefacts, such as among-lineage or among-site compositional heterogeneity, codon-usage bias, substitutional saturation, and so on. Several interesting patterns of base compositional heterogeneity are revealed using disparity index values calculated from individual codon positions and matched-pairs tests of symmetry. First, if a taxon evolves non stationarily, it can be expressed in all three codon positions. Second, a taxon with compositional bias may express more heterogeneity in certain codon positions than others, which deviates from the first pattern. Third, the 3rd codon position shows the majority of the

base compositional heterogeneity, followed by the 1st and 2nd codon positions. Indeed, the 2nd position is the least affected by compositional bias, most likely due to functional constraints; a similar pattern has been observed in other mitogenomes[18,42]. Typical phylogenetic inference methods consistently perform poorly when the data is biased. Among these methods, we encounter that a parsimony analysis without any data transformation is the most severely affected[42,178]. The LogDet transformation is well-known for dealing with base compositional heterogeneity; however, after bootstrapping analysis, the LogDet transformed data can group *E. sorbilans* with the Oestroidea superfamily at the expense of polytomy of several other branches[15].

Substitutional saturation reduces the accuracy of phylogenetic signals and has an impact on nucleotide sequences due to genetic code redundancy. Saturation is a consequence of time and mutation rate, so it affects mainly the resolution of deep nodes, as demonstrated by the reconstruction of arthropod[32] or animal phylogeny[40]. When sequences have reached full saturation, the similarity between them may entirely depend on nucleotide frequencies, so phylogenetic inferences based on such data will present ambiguous relationships based on compositional similarities among sites and lineages[39,179]. Substitution models in phylogenetic analyses can only account for saturation to a limited extent by modelling site rates (typically a mixture model of gamma-distributed among-site rates) and compositional biases (i.e., CAT and NDCH models). To minimize saturation issues in protein-coding data sets, it is typical practice to eliminate the third codon sites, which are mostly redundant in defining the codon and so have the fastest substitution rate. Alternatively, translation of the nucleotide sequence into an amino-acid sequence, can lessen the negative evolutionary consequences of saturation at the nucleotide level because many mutations in protein-coding genes result in synonymous codons[39]. Within our data set, the third codon positions are more severely saturated than first

and second codon positions (Fig. 3.7), and such obvious differences may explain the conflict in phylogenetic signal.

Another origin of inconsistency in the phylogenetic signal of nucleotide data sets could be lineage-specific base compositional heterogeneity, that has been shown to cause artefacts when not accommodated in the models [15,29,121]. Nonstationary heterogeneous composition models that allow base composition to vary across lineages may help to avoid such topological artefacts [121]. In the current mitochondrial data set, the nucleotide content plot (Fig. 3.5) and t-tests (data not shown) revealed that major Diptera lineages differ significantly in their GC composition, implying that compositional biases could also reason behind the incongruence[39].

Degeneracy in the universal genetic code allows synonymous codons to be used preferentially within a genome and across species, possibly resulting in highly distinct forms of codon-usage among lineages[180,181].Codon-usage bias among lineages can lead to dispersion of homoplasious character and, as a result, anomalous phylogenetic inferences[18,121,182]. Since most substitutions in 3rd codon positions are synonymous, excluding these positions can partially alleviate the problem of compositional convergence, but not adequately to eliminate all influences owing to codon degeneracy at the first or second codon position. Thus, recoding of nucleotide sequences to codons by using IUPAC codes that reflect alternative "synonymous" bases at the three codon positions to assess and reduce the impact of convergent compositional heterogeneity driven by codon redundancy[18,183]. However, degenerating nucleotide data by codons to remove "noisy" phylogenetic signals caused by codon/composition biases may allow an actual, and accurate, phylogenetic signal to take precedence[39].

As shown in this study, Insect mitochondrial genomes display considerable base compositional and mutational rate variability between genes, codon sites within a gene, and taxonomic levels [71,184,185]. Such heterogeneities defy the widely used site-homogeneous nucleotide substitution

models' assumption of stationarity[13,23,69]. Despite the fact that evolutionary processes vary by position, site-homogeneous models assume the same evolutionary process for every site in the data set [40]. In PhyloBayes, the CAT + GTR is better adapted to account for this variation in the process of evolution across sites[40,69]. After removing 3rd codon positions, the site-heterogeneous mixture model corresponded worsen the phylogenetic inference to Dipteran mitochondrial GB data sets in this study.

However, data biases caused by saturation of variation and high degrees of homoplasy still influence the site-heterogeneous model. The fast evolving regions in mitochondrial genomes, especially the 3rd codon position of protein-coding genes, are predicted to be the most varied in composition and saturated in substitutions, and they frequently lead to phylogenetic artefacts [38,186]. Previous study has shown that eliminating these sites or include two rRNAs, as well as using amino acid data to deal with systematic errors, can improve phylogenetic signal and signal-to-noise ratio[39,59,185,186].

Investigation of gene tree discordance has become a vital step in gaining a better understanding of intractable relationships across the Tree of Life. Recently, new techniques for detecting and visualizing gene tree discordance have been created[143–145]. However, downstream techniques for analyzing the processes that cause observable patterns of gene tree discordance are still in their early stages. We were able to interpret the causes of conflict throughout the backbone phylogeny of Diptera using mitogenomes. We found that gene tree heterogeneity in the Diptera mitogenome may be explained by a number of phenomena, such as ancient hybridization and uninformative genes, that might have occurred concurrently and/or cumulatively[173].

All methods that we used to detect ancient hybridization inferred the presence of reticulation events. However, our results suggest that these methods all struggle with ancient, rapid radiations. Species network inference, on the other hand, is still computationally costly and

confined to a small number of species and a few hybridization events at the moment (mentioned earlier). Furthermore, analyses comparing the effectiveness of various phylogenetic network inference methodologies are limited to basic hybridization situations[187]. Applying three methods on our 10-taxon(net) data set, we were able to find multiple reticulation events involving our target taxa (*E. sorbilans* and *P. steinenii*) (Fig. 3.17). Furthermore, the neighbour-net analysis in SplitsTree network (Fig. 3.16) revealed the presence of box like structure between the *E. sorbilans* of Tachinidae family and *D. melanogaster* of Drosophilidae family. The networked relationships between the sequences with box-like structures rather than bifurcations confirmed the notion that reticulation might have happened during their evolution. Reticulations have been hypothesized to promote subsequent rapid diversification events in a diverse range of species[94,188,189]. Reticulate evolution is most likely to induce rapid diversification when parental lineages occupy extremely dissimilar ecological niches but are only marginally genetically segregated[190]. Rapid evolutionary radiations have also been proposed to explain poorly resolved phylogenies across a wide range of species group, including aphids, black flies, bees, birds, turtles, mammals, and higher plants, as observed in our work[191–196].

We also identified signals of introgression in our taxa of interest using the *D*-statistic[152,153]. The estimated introgression events agreed with at least one of the phylogenetic network analysis's reticulation hypotheses. The *D*-statistic technique measures the amount of ABBA and BABA sites, parsimony-informative sites that support a phylogeny other than the species tree to see whether they are statistically equal. The two discordant genealogies with the species tree, ABBA and BABA, are equally likely to occur by ILS (incomplete lineage sorting); hence, their numbers should not vary if it's only ILS, but not gene flow, is present[197]. A substantial disparity between ABBA and BABA sites implies that two unrelated species are more similar than expected, which is regarded as an evidence of gene flow. In this work, *D*-statistic analysis using

two distinct approaches revealed that *E. sorbilans*, also known as the Tachinidae fly, has an excess of ABBA sites, which leads to introgression with the fruit fly *D. melanogaster* (Table 3.5, Fig. 3.19 (left)). Previously the *D*-statistic has been used in several studies to detect gene flow between related species of bears[198] , equids[199] , butterflies[200] as well as hominids[153]. Similarly, the *f*4-ratio[155] statistics which used to detect the proportion of ancestry shared by any potential admixed species also in line with the *D*-statistics result suggesting that extant *E. sorbilans* carry significant amount of Drosophila ancestry (Fig. 3.19 (right)). Further, the *f*-branch statistic matrix displayed the amount of probable gene flow between donor-recipient combinations indicated significant amount of gene flow occurred between *E. sorbilans* Drosophila flies (Fig. 3.20).

Hybridization, traditionally thought to be an evolutionary dead end, is more widespread than previously anticipated, occurring in plants, fungi, and mammals[201–203]. Oversimplifying evolutionary assumptions or leading to erroneous or inconsistent interpretations might result from failing to acknowledge the possible effect of hybridization on speciation and delineating species boundaries[204]. Phylogenomic analysis can provide a robust tree, however the well-supported result might still be incorrect and lacking in phylogenetic signal. The loss of signal can be attributed to numerous reasons as we analyzed and discussed throughout this study such as, base compositional heterogeneity, substitutional saturation, codon usage bias, among-site rate variation and so on. By analyzing phylogenomic data from mitochondrial genomes of Diptera flies and reconstructing a phylogenetic network, we found some evidence of a gene tree discordance as well as a reticulated evolutionary history. The magnitude of hybridization's impact on speciation and the ability to identify it are highly debated in the scientific community[205,206]; yet, studying the effects of interspecific gene flow on species as a whole is still worthwhile.

3.5 References:

1. Bertone, Matthew A., and B. M. W. *True fies (Diptera). The timetree of life* (2009).
2. Wiegmann, B. M. *et al.* Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* **108**, 5690–5695 (2011).
3. Grimaldi, D. A. & Engel, M. S. *Evolution of the insects*. (Cambridge University Press, 2005).
4. Skevington, J. H. & Dang, P. T. Exploring the diversity of flies (Diptera). *Biodiversity* **3**, 3–27 (2002).
5. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* vol. 6 361–375 (2005).
6. Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13**, 122–134 (2012).
7. Klopfstein, S., Kropf, C. & Quicke, D. L. J. An Evaluation of Phylogenetic Informativeness Profiles and the Molecular Phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). *Syst. Biol.* **59**, 226–241 (2010).
8. Kainer, D. & Lanfear, R. The Effects of Partitioning on Phylogenetic Inference. *Mol. Biol. Evol.* **32**, 1611–1627 (2015).
9. Abadi, S., Azouri, D., Pupko, T. & Mayrose, I. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun. 2019 101* **10**, 1–11 (2019).
10. Townsend, J. P. Profiling Phylogenetic Informativeness. *Syst. Biol.* **56**, 222–231 (2007).
11. Rosenberg, M. S. & Kumar, S. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10751 (2001).
12. López-Giráldez, F., Moeller, A. H. & Townsend, J. P. Evaluating Phylogenetic Informativeness as a Predictor of Phylogenetic Signal for Metazoan, Fungal, and Mammalian Phylogenomic Data Sets. *Biomed Res. Int.* **2013**, (2013).
13. Rosenberg, M. S. & Kumar, S. Taxon Sampling, Bioinformatics, and Phylogenomics. *Syst. Biol.* **52**, 119–124 (2003).
14. Brinkmann, H., van der Giezen, M., Zhou, Y., de Raucourt, G. P. & Philippe, H. An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics. *Syst. Biol.* **54**, 743–757 (2005).
15. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11**, 605–612 (1994).
16. Galtier, N. & Gouy, M. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 11317–11321 (1995).
17. Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J. & Larkum, A. W. D. The Biasing Effect of Compositional Heterogeneity on Phylogenetic Estimates May be Underestimated. *Syst. Biol.* **53**, 638–643 (2004).
18. Gibson, A., Gowri-Shankar, V., Higgs, P. G. & Rattray, M. A Comprehensive Analysis of Mammalian Mitochondrial Genome Base Composition and Improved Phylogenetic Methods. *Mol. Biol. Evol.* **22**, 251–264 (2005).
19. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372 (1996).
20. Felsenstein, J. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.* **53**, 447–455 (2001).

21. Mayrose, I., Friedman, N. & Pupko, T. A Gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* **21**, 151–158 (2005).

22. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol. 2005 51* **5**, 1–8 (2005).

23. Bryan Kolaczkowski & Joseph W. Thornton. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984 (2004).

24. Felsenstein, J. Confidence Limits on Phylogenies : An Approach Using the Bootstrap. *Evolution (N. Y).* **39**, 783–791 (1985).

25. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinforma. Appl. NOTE* **17**, 754–755 (2001).

26. Huelsenbeck, J. P. When are Fossils Better than Extant Taxa in Phylogenetic Analysis? *Syst. Biol.* **40**, 458–469 (1991).

27. Farris, J. S. The Retention Index and The Rescaled Consistency Index. *Cladistics* **5**, 417–419 (1989).

28. Townsend, J. P., Su, Z. & Tekle, Y. I. Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny. *Syst. Biol.* **61**, 835 (2012).

29. Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J. & Larkum, A. W. D. Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.* **34**, 153–162 (1992).

30. Felsenstein, J. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Syst. Zool.* **27**, 401–410 (1978).

31. Yang, Z. Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites. *Mol. Biol. Evol.* **10**, 1396–1401 (1993).

32. Xia, X., Xie, Z., Salemi, M., Chen, L. & Wang, Y. An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* **26**, 1–7 (2003).

33. Ho, S. Y. W. & Jermiin, L. S. Tracing the Decay of the Historical Signal in Biological Sequence Data. *Syst. Biol.* **53**, 623–637 (2004).

34. Lake, J. A. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1455–1459 (1994).

35. Hasegawa, M. & Fujiwara, M. Relative Efficiencies of the Maximum Likelihood, Maximum Parsimony, and Neighbor-Joining Methods for Estimating Protein Phylogeny. *Mol. Phylogenet. Evol.* **2**, 1–5 (1993).

36. Crispell, J., Balaz, D. & Gordon, S. V. Homoplasyfinder: A simple tool to identify homoplasies on a phylogeny. *Microb. Genomics* **5**, e000245 (2019).

37. Goremykin, V. V. *et al.* The Evolutionary Root of Flowering Plants. *Syst. Biol.* **62**, 50–61 (2013).

38. Pisani, D. Identifying and Removing Fast-Evolving Sites Using Compatibility Analysis: An Example from the Arthropoda. *Syst. Biol.* **53**, 978–989 (2004).

39. Liu, Y., Cox, C. J., Wang, W. & Goffinet, B. Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Syst. Biol.* **63**, 862–878 (2014).

40. Philippe, H. *et al.* Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol.* **9**, e1000602 (2011).

41. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).

42. Song, H., Sheffield, N. C., Cameron, S. L., Miller, K. B. & Whiting, M. F. When phylogenetic

assumptions are violated: Base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Syst. Entomol.* **35**, 429–448 (2010).

43. Reed, R. D. & Sperling, F. A. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus Papilio. *Mol. Biol. Evol.* **16**, 286–297 (1999).

44. Kelava, S. *et al.* Phylogenies from mitochondrial genomes of 120 species of ticks: Insights into the evolution of the families of ticks and of the genus Amblyomma. *Ticks Tick. Borne. Dis.* **12**, 101577 (2021).

45. Lin, C. P. & Danforth, B. N. How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. *Mol. Phylogenet. Evol.* **30**, 686–702 (2004).

46. Baurain, D., Brinkmann, H. & Philippe, H. Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? *Mol. Biol. Evol.* **24**, 6–9 (2007).

47. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nat. 2008 4527188* **452**, 745–749 (2008).

48. Nardi, F. *et al.* Hexapod origins: Monophyletic or paraphyletic? *Science (80-.).* **299**, 1887–1889 (2003).

49. Cameron, S. L., Miller, K. B., D'Haese, C. A., Whiting, M. F. & Barker, S. C. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda). *Cladistics* **20**, 534–557 (2004).

50. Castro, L. R. & Dowton, M. Mitochondrial genomes in the Hymenoptera and their utility as phylogenetic markers. *Syst. Entomol.* **32**, 60–69 (2007).

51. Delsuc, F., Phillips, M. J. & Penny, D. Comment on 'Hexapod origins: monophyletic or paraphyletic?'. *Science (80-.).* **301**, (2003).

52. Phillips, M. J., Delsuc, F. & Penny, D. Genome-Scale Phylogeny and the Detection of Systematic Biases. *Mol. Biol. Evol.* **21**, 1455–1458 (2004).

53. Beckenbach, A. T. Mitochondrial genome sequences of nematocera (lower diptera): Evidence of rearrangement following a complete genome duplication in a winter crane fly. *Genome Biol. Evol.* **4**, 89–101 (2012).

54. Cameron, S. L. & Whiting, M. F. The complete mitochondrial genome of the tobacco hornworm, Manduca sexta, (Insecta: Lepidoptera: Sphingidae), and an examination of mitochondrial gene variability within butterflies and moths. *Gene* **408**, 112–123 (2008).

55. Dowton, M., Castro, L. R. & Austin, A. D. Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: The examination of genome 'morphology'. *Invertebrate Systematics* vol. 16 345–356 (2002).

56. Macey, J. R., Larson, A., Ananjeva, N. B. & Papenfuss, T. J. Evolutionary Shifts in Three Major Structural Features of the Mitochondrial Genome Among Iguanian Lizards. *J. Mol. Evol.* **44**, 660–674 (1997).

57. Kostka, M., Uzlikova, M., Cepicka, I. & Flegr, J. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. *BMC Bioinforma. 2008 91* **9**, 1–6 (2008).

58. Goremykin, V. V., Nikiforova, S. V. & Bininda-Emonds, O. R. P. Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol.* **71**, 319–331 (2010).

59. Zhong, B. *et al.* Systematic Error in Seed Plant Phylogenomics. *Genome Biol. Evol.* **3**, 1340 (2011).

60. Kim, J. Large-Scale Phylogenies and Measuring the Performance of Phylogenetic Estimators. *Syst. Biol.* **47**, 43–60 (1998).

61. Yang, Z. & Rannala, B. Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics* vol. 13 303–314 (2012).

62. Hillis, D. M., Huelsenbeck, J. P. & Swofford, D. L. Hobgoblin of phylogenetics? *Nature* vol. 369 363–364 (1994).

63. Lockhart, P. J., Larkumt, A. W. D., Steelt, M. A., Waddell, P. J. & Penny, D. *Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis*. vol. 93 (1996).

64. Chang, J. T. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosci.* **134**, 189–215 (1996).

65. Nishihara, H., Okada, N. & Hasegawa, M. Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol.* **8**, 199 (2007).

66. Philips, T. K., Pretorius, E. & Scholtz, C. A phylogenetic analysis of dung beetles (Scarabaeinae:Scarabaeidae): unrolling an evolutionary history. *Invertebr. Syst.* **18**, 53–88 (2004).

67. Fiala, K. L. & Sokal, R. R. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. *Evolution (N. Y).* **39**, 609–622 (1985).

68. Poe, S. Points of View Evaluation of the Strategy of Long-Branch Subdivision to Improve the Accuracy of Phylogenetic Methods. *Syst. Biol* **52**, 423–428 (2003).

69. Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).

70. Tarrío, R., Rodríguez-Trelles, F. & Ayala, F. J. Shared Nucleotide Composition Biases Among Species and Their Impact on Phylogenetic Reconstructions of the Drosophilidae. *Mol. Biol. Evol* **18**, 1464–1473 (2001).

71. Sheffield, N. C., Song, H., Cameron, S. L. & Whiting, M. F. Nonstationary Evolution and Compositional Heterogeneity in Beetle Mitochondrial Phylogenomics. *Syst. Biol.* **58**, 381–394 (2009).

72. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).

73. Timmermans, M. J. T. N. *et al.* Family-Level Sampling of Mitochondrial Genomes in Coleoptera: Compositional Heterogeneity and Phylogenetics. *Genome Biol. Evol.* **8**, 161 (2016).

74. Li, H. *et al.* Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. *Sci. Reports 2015 51* **5**, 1–10 (2015).

75. Sullivan, J. & Joyce, P. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst* **36**, 445–66 (2005).

76. Brown, J. M. & Lemmon, A. R. The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics. *Syst. Biol.* **56**, 643–655 (2007).

77. Lanfear, R., Calcott, B., Kainer, D., Mayer, C. & Stamatakis, A. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* (2014) doi:10.1186/1471-2148-14-82.

78. Powell, A. F. L. A., Barker, F. K. & Lanyon, S. M. Empirical evaluation of partitioning schemes for phylogenetic analyses of mitogenomic data: An avian case study. *Mol. Phylogenet. Evol.* (2013) doi:10.1016/j.ympev.2012.09.006.

79. Smith, S. A. *et al.* Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* **480**, 364–367 (2011).

80. Chaudhary, R., Burleigh, J. G. & Fernández-Baca, D. Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol. Biol. 2013 81* **8**, 1–12 (2013).

81. Sharma, P. P. *et al.* Phylogenomic Interrogation of Arachnida Reveals Systemic Conflicts in Phylogenetic Signal. *Mol. Biol. Evol.* **31**, 2963–2984 (2014).

82. Xi, Z., Liu, L., Rest, J. S. & Davis, C. C. Coalescent versus Concatenation Methods and the Placement

of Amborella as Sister to Water Lilies. *Syst. Biol.* **63**, 919–932 (2014).

83. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–31 (2014).

84. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).

85. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804 (2003).

86. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Mol. Biol. Evol.* **33**, 1606–1617 (2016).

87. Ané, C., Larget, B., Baum, D. A., Smith, S. D. & Rokas, A. Bayesian Estimation of Concordance among Gene Trees. *Mol. Biol. Evol.* **24**, 412–426 (2007).

88. Allman, E. S., Kubatko, L. S. & Rhodes, J. A. Split Scores: A Tool to Quantify Phylogenetic Signal in Genome-Scale Data. *Syst. Biol.* **66**, 620–636 (2017).

89. Galtier, N. & Daubin, V. Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 4023–4029 (2008).

90. Morales-Briones, D. F., Liston, A. & Tank, D. C. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). *New Phytol.* **218**, 1668–1684 (2018).

91. Tigano, Anna, and V. L. F. Genomics of local adaptation with gene flow. *Mol. Ecol.* **25**, 2144–2164 (2016).

92. Gladieux, Pierre, Jeanne Ropars, Hélène Badouin, Antoine Branca, Gabriela Aguileta, Damien M. De Vienne, Ricardo C. Rodríguez de la Vega, Sara Branco, and T. G. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* **23**, 753–773 (2014).

93. Rieseberg, L. H. *et al.* Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. *Science (80-.).* **301**, 1211–1216 (2003).

94. Widhelm, T. J. *et al.* Multiple historical processes obscure phylogenetic relationships in a taxonomically difficult group (Lobariaceae, Ascomycota). *Sci. Rep.* **9**, 1–16 (2019).

95. Nakhleh, L., Warnow, T. & Linder, C. R. Reconstructing Reticulate Evolution in Species-Theory and Practice. *J. Comput. Biol.* **12**, 796–811 (2005).

96. Yu, Y., Dong, J., Liu, K. J. & Nakhleh, L. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci.* **111**, 16448–16453 (2014).

97. He, L., Wang, S., Miao, X., Wu, H. & Huang, Y. Identification of necrophagous fly species using ISSR and SCAR markers. *Forensic Sci. Int.* **168**, 148–153 (2007).

98. Núñez-Vázquez, C., Tomberlin, J. & García-Martínez, O. First Record of the Blow Fly *Calliphora grahami* [1] from Mexico. *Southwest. Entomol.* **35**, 313–316 (2010).

99. Yan, J., Liao, H., Xie, K. & Cai, J. The complete mitochondria genome of *Chrysomya pinguis* (Diptera: Calliphoridae). *Mitochondrial DNA Part A* **27**, 3852–3854 (2016).

100. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

101. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).

102. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic

Analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

103. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).

104. Sievers, F. & Higgins, D. G. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. in 105–116 (Humana Press, Totowa, NJ, 2014). doi:10.1007/978-1-62703-646-7_6.

105. Jian, S. *et al.* Resolving an ancient, rapid radiation in saxifragales. *Syst. Biol.* **57**, 38–57 (2008).

106. Barrett, C. F., Davis, J. I., Leebens-Mack, J., Conran, J. G. & Stevenson, D. W. Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* **29**, 65–87 (2013).

107. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

108. Mcguire, J. A., Witt, C. C., Altshuler, D. L. & Remsen, J. V. Phylogenetic Systematics and Biogeography of Hummingbirds: Bayesian and Maximum Likelihood Analyses of Partitioned Data and Selection of an Appropriate Partitioning Strategy. *Bio!* **56**, 837–856 (2007).

109. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. (2012) doi:10.1093/molbev/mss020.

110. Schmidt, H. A. & Haeseler, A. Maximum-Likelihood Analysis Using TREE-PUZZLE. *Curr. Protoc. Bioinforma.* **17**, 6.6.1-6.6.23 (2007).

111. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

112. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).

113. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **61**, 539–542 (2012).

114. Rambaut, Andrew, Marc A. Suchard, D. Xie, and A. J. D. Tracer v1. 6. 2014. (2015).

115. Swofford, D. L. PAUP*: phylogenetic analysis using parsimony. v. 4.0 b10. *Sinauer, Sunderl.* (2002).

116. Yang, Z. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J Mol Evol* **42**, 587–596 (1996).

117. 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. *Mol. Biol. Evol.* **32**, 268–274 (2015).

118. Lápez-Giráldez, F. & Townsend, J. P. PhyDesign: An online application for profiling phylogenetic informativeness. *BMC Evol. Biol.* **11**, 152 (2011).

119. Kosakovsky Pond, S. L., Frost, S. D. W. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005).

120. Moeller, A. H. & Townsend, J. P. Phylogenetic informativeness profiling of 12 genes for 28 vertebrate taxa without divergence dates. *Molecular Phylogenetics and Evolution* vol. 60 271–272 (2011).

121. Foster, P. G. Modeling Compositional Heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).

122. Inagaki, Y. & Roger, A. J. Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. *Mol. Phylogenet. Evol.* **40**, 428–434 (2006).

123. Stenøien, H. K. Adaptive basis of codon usage in the haploid moss Physcomitrella patens. *Heredity (Edinb).* **94**, 87–93 (2005).

124. Xia, X. DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *J. Hered.* **108**, 431–437 (2017).

125. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

126. Jermiin, L. S., Ho, J. W. K., Lau, K. W. & Jayaswal, V. SeqVis: A tool for detecting compositional heterogeneity among aligned nucleotide sequences. *Methods Mol. Biol.* **537**, 65–91 (2009).

127. Ababneh, F., Jermiin, L. S., Ma, C. & Robinson, J. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* **22**, 1225–1231 (2006).

128. Kück, P., Meid, S. A., Groß, C., Wägele, J. W. & Misof, B. AliGROOVE - visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support. *BMC Bioinformatics* **15**, 294 (2014).

129. Liu, Y. *et al.* Compositional heterogeneity in true bug mitochondrial phylogenomics. *Mol. Phylogenet. Evol.* **118**, 135–144 (2018).

130. Cutter, A. D., Wasmuth, J. D. & Blaxter, M. L. The evolution of biased codon and amino acid usage in nematode genomes. *Mol. Biol. Evol.* **23**, 2303–2315 (2006).

131. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).

132. Sun, X., Yang, Q. & Xia, X. An Improved Implementation of Effective Number of Codons (Nc). *Mol. Biol. Evol.* **30**, 191–196 (2013).

133. Li, H. *et al.* Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs. *Proc. R. Soc. B Biol. Sci.* **284**, (2017).

134. Termier, Alexandre, M-C. Rousset, and M. S. Treefinder: a first step towards xml data mining. in *IEEE International Conference on Data Mining, 2002. Proceedings.IEEE* 450–457 (2002).

135. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an Important Process of Protein Evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).

136. Crotty, S. M. *et al.* GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Syst. Biol* **69**, 249–264 (2020).

137. Fitch, W. M. Networks and Viral Evolution. *J. Mol. Evol.* **44**, 65–75 (1997).

138. Bryant, D. & Moulton, V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Mol. Biol. Evol.* **21**, 255–265 (2004).

139. Whitfield, J. B. & Lockhart, P. J. Deciphering ancient rapid radiations. *Trends in Ecology and Evolution* vol. 22 258–265 (2007).

140. Huson, D. H. & Bryant, D. Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).

141. Whitfield, J. B., Cameron, S. A., Huson, D. H. & Steel, M. A. Filtered Z-Closure Supernetworks for Extracting and Visualizing Recurrent Signal from Incongruent Gene Trees. *Syst. Biol.* **57**, 939–947 (2008).

142. Bandelt, H. J. & Dress, A. W. M. A canonical decomposition theory for metrics on a finite set. *Adv. Math. (N. Y).* **92**, 47–105 (1992).

143. Salichos, L., Stamatakis, A. & Rokas, A. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).

144. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 1–15 (2015).

145. Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E. & Smith, S. A. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.* **105**, 385–403 (2018).

146. Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L. & Tamura, K. Statistics and Truth in Phylogenomics. *Mol. Biol. Evol.* **29**, 457–472 (2012).

147. Than, C., Ruths, D. & Nakhleh, L. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9**, 1–16 (2008).

148. Yu, Y. & Nakhleh, L. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* **16**, 1–10 (2015).

149. Shimodaira, H. & Hasegawa, M. *CONSEL: for assessing the confidence of phylogenetic tree selection*. *Bioinformatics* vol. 17 http://www.ism.ac.jp/ (2001).

150. Mallet, J., Besansky, N. & Hahn, M. W. How reticulated are species? *BioEssays* **38**, 140–149 (2016).

151. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* **468**, 1053–1060 (2010).

152. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).

153. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science (80-.).* **328**, 710–723 (2010).

154. Blackmon, H. & Adams, R. H. EvobiR: Tools for comparative analyses and teaching evolutionary biology. (2015) doi:10.5281/ZENODO.30938.

155. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

156. Petr, M., Vernot, B. & Kelso, J. admixr—R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics* **35**, 3194–3195 (2019).

157. Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.* **2**, 1940–1955 (2018).

158. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23**, 1817–1828 (2013).

159. Hasegawa, M., Kishino, H. & Yano, T. aki. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).

160. Strimmer, K. & Von Haeseler, A. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 6815–6819 (1997).

161. Kabiraj, D. *et al.* Mitogenome-wise codon usage pattern from comparative analysis of the first mitogenome of Blepharipa sp. (Muga uzifly) with other Oestroid flies. *Sci. Rep.* **12**, 1–33 (2022).

162. Saccone, C., Pesole, G. & Preparata, G. DNA Microenvirenments and the Molecular Clock. *J. Mol. Evol.* **29**, 407–411 (1989).

163. Penny, D., Hendy, M. D., Zimmer, E. A. & Hamby, R. K. Trees from sequences: panacea or pandora's box? *Aust. Syst. Bot.* **3**, 21–38 (1990).

164. Steel, M. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.* **7**, 19–23 (1994).

165. Fitch, W. M. & Margoliash, E. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem. Genet.* **1**, 65–71 (1967).

166. Holmquist, R., Goodman, M., Conroy, T. & Czelusniak, J. The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.* **19**, 437–448 (1983).

167. Jayaswal, V., Wong, T. K. F., Robinson, J., Poladian, L. & Jermiin, L. S. Mixture Models of Nucleotide Sequence Evolution that Account for Heterogeneity in the Substitution Process Across Sites and Across Lineages. *Syst. Biol.* **63**, 726–742 (2014).

168. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7**, 1–14 (2007).

169. Whelan, N. V. & Halanych, K. M. Who Let the CAT Out of the Bag? Accurately Dealing with Substitutional Heterogeneity in Phylogenomic Analyses. *Syst. Biol.* **66**, 232–255 (2017).

170. N Saitou, M. N. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425 (1987).

171. Andreas W.M. Dress & Daniel H. Huson. Constructing Splits Graphs. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **1**, 109–115 (2004).

172. Wen, D., Yu, Y., Zhu, J. & Nakhleh, L. Inferring Phylogenetic Networks Using PhyloNet. *Syst. Biol.* **67**, 735–740 (2018).

173. Than, C. V. & Rosenberg, N. A. Consistency Properties of Species Tree Inference by Minimizing Deep Coalescences. *J. Comput. Biol.* **18**, 1–15 (2011).

174. Morales-Briones, D. F. *et al.* Disentangling Sources of Gene Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in Amaranthaceae s.l. *Syst. Biol.* **70**, 219–235 (2021).

175. White, W. T., Hills, S. F., Gaddam, R., Holland, B. R. & Penny, D. Treeness Triangles: Visualizing the Loss of Phylogenetic Signal. *Mol. Biol. Evol.* **24**, 2029–2039 (2007).

176. Graybeal, A. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. *Syst. Biol.* **43**, 174–193 (1994).

177. Buckley, T. R. Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets. *Syst. Biol.* **51**, 509–523 (2002).

178. Bellot, S., Mitchell, T. C. & Schaefer, H. Phylogenetic informativeness analyses to clarify past diversification processes in Cucurbitaceae. *Sci. Rep.* **10**, 1–13 (2020).

179. Collins, T. M., Wimberger, P. H. & Naylor, G. J. P. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Syst. Biol.* **43**, 482–496 (1994).

180. Rapoport, B. & Xuhua, X. Data Analysis in Molecular Biology and Evolution. *Springer Sci. Bus. Media* (2000) doi:10.1007/B113439.

181. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).

182. Liu, Q., Feng, Y. & Xue, Q. Analysis of factors shaping codon usage in the mitochondrion genome of Oryza sativa. *Mitochondrion* **4**, 313–20 (2004).

183. Rota-Stabelli, O., Lartillot, N., Philippe, H. & Pisani, D. Serine Codon-Usage Bias in Deep Phylogenomics: Pancrustacean Relationships as a Case Study. *Syst. Biol.* **62**, 121–133 (2013).

184. Regier, J. C. *et al.* Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079–1083 (2010).

185. Bernt, M. *et al.* A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. *Mol. Phylogenet. Evol.* **69**, 352–364 (2013).

186. Cameron, S. L. Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny mt: mitochondria PCG: protein-coding gene. *Annu. Rev. Entomol.* **59**, 95–117 (2014).

187. Song, F. *et al.* Capturing the phylogeny of holometabola with mitochondrial genome data and Bayesian site-heterogeneous mixture models. *Genome Biol. Evol.* **8**, 1411–1426 (2016).

188. Kamneva, Olga K., and N. A. R. Simulation-Based Evaluation of Hybridization Network Reconstruction Methods in the Presence of Incomplete Lineage Sorting. *Evol. Bioinforma.* **13**, (2017).

189. Milani, L., Ghiselli, F., Pellecchia, M., Scali, V. & Passamonti, M. Reticulate evolution in stick insects: The case of Clonopsis (Insecta Phasmida). *BMC Evol. Biol.* **10**, 1–15 (2010).

190. Wen, D., Yu, Y., Hahn, M. W. & Nakhleh, L. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol. Ecol.* **25**, 2361–2372 (2016).

191. Kagawa, K. & Takimoto, G. Hybridization can promote adaptive radiation by means of transgressive segregation. *Ecol. Lett.* **21**, 264–274 (2018).

192. Von Dohlen, C. D. & Moran, N. A. Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. *Biol. J. Linn. Soc.* **71**, 689–717 (2000).

193. Banks, J. C. & Whitfield, J. B. Dissecting the ancient rapid radiation of microgastrine wasp genera using additional nuclear genes. *Mol. Phylogenet. Evol.* **41**, 690–703 (2006).

194. Cooper, A. & Penny, D. Mass Survival of Birds Across the Cretaceous- Tertiary Boundary: Molecular Evidence. *Science (80-.).* **275**, 1109–1113 (1997).

195. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science (80-.).* **294**, 2348–2351 (2001).

196. Shaffer, H. B., Meylan, P. & Mcknight, M. L. TESTS OF TURTLE PHYLOGENY: MOLECULAR, MORPHOLOGICAL, AND PALEONTOLOGICAL APPROACHES. *Syst. Biol.* **46**, 235–268 (1997).

197. Moulton, J. K. Can the current molecular arsenal adequately track rapid divergence events within Simuliidae (Diptera)? *Mol. Phylogenet. Evol.* **27**, 45–57 (2003).

198. Zheng, Y. & Janke, A. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics* **19**, 1–19 (2018).

199. Kumar, V. *et al.* The evolutionary history of bears is characterized by gene flow across species. *Sci. Rep.* **7**, 1–10 (2017).

200. Jónsson, H. *et al.* Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18655–18660 (2014).

201. Heliconius Genome Consortium, T. & Consortium, G. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).

202. Chan, K. M. A. & Levin, S. A. Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. *Evolution* **59**, 720–729 (2005).

203. Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* **22**, 4606–4618 (2013).

204. Huang, J. P. Parapatric genetic introgression and phenotypic assimilation: testing conditions for introgression between Hercules beetles (Dynastes, Dynastinae). *Mol. Ecol.* **25**, 5513–5526 (2016).

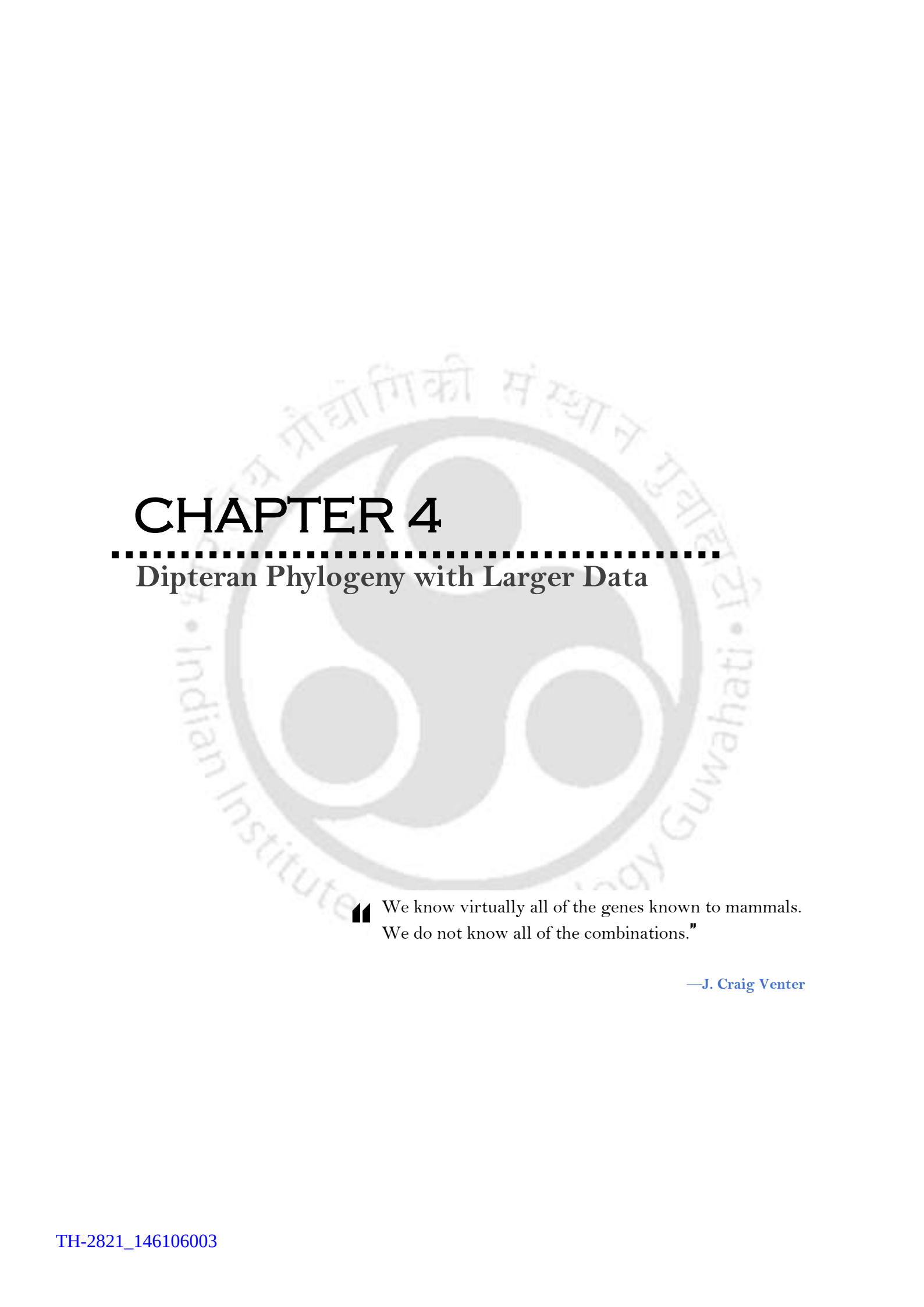
205. Keuler, R. *et al.* Genome-scale data reveal the role of hybridization in lichen-forming fungi. *Sci. Rep.* **10**, 1–14 (2020).

206. Schumer, M., Rosenthal, G. G. & Andolfatto, P. How common is homoploid hybrid speciation? *Evolution (N. Y).* **68**, 1553–1560 (2014).

207. Albach, D. *et al.* Hybridization and speciation*. *J. Evol. Biol.* **26**, 229–246 (2013).

CHAPTER 4

Dipteran Phylogeny with Larger Data



"We know virtually all of the genes known to mammals. We do not know all of the combinations."

—J. Craig Venter

TH-2821_146106003

CHAPTER 4

Reconstruction of Dipteran phylogeny and molecular dating using larger dataset (Orthologous genes) and comparative analysis

Abstract:

Diptera is one of the most successful organisms on the globe, accounting for about one-tenth of all living species. Phylogenetic relationships in species complexes and lineages derived from rapid diversifications are often difficult to resolve using morphology, regular DNA barcoding, or mitochondrial markers. A lack of strongly resolved phylogenetic relationships, currently hinders the reliable reconstruction of the underlying evolutionary processes. In this study, after discovering insect orthologous genes and conducting a basic comparative analysis, we evaluate the phylogenetic performance of 335 shared nuclear genes from 52 Diptera species using both coalescent- and concatenation-based approaches. We show that the coding regions of orthologous genes are generally conserved in size, although gene size varies with genome size. The average nucleotide identity assessment of entire orthologous Diptera genes reveals that closely related taxa share highly similar nucleotides. We also establish correlation among codon usage indices of all orthologous genes. Using both concatenation and coalescence based phylogenetic methods, we found significant conflict across the phylogeny, particularly along the backbone. This study also shows that estimation of ω = (dN/dS) and corresponding κ

(Ts/Tv) influenced by the choice of reference tree used. Using species trees and gene trees, we investigated tree space to show links between trees in a continuous, low-dimensional Euclidean space to any tree metric. Meanwhile, we deduced that the Diptera evolved between ~247 and 290 million years ago, during the Late Permian to Late Triassic period.

4.1 Introduction:

Diptera (true flies) are among the largest and most structurally and functionally diverse animal groups and no less than 1 in 10 species on Earth[1,2]. They are responsible for playing numerous roles in the planet's ecological interactions—hundreds of flies feed on plants, manage pest arthropods, break down decaying vegetation and faeces, pollinate flowers, supply food for other species, and, of course, spread diseases[3]. The limited and contradicting morphological and genetic evidence, as well as the challenge of capturing the immense species diversity in a single complete phylogenetic analysis, muddle our understanding of the evolution of flies[4]. Even well-studied species like fruit flies, mosquitoes, and house flies belong to incredibly diverse lineages that are difficult to resolve[2].

A fully resolved and well-supported phylogeny is important for understanding the evolutionary history of Diptera, and provides a foundation for research on gene function, and phenotypic evolution, and molecular dating. Diptera are an excellent example of a mega-diverse insect order, with recent explosive diversifications resulting in a bottleneck of species to identify[2,5–7]. In this species rich lineages, common DNA molecular markers are insufficient for resolving phylogenetic relationships, molecular data has been used mostly to examine generic and supra-generic relationships in this species-rich lineage, but not as regularly for analyzing species boundaries[2,8–10]. Molecular-based techniques to species demarcation may result in incorrect inferences when single or few-locus datasets are studied because of topological discordances between gene trees and species trees[11]. Over the last four decades, mitochondrial DNA has been the most widely used marker of genetic variety, owing to a combination of technical

convenience and the alleged biological and evolutionary qualities of clonality, near-neutrality, and the clock-like nature of its substitution rate[8,12–14]. However, a number of recent studies have expressed scepticism about the utility of mitochondrial DNA for evolutionary research, as mitochondrial DNA is not necessarily clonal, is far from neutrally evolving, is probably not clock-like, and low informative sites than the genomic DNA[14–16]. Many evolutionary research have recently focused on reconstructing species trees using data from hundreds or thousands of genes, which gives adequate information to overcome phylogenetic noise and ambiguity in resolving relationships[16–18].

Despite these phylogenetic breakthroughs with large-genomic data, little research has been done on the distribution of topological conflict and concordance across individual gene tree histories[19–21]. As more genomic information become available, it's essential that we begin to look into conflicting signals within gene trees, not just to better understand species trees, but also because such conflict might be a doorway into the genome's molecular evolution[22]. Furthermore, by deeper understanding the conflict within these assessments, we may be able to model the mechanisms that cause discordance more accurately. The conflict among gene trees potentially arises due to plenty of causes namely, hybridization, concealed paralogy, incomplete lineage sorting (ILS) owing to rapid radiation and/or recent divergence, lack of signal due to saturation, horizontal gene transfer, and recombination[22,23]. In the last decade there are many sophisticated tools have been developed to address the issue of gene tree/species tree reconciliation particularly in phylogenomic datasets[24,25]. There's a binning process for dealing with the combination of weak signal from individual genes, a filtering procedure for excluding low-signal genes, and a combined gene tree/species tree estimation procedure[26–28]. L. Salichos et al. devised a method for calculating the distribution of conflict across various topologies by considering only a subset of potential conflict sources, as well as efforts to accommodate numerous sources of conflict[29]. It appears that increasing the number of tested

genes does not ensure the inference of a correct species tree, because topological discordance in species/gene trees is prevalent in multi-locus inferences[30].

The first Diptera fossil discovered in France, *Grauvogelia arzvilleriana*, suggests that earlier Diptera may have lived as early as the Middle Triassic period (Lower Anisian)[31]. Moreover, multiple contemporary studies that assessed the timing of Diptera divergence imply that flies evolved either before or after the catastrophic End-Permian Extinction (EPE) shown by Montagna, Matteo, et al.[32] or after EPE described by Misof et al.[33]. However, dedicated study on Diptera by Wiegmann et al. supported the post-EPE origin of Diptera[2], but, Zhao Z. et al. shown a pre-EPE origin of Diptera[8]. Additionally, previous research suggests that the divergence time of other prominent Diptera clades are contradictory, providing plenty of potential for more investigation[2,8,13,34].

Our major goal in this work is to analyse the causes of the lack of resolution around the backbone of Diptera phylogeny, as well as compare distinct phylogenetic inconsistencies imposed by different methodologies. We sampled 335 protein-coding nuclear genes from 52 Diptera taxa, including the earliest diverging Diptera lineages and sufficient representatives of major groups of Brachycera and Nematocera. The large nuclear dataset allows us to investigate whether inadequate resolution of phylogenetic relationships across Diptera's major lineages is owing to a lack of phylogenetic signal or gene tree conflict, and if Diptera evolution is absolutely bifurcating. We applied both coalescent and concatenation -based methods to infer Diptera phylogenetic trees. We also employed tree space to compare different phylogenetic trees by mapping tree topology or branch length variability onto a low-dimensional, Euclidean space, which may then be used to show relationships across phylogenies and possibly form clusters of similar trees[35]. We also conducted molecular dating studies to infer evolutionary chronology of Diptera evolution. Additionally, we have done some comprehensive comparative analysis on entire orthologous genes.

4.2 Materials and Method:

4.2.1 Genome data acquisition and sorting:

First, we downloaded list of insect genome available in NCBI by selecting the option. Then we sorted only Diptera genome available in the list. Two midge genomes were acquired from midgebase. A total of 159 Diptera Genome applied for annotation and orthologous gene annotation. *Anoplophora glabripennis*, *Manduca sexta* genomes were also downloaded as outgroup for this analysis.

4.2.2 Orthologus gene (OG) identification and average nucleotide identity (ANI):

We used BUSCO's workflow to build our phylogenomic dataset, with the goal of rigorously selecting a set of single copy orthologous genes while simultaneously minimizing the amount of missing data in the dataset[36]. To identify the Orthologus genes (OGs) and assess the completeness of the retrieved genomes, we used OrthoDB's insecta_odb9 dataset of 1658 single-copy orthologs of insect lineage by applying the genome mode[37]. In the initial phase of BUSCO pipeline applies TBLASTN to identify a subset of sequences in the genome using amino acid consensus sequences as queries; then AUGUSTUS creates the precise gene structures on these regions to derive the protein sequence, and HMMER is used to assign a score to the candidate amino acid sequence[38–40]. The following phase of the BUSCO genome mode comprises a retraining process, which results in improved set of parameters for AUGUSTUS based on the single-copy BUSCO genes identified as complete during the first phase. The final phase focuses on identifying the missing BUSCO genes by TBLASTN utilizing additional variants of the amino acid consensus, followed by an AUGUSTUS step to apply the retrained parameters and a further HMMER run to generate a final classification[36,41]. The average nucleotide identity (ANI) of all identified OGs of 54 species was estimated using pyani software and the heatmap of ANI matrix was generated using superheat package in R [42,43]. For phylogenomic study we retained genomes with at least 80% completeness and at least

1000 single copy orthologs for further investigation. We also reduced the number of genomes from families with abundant genomes (e.g., Culicidae and Drosophilidae), resulting in a final set of 335 single copy orthologs shared by 52 Diptera genomes (Table 4.1).

4.2.3 Calculation of Codon usage indices of all orthologous genes:

The degeneracy of the genetic code allows for multiple codons to encode the same amino acid. However, degenerate codons are not present at equal frequencies in genes, a phenomenon termed codon-usage bias (CUB) [44]. The measures of CUB are known under a broad term of codon indices and different kind of such indices had emerged to capture different biological phenomenon. Here we estimated few such codon indices for all orthologous protein coding genes identified by BUSCO. The software BioCUA was used for calculation of GC, GC3 and frequency of optimal codons (fop) [45]. The sequence length, codon adaptation index (CAI), effective number of codons (ENc) were calculated using DAMBE[46]. The translation adaptation index (tAI) was measured through codonR (https://github.com/santiago1234/codonr).

4.2.4 Orthologus gene alignment:

The resulting set of 335 orthologs were aligned using MAFFT algorithm subjected for TranslatorX guided by amino acid alignment which shows that some of the orthologs among 335 orthologs persist internal stop codons [47,48]. To overcome this issue, we aligned 335 orthologs fasta files by MACSE software [49]. MACSE further allows to align orthologs by eliminating internal and external stop codons. Now the resultant CDS were again aligned through amino acid guided MAFFT algorithm using stringent GBLOCK for filtering gaps and ambiguously aligned sites.

Table 4.1: List of Diptera (n=52) genomes used in this study for comparative genomics and phylogenetic analysis.

Species Name	Size (Mb)	Cove -rage	Releas e Year	Sub-order	Infraorder	Family	Assembly	Assembly method	Sequencing technology	Reference
Bactrocera latifrons	462.505	101.0	2016	Brachycera	Muscomorpha	Tephritidae	GCA_001853355.1	AllPaths v. r44837	Illumina HiSeq	
Bactrocera oleae	471.780	50	2015	Brachycera	Muscomorpha	Tephritidae	GCA_001188975.2	Ray + SSPACE + gap closing with PacBio v. 2014-2015	PacBio; Illumina HiSeq	50
Bactrocera tryoni	519.006	96.0	2014	Brachycera	Muscomorpha	Tephritidae	GCA_000695345.1	ABySS v. AUG-2013; SSPACE v. AUG-2013	Illumina HiSeq; Illumina GAII; Illumina MiSeq; 454	51
Ceratitis capitata	436.491	152.5	2013	Brachycera	Muscomorpha	Tephritidae	GCA_000347755.4	AllPaths v. 35218; ATLAS-link v. 1.0; ATLAS-gapfill v. 2.2; redundans v. 0.12c	Illumina HiSeq	52
Drosophila arizonae	141.387	52.0	2016	Brachycera	Muscomorpha	Drosophilidae	GCA_001654025.1	AllPaths v. R48777	Illumina HiSeq	53
Drosophila athabasca	193.536	45.0	2019	Brachycera	Muscomorpha	Drosophilidae	GCA_008121225.1	canu v. 1.6	PacBio	54
Drosophila biarmipes	169.379	186.9	2011	Brachycera	Muscomorpha	Drosophilidae	GCA_000233415.2	Celera Assembler v. 6.1; BWA v. 0.6.0; Samtools v. 0.1.14; GATK v. 1.1-9	Illumina; 454	55
Drosophila busckii	135.749	172.9	2015	Brachycera	Muscomorpha	Drosophilidae	GCA_001277935.1	ALLPATHs-LG r44206	Illumina	56
Drosophila elegans	171.268	204.9	2011	Brachycera	Muscomorpha	Drosophilidae	GCA_000224195.2	Celera Assembler v. 6.1; BWA v. 0.6.0; Samtools v. 0.1.14; GATKv. 1.1-9	454; Illumina	55
Drosophila melanogaster	144.125	147.0	2017	Brachycera	Muscomorpha	Drosophilidae	GCA_002300595.1	WGS v. 8.3rc1; DBG2OLC v. 1	PacBio	57
Drosophila montana	183.585	40.0	2018	Brachycera	Muscomorpha	Drosophilidae	GCA_003086615.1	CLC NGS Cell v. May-2012	Illumina HiSeq; Illumina MiSeq	58
Drosophila nasuta	137.224	55.0	2017	Brachycera	Muscomorpha	Drosophilidae	GCA_002222885.1	CLC Genomics Workbench v. 6.0	Illumina NextSeq 500	59
Drosophila obscura	181.869	425	2017	Brachycera	Muscomorpha	Drosophilidae	GCA_002217835.1	ALLPATHS-LG v. 50927	HiSeq2000	60
Scaptodrosophila lebanonensis	247.078	100.0	2018	Brachycera	Muscomorpha	Drosophilidae	GCA_003285725.2	CANU v. 1.5	PacBio Sequel; Illumina	61
Scaptomyza flava	214.837	90.0	2018	Brachycera	Muscomorpha	Drosophilidae	GCA_003952975.1	AllPaths v. MARCH-2013	Illumina HiSeq	62
Ephydra gracilis	410.873	9.4	2015	Brachycera	Muscomorpha	Ephydridae	GCA_001014675.1	SOAPdenovo v. 1.05	Illumina	61

TH-2821_146106003

Glossina austeni	370.265	62	2014	Brachycera	Muscomorpha	Glossinidae	GCA_000688735.1	ALLPATHS-LG v. August 2013	Illumina	63
Glossina brevipalpis	315.360	81	2014	Brachycera	Muscomorpha	Glossinidae	GCA_000671755.1	ALLPATHS-LG v. August 2014	Illumina	63
Glossina fuscipes	374.775	120	2014	Brachycera	Muscomorpha	Glossinidae	GCA_000671735.1	ALLPATHS-LG v. August 2015	Illumina	63
Glossina morsitans	348.063	13.9	2015	Brachycera	Muscomorpha	Glossinidae	GCA_001014515.1	SOAPdenovo v. 1.05	Illumina	61
Glossina pallidipes	357.332	78	2014	Brachycera	Muscomorpha	Glossinidae	GCA_000688715.1	ALLPATHS-LG v. August 2013	Illumina	63
Glossina palpalis gambiensis	380.104	138	2015	Brachycera	Muscomorpha	Glossinidae	GCA_000818775.1	ALLPATHS-LG November 2014	Illumina	63
Lucilia cuprina	378.290	158.3	2014	Brachycera	Muscomorpha	Calliphoridae	GCA_000699065.2	AllPaths LG v. 44620; Atlas Link v. 1.0; Atlas GapFill v. 2.2; redundans v. 0.12c	Illumina	64
Phormia regina	549.933	44.0	2016	Brachycera	Muscomorpha	Calliphoridae	GCA_001735545.1	CLC Genomics Workbench v. 6.0.5	Illumina HiSeq; PacBio	65
Paykullia maculata	422.395	96.0	2018	Brachycera	Muscomorpha	Rhinophoridae	GCA_003055125.1	Platanus v. 1.2.4	Illumina HiSeq	66
Proctacanthus coquilletti	208.908	132.0	2017	Brachycera	Muscomorpha	Asilidae	GCA_001932985.1	Discovar v. MAR-2015; w2rap-contigger v. MAR-2016	Illumina HiSeq	67
Stomoxys calcitrans	971.189	66	2015	Brachycera	Muscomorpha	Muscidae	GCA_001015335.1	ALLPATHS-LG v. April 2015	Illumina	68
Zaprionus indianus	123.675	36.0	2016	Brachycera	Muscomorpha	Drosophilidae	GCA_001752445.1	CLC Genomics Workbench v. 6.0	Illumina NextSeq	69
Zeugodacus cucurbitae	374.820	66.2	2014	Brachycera	Muscomorpha	Tephritidae	GCA_000806345.1	AllPaths v. 49856	Illumina HiSeq	
Aedes aegypti	1278.730	110.0	2017	Nematocera	Culicomorpha	Culicidae	GCA_002204515.1	Falcon_UNZIP v. 0.7.0	PacBio	70
Anopheles aquasalis	162.944	340.0	2017	Nematocera	Culicomorpha	Culicidae	GCA_002846955.1	SOAPdenovo v. February-2014	Illumina HiSeq	
Anopheles arabiensis	246.568	100.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000349185.1	allpaths v. R43436	Illumina	71
Anopheles atroparvus	224.290	98.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000473505.1	allpaths v. R46504	Illumina	71
Anopheles christyi	172.659	35.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000349165.1	allpaths v. R44024	Illumina	71
Anopheles culicifacies	202.999	31.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000473375.1	allpaths v. R46449	Illumina	71

Anopheles dirus	216.308	184.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000349145.1	allpaths v. R43500	Illumina	71
Anopheles epiroticus	223.487	49.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000349105.1	allpaths v. R43500	Illumina	71
Anopheles farauti	183.103	233.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000473445.2	allpaths v. R47616	Illumina	71
Anopheles funestus	444.544	234.0	2018	Nematocera	Culicomorpha	Culicidae	GCA_003951495.1	Canu v. 1.3	PacBio RSII	72
Anopheles merus	288.049	147.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000473845.2	allpaths v. R47616	Illumina	71
Anopheles minimus	201.793	211.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000349025.1	allpaths v. R43460	Illumina	71
Anopheles quadriannulatus	283.829	93.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000349065.1	allpaths v. R43436	Illumina	71
Anopheles sinensis	375.764	86.0	2013	Nematocera	Culicomorpha	Culicidae	GCA_000472065.2	allpaths v. R47616	Illumina	73
Mochlonyx cinctipes	441.264	12.9	2015	Nematocera	Culicomorpha	Chaoboridae	GCA_001014845.1	SOAPdenovo v. 1.05	Illumina	61
Belgica antarctica	89.584	150.0	2014	Nematocera	Culicomorpha	Chironomidae	GCA_000775305.1	Velvet v. 1.1.05; ERANGE v. June 7, 2011; PBJelly v. Jelly 14.1.14	Illumina; PacBio	74
Chironomus riparius	154.534	17.5	2015	Nematocera	Culicomorpha	Chironomidae	GCA_001014505.1	SOAPdenovo v. 1.05	Illumina	61
Chironomus tentans	213.463	50	2014	Nematocera	Culicomorpha	Chironomidae	GCA_000786525.1	CLC	Illumina paired-end (PE), Illumina 5 kb mate-pair (MP), and 454 single-end (SE)	75
Clunio marinus	85.491	140	2016	Nematocera	Culicomorpha	Chironomidae	GCA_900005825.1	Velvet	Illumina HiSeq2000	76
Polypedilum nubifer	118.172	58	2013	Nematocera	Culicomorpha	Chironomidae	http://bertone.nises-f.affrc.go.jp/midgebase/	Platanus	Illumina and SOLiD4	77
Polypedilum vanderplanki	133.777	562	2013	Nematocera	Culicomorpha	Chironomidae	http://bertone.nises-f.affrc.go.jp/midgebase/	Platanus	Illumina and SOLiD4	77
Clogmia albipunctata	256.249	17.8	2015	Nematocera	Psychodo-morpha	Psychodidae	GCA_001014945.1	SOAPdenovo v. 1.05	Illumina	61
Coboldia fuscipes	98.759	53.0	2015	Nematocera	Psychodo-morpha	Scatopsidae	GCA_001014335.1	SOAPdenovo v. 1.05	Illumina	61

4.2.5 Data partitioning and phylogenetic analysis:

The concatenated sequence of 335 gene was first partitioned into gene wise three codon position (335*3). Then PartitionFinder was used for assigning model for different partitions using rcluster algorithm [78,79].

Phylogenetic analysis was performed using two different concatenation and coalescent based method. In data concatenation analysis Maximum likelihood (RaxML[80], IQTree[81]) and Bayesian inference (ExaBayes[82], PhyloBayes [83]) were used for inferring phylogenetic relationship. In coalescent based method ASTRAL and MP-EST software were used [84,85].

In RAXML, GTR model was used for all subset of partitions and run for 1000 generation. We also performed ML analysis using Edge-Linked-proportional and the Edge-Unlinked model. Bayesian phylogenetic inference were conducted using ExaBayes and PhyloBayes. ExaBayes was run using number of runs 2, number Coupled Chains 2, Number of generations $2e^6$ (2000000) mentioning DNA as the data type on 1005 partitions. The PhyloBayes analysis was done under site-heterogeneous CAT-GTR G4 model for each nucleotide partition on 2 parallel chains. In this mixed model analysis, all model parameters including branch length were unlinked between partitions.

For coalescent based method 335 gene trees were first created in RaxML using GTR model and 1000 bootstrap replicates. ASTRAL and MP-EST method was used to build species tree from the gene trees.

4.2.6 Gene and species tree concordance/discordance:

We evaluated the topological concordance of gene trees by mapping reference species trees generated by previous analysis (RaxML, IQtree edge proportional, IQtree edge unlinked, ExaBayes, PhyloBayes, ASTRAL, MP-EST) with the program PhyParts (https://bitbucket.org/blackrim/phyparts)[22]. Trees were previously rooted in R with the package

APE and *Manduca sexta* as outgroup. The output obtained with Phyparts was visualized by plotting pie charts on the all-coalescent species tree and concatenated tree with the script PhyPartsPieCharts (https://github.com/mossmatters/MJPythonNotebooks) using the ETE3 Python toolkit [86].

Internode Certainty All (ICA): Internal edge certainty (ICA) metric that considers the frequency of all conflicting bipartitions to calculate the degree of certainty for specific focal bipartitions (internal edges)[29]. This is calculated for each internal edge, i, as.

$$ICAi = 1 + \sum_{n=1}^{b} P(Xn) logb[P(Xn)]$$

where, $P(Xn)$ is the proportional frequency of bipartition n in the set of bipartitions being studied, and b is the number of unique conflicting bipartitions (including the bipartition of interest *i*). ICA values near 0 imply maximum conflict (i.e., conflicting bipartitions occur often), whereas values near 1 suggest strong certainty in the bipartition of interest[22]. Here Phyparts was used estimate the ICA score to describe the conflict and certainty in the bipartitions of the species trees[22].

4.2.7 Substitution rate calculation on reference tree:

In order to calculate substitution rate ω for all 335 aligned genes, we used basic model (M0) using seven previously created species trees as reference tree using fixed branch length. The M0 model will provide single dN/dS (Nonsynonymous substitution/Synonymous substitution) and Ts/Tv (Transition/Transversion) for each gene. This analysis was performed in PAML using Param-Ishan HPC [87]. R package ggplot2 was used to plot dN/dS and Ts/Tv for each phylogenetic tree and calculating mean of dN/dS and Ts/Tv for each phylogenetic tree. Further we also plot correlation of dN/dS and Ts/Tv.

4.2.8 Molecular dating and divergence time estimation:

We estimated divergence times for our Diptera-wide, 335-gene dataset using Bayesian inference in MCMCTree. For analysis using MCMCTree, we used ML approximation by first calculating the ML estimates of the branch lengths, the gradient vector and Hessian matrix, using BaseML and CodeML programs in PAML[87]. The uncorrelated relaxed clock model (clock = 2) was used and the prior on the root age set to the first appearance of the winged insect at 325 Mya in Carboniferous. The dating estimations were conducted by running 2 independent MCMC chains were run with following parameters: number of samples = 20000; sampling frequency = 1000; burn in = 20000. 19980000 samples were then summarized to estimate mean divergence date and 95% credibility intervals.

4.2.9 Exploration of phylogenetic tree space:

We assessed the topological harmony among gene trees and species trees through supermatrix approach with the R package TreeSpace v. 1.10.19 [35]. First the species trees were compared according to their distance based on four distinct metrics namely, Robinson-Foulds (RF) symmetric difference [88], Kendall-Colijn metric [89], branch score distance [90], Abouheif's dissimilarity [91]. Then derived distance metrices of species trees were visualized through heatmap in superheat package of R [42]. We also calculated and compared mean of MRCA tree depth of different species trees. Further we identified clusters of similar trees with Metric Multidimensional Scaling (MDS) based on aforementioned distance methods.

4.3 Result and Discussion:

4.3.1 Assessment of the completeness of the candidate Genomes:

The BUSCO statistics of 52 candidate Diptera genomes are displayed in bar graphs to emphasize different standards of assembly and annotation work deposited in public databases (Fig. 4.1). Herein, it reveals the amount of core insect OGs present in any Diptera genomes. According the statistics, the 35 Diptera genomes (67%) out of 52 included in the study had

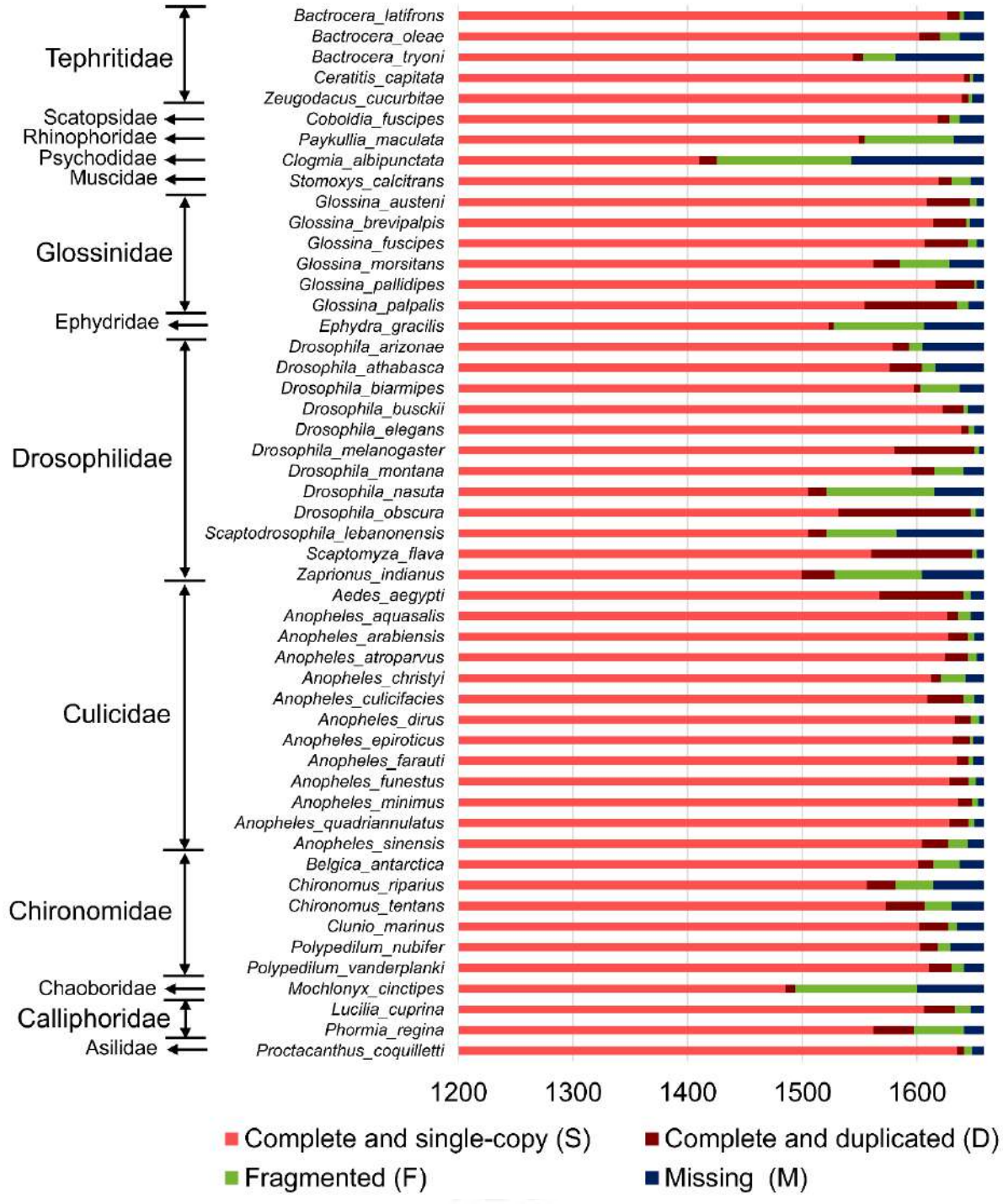


Figure 4.1: BUSCO completeness assessments for genomics data of 52 Diptera collected from different databases (mainly NCBI); bar charts show number of classified Orthologous genes (OGs) presented as complete and single-copy (S, red), complete duplicated (D, brown), fragmented (F, green), and missing (M, dark blue)

more than 95% complete and single copy of the core insect OGs. In this study, *Ceratitis capitata* has 1641 OGs, which covers 99% of all core set of insect OGs (highest), whereas *Clogmia albipunctata* has 1410 OGs, which covers only ~85% core set of insect OGs (lowest).

4.3.2 Comparison among the size of Genome, OGs and CDS:

In this section the genome size was documented from NCBI (https://www.ncbi.nlm.nih.gov/) and the size of gene and coding sequence were derived from .gff file of BUSCO analysis. A comparison of complete genome size, size of orthologous genes (OGs) and size of coding

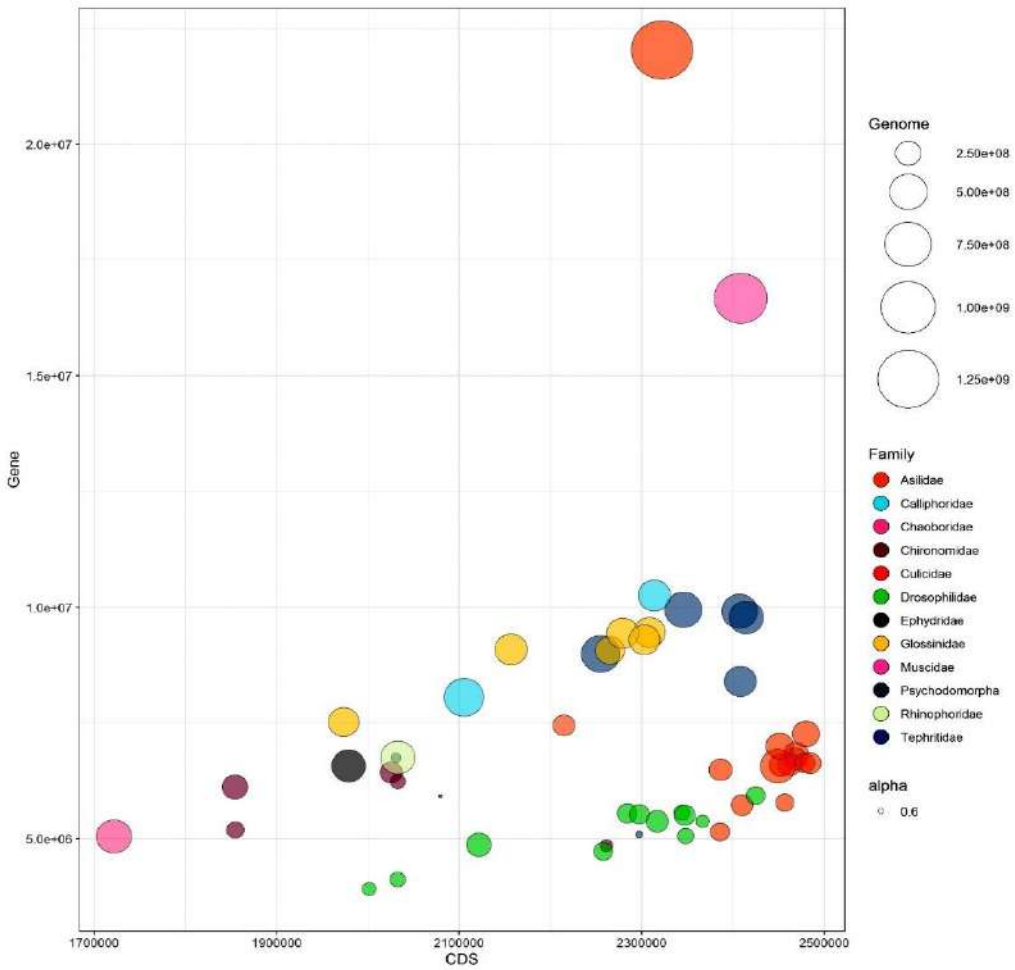


Figure 4.2: Correlation among the size of Genome, Orthologous Genes and Coding Sequence. The circle size defines the size of the genome, y-axis: Gene size and x-axis: size of coding sequence (CDS) and colour of the circle denote different families included in this study

sequences (CDS) depicted in Figure 4.2. It reveals that *Aedes aegypti* (red colour top) has the largest genome (1.28e^{09} bp) among the compared species which is due to presence of large number transposons[92]. Simultaneously, the overall size of OGs in Aedes aegypti is the largest among other species, while the total size of CDS is within the range of other species, and CDS accounts for 0.18% of total genome (Fig. 4.2). Similarly, for *Stomoxys calcitrans* (pink colour top) the CDS size (0.24% of the genome) not as high as genome of gene size. *Belgica antarctica*'s CDS accounts for 2.56% of the whole genome, the highest in this study. *Clunio marinus* has the smallest genome (8.55e^{07} bp) and CDS accounts for 2.43% of the whole genome. The length of the CDS ranges from 1.72e^{06} bp in *Mochlonyx cinctipes* to 2.48e^{06} bp in *Anopheles minimus*. Whereas, entire OGs length varies from 2.20e^{07} bp in *A. aegypti* to 3.91e^{06} bp in *Zaprionus indianus*. It implies that the size of coding sequences is more or less conserved among species than gene or genome.

4.3.3 Average nucleotide identity (ANI) of OGs:

Average nucleotide identity (ANI) of core orthologous protein coding sequence analysed and presented in Figure 4.3. The heatmap of ANI show that similar group of species have similarity in protein coding sequence and they created blocks with lighter colour shed (Fig. 4.3). This investigation revealed that *Anopheles arabiensis* (Aara) has ~98% nucleotide identical with *Anopheles merus* (Amer) and *Anopheles quadriannulatus* (Aqua) in the Culicidae family, while Aqua likewise shares ~98% nucleotide identity with Amer. With the exception of *Glossina brevipalpis*, all other species in the Glossinidae family have >95% nucleotide identity with each other; moreover, OGs of *Glossina fuscipes* (Gfus) are >99% identical with *Glossina palpalis gambiensis* (Gpalg).

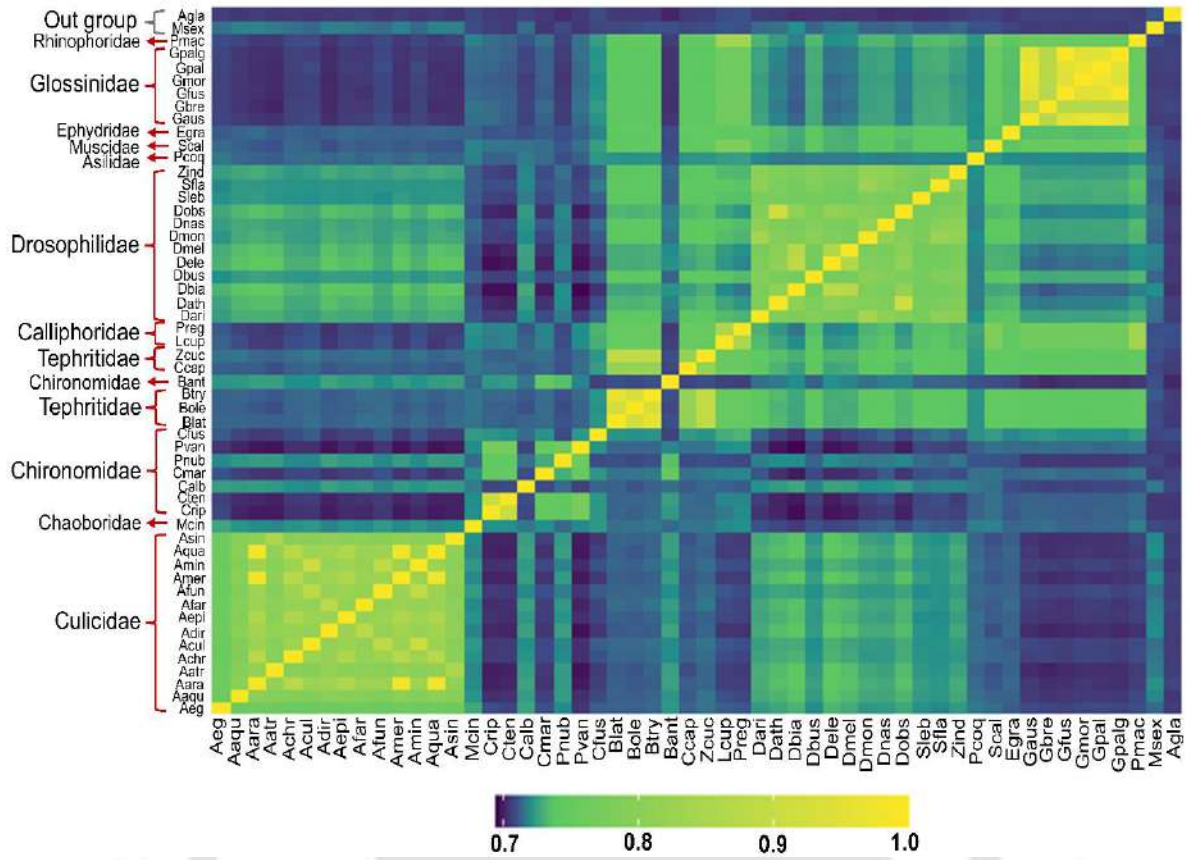


Figure 4.3: A heat map of Average nucleotide identity (ANI) of Orthologous genes of 52 Diptera and 2 outgroups. The coloured bar represents the species as indicated in supported ANI values shown here. ANI is depicted as the colour gradient indicated by the legend: lighter = 1 (100% ANI), darker = 0.7 (70% ANI)

4.3.4 Correlation among codon usage indices:

A single amino acid can be encoded by more than one synonymous codon, and synonymous codons are employed in an uneven manner. Some codons, in particular, are utilized more frequently than their synonymous counterparts in highly expressed genes[93]. Several measures of codon usage bias have been developed to assess the unevenness of codon usage namely, Effective number codons (ENc)[94], codon adaptation index (CAI)[93], translation adaptation index (tAI)[95], frequency of optimal codons (FOp)[96], GC, GC content at 3rd codon position (GC3). Here in this section, we calculated different codon usage indices of all orthologous genes of 52 Diptera species. The correlation matrix of all codon usage indices is depicted in Figure 4.4. The correlation among different codon usage indices shows that ENc is significantly negatively

correlated with other codon indices like CAI, GC, GC3, FOp, tAI. This suggests that when the GC, GC3, FOp, tAI, and CAI of any gene increase, the ENc value decreases, implying an increase in codon usage bias of that particular gene. Whereas, other codon usage indices are significantly positively correlated with each other. Therefore, it means that increasing the proportion of GC and GC3 of any gene the CAI, tAI and FOp increases whereas effective number of codons (ENc) decreases. This indicates that high GC content in any gene is associated with an increase in codon usage bias, an increase in the quantity of GC biased codons, an increase in gene expression, as well as an enhancement in translational efficiency. Except for FOp, the sequence length of any gene is positively correlated with codon usage indices, however these correlations are non-significant.

4.3.5 Phylogenetic analyses based on 335 nuclear genes:

We applied both coalescent and concatenation methods to reconstruct the Diptera phylogeny using the 335 gene dataset (Fig. 4.5). Our phylogenomic analyses recovered full support for Diptera as monophyletic group sister to well distinguished outgroup (*Manduca sexta* (Lepidoptera) and *Anoplophora glabripennis* (Coleoptera)). Major families of Diptera such as Chironomidae, Culicidae, Glossinidae, Tephritidae, and Drosophilidae were recovered as monophyletic groups with maximum support. However, the relationships between Calliphoridae and Rhinophoridae family less robustly resolved as our analysis unable recover Calliphoridae as monophyletic group. *Phormia regina*, a Calliphoridae fly clustered with Rhinophoridae fly, *Paykullia maculate* and other Calliphoridae fly, *Lucilia cuprina* paraphyletic to *Phormia regina*. Although, the node between *Phormia regina* and *Paykullia maculate* got no bootstrap Support (bs = 0) from RaxML analysis and relatively low posterior probability from ASTRAL analysis (pp = 0.92). The position of a *Coboldia fuscipes* not well resolved, as it is conventionally a Nematocera fly belong to Scatopsidae family and Psychodomorpha infraorder situated as sister to all Brachycera group. While another

Psychodomorpha fly, *Clogmia albipunctata* located as sister to all Diptera by RaxML and ASTRAL analysis. In PhyloBayes analysis both the species situated as sister to all Brachycera group with maximum posterior probability (pp = 1). That might be an impediment to enough support in the phylogeny's backbone node, which separates Brachycera from Nematocera. RaxML analysis found relatively low support (bs = 98) and ASTRAL analysis found weakest posterior probability (pp = 0.78) in this node.

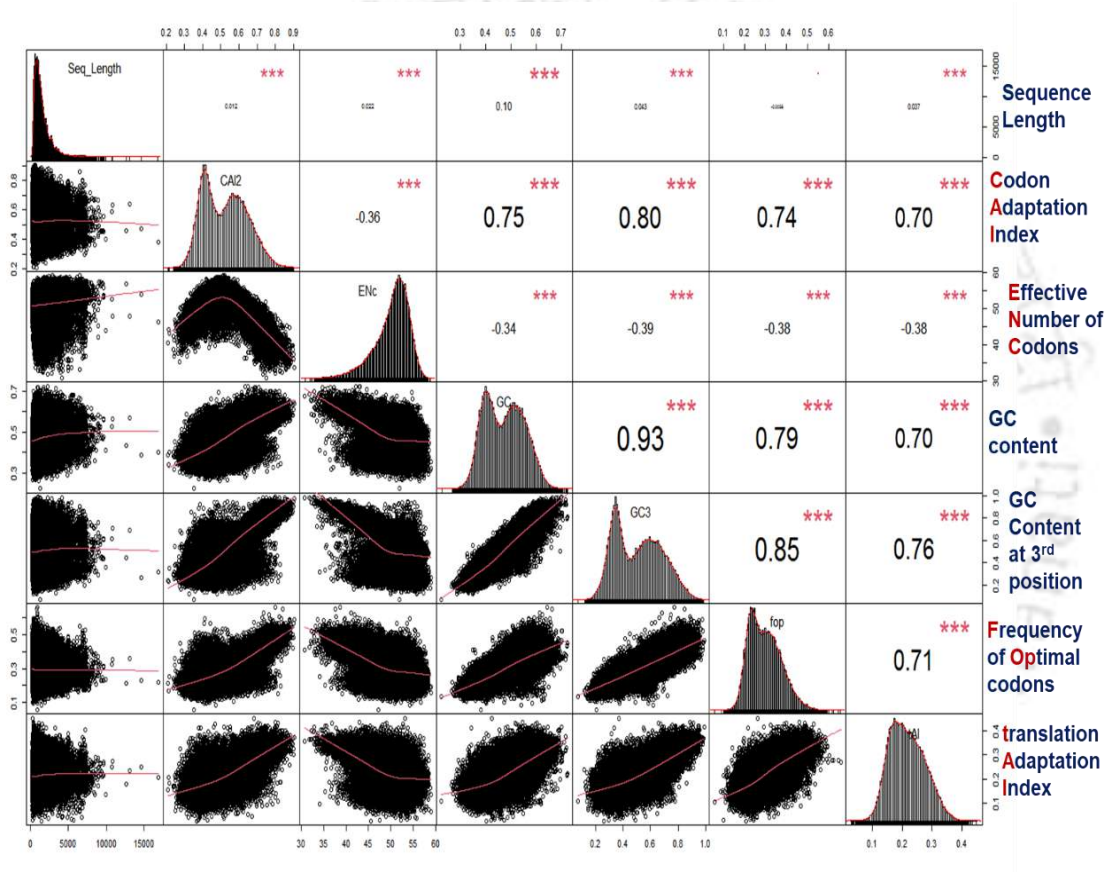


Figure 4.4: Correlation matrix among different codon usage indices of all orthologous genes of 52 Diptera species.

4.3.6 Evaluation of Gene Tree Conflict:

We observed certain nodes in our phylogenomic analysis that were not adequately confirmed by bootstrap support or posterior probability. To understand further how much evolution of species supported by the genes we calculated topological concordance and discordance

between 335 gene trees and the species tree. The conflicts between gene trees and species tree were prevalent at the nodes along the backbone of the Diptera phylogeny (Fig. 4.6). The number of homolog groups concordant with each clade in the species tree varied significantly. Specially, the node connecting *C. fuscipes* with Brachycera displayed only 55 concordant homologs supporting all three species tree topologies (RaxML, PhyloBayes and ASTRAL).

Our result revealed that 43 homologs in the node linking Nematocera and Brachycera coincide with RaxML and ASTRAL species trees. Whereas, only 26 homologs in the PhyloBayes species tree agreed with the species tree in linking node Nematocera to Brachycera, where *C. albipunctata* was also positioned as sister to all Brachycera lineages, with *C. fuscipes*. Another place where *P. regina*, a Calliphoridae fly clustered with Rhinophoridae fly, *P. maculate* we found only 110 concordant homologs supporting all three species tree topologies. The node connecting Culicidae and Chironomidae where a Chaoboridae fly, *Mochlonyx cinctipes* positioned as sister to Culicidae showed low support only by 83 concordant homologs. The node connecting Tephritidae to Drosophilidae *Ephydra gracilis* situated as sister to Drosophilidae lineages displayed 83 concordant homologs in concatenation-based species trees RaxML and PhyloBayes. At contrast, a coalescence-based species tree placed Drosophilidae as the stem lineage with a sufficient number of concordant homologs (239), but only 55 concordant homologs in the node Tephritidae linked to Glossinidae. Aside from that, for all three species trees, four nested nodes within the Drosophilidae family had a low number of concordant homologs (109, 54, 98, and 72) (Fig. 4.6).

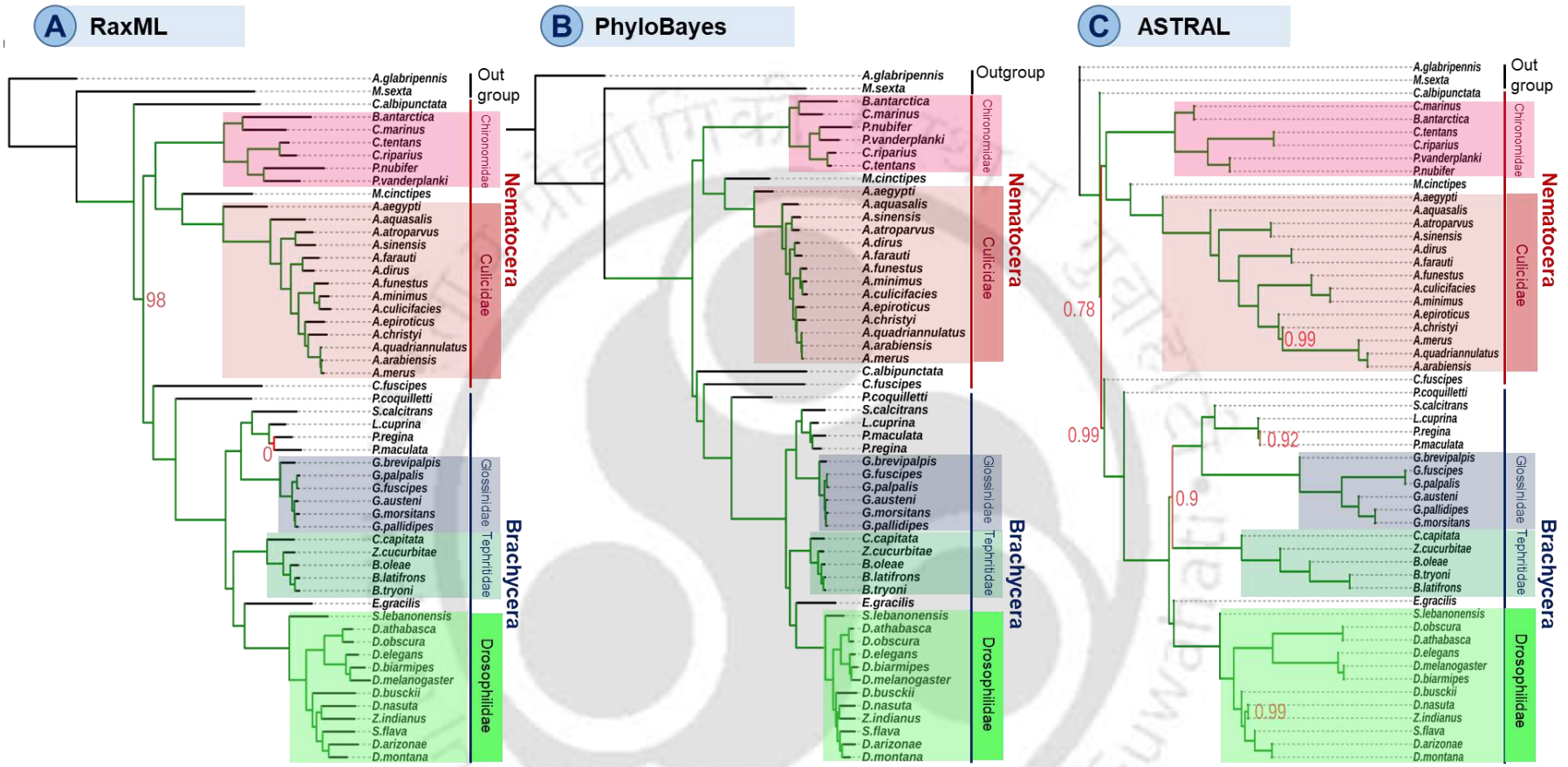


Figure 4.5: The Species Trees Inferred from the Dataset of 335 Nuclear Genes. (A) The concatenation-based species tree inferred by RaxML (Maximum-likelihood). Colour of the branches indicate the bootstrap support (bs); green: maximum (100%) bs, red: minimum bs. Number in red colour associated with node denote < 100% bs. (B) The concatenation-based species tree inferred by PhyloBayes (Bayesian inference). Colour of the branches indicate the posterior probability (pp); green: maximum pp (C) The coalescent-based species tree was inferred by ASTRAL. Colour of the branches indicate the posterior probability (pp); green: maximum pp, red: minimum pp. Number in red colour associated with node denote < 1 pp.

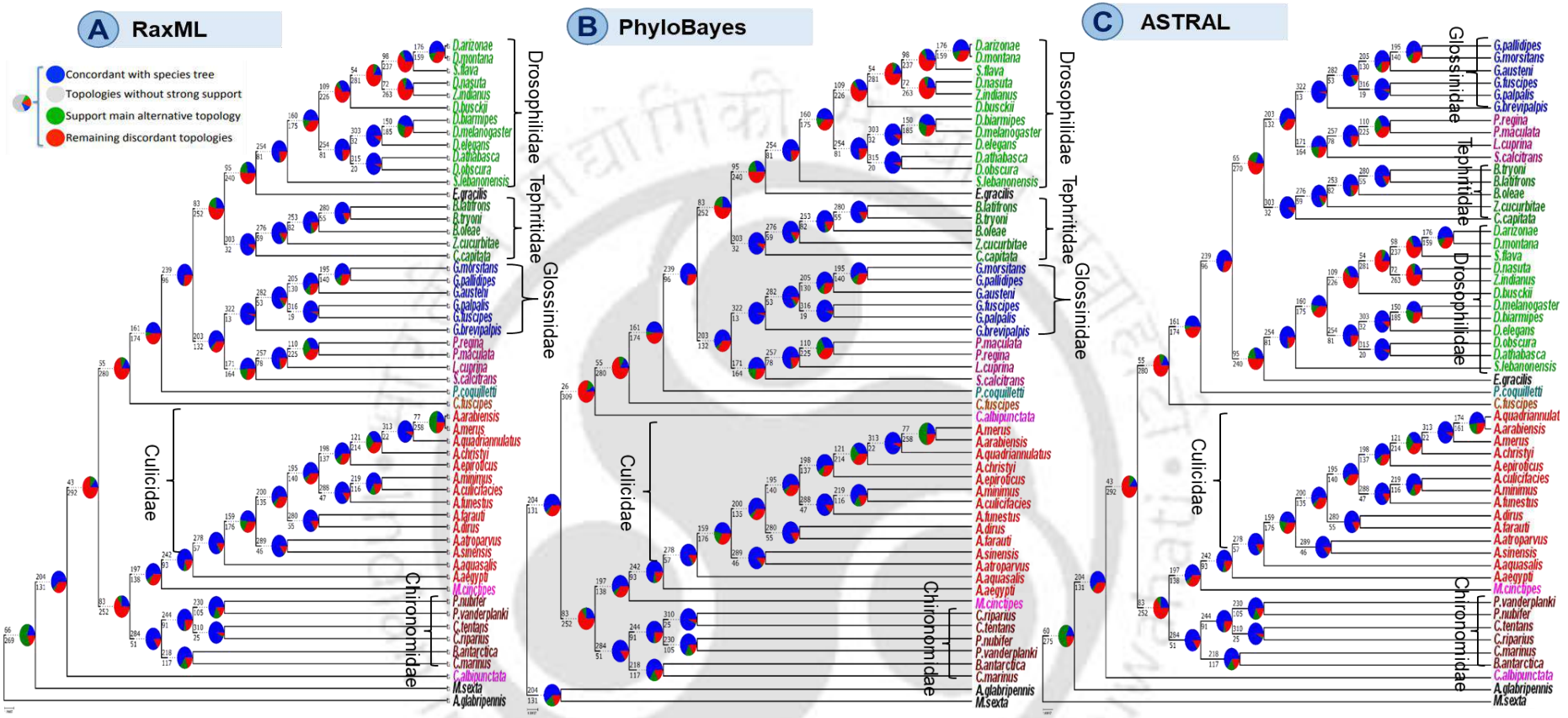


Figure 4.6: Summary of conflicting and concordant homologs. The species trees are generated by RaxML(A), PhyloBayes (B), and ASTRAL (C). For each branch, the top number indicates the number of homologs concordant with the species tree at that node, and the bottom number indicates the number of homologs in conflict with that clade in the species tree. The pie charts at each node present the proportion of homologs that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion that inform (conflict or support) this clade that have less than 50% bootstrap support (grey).

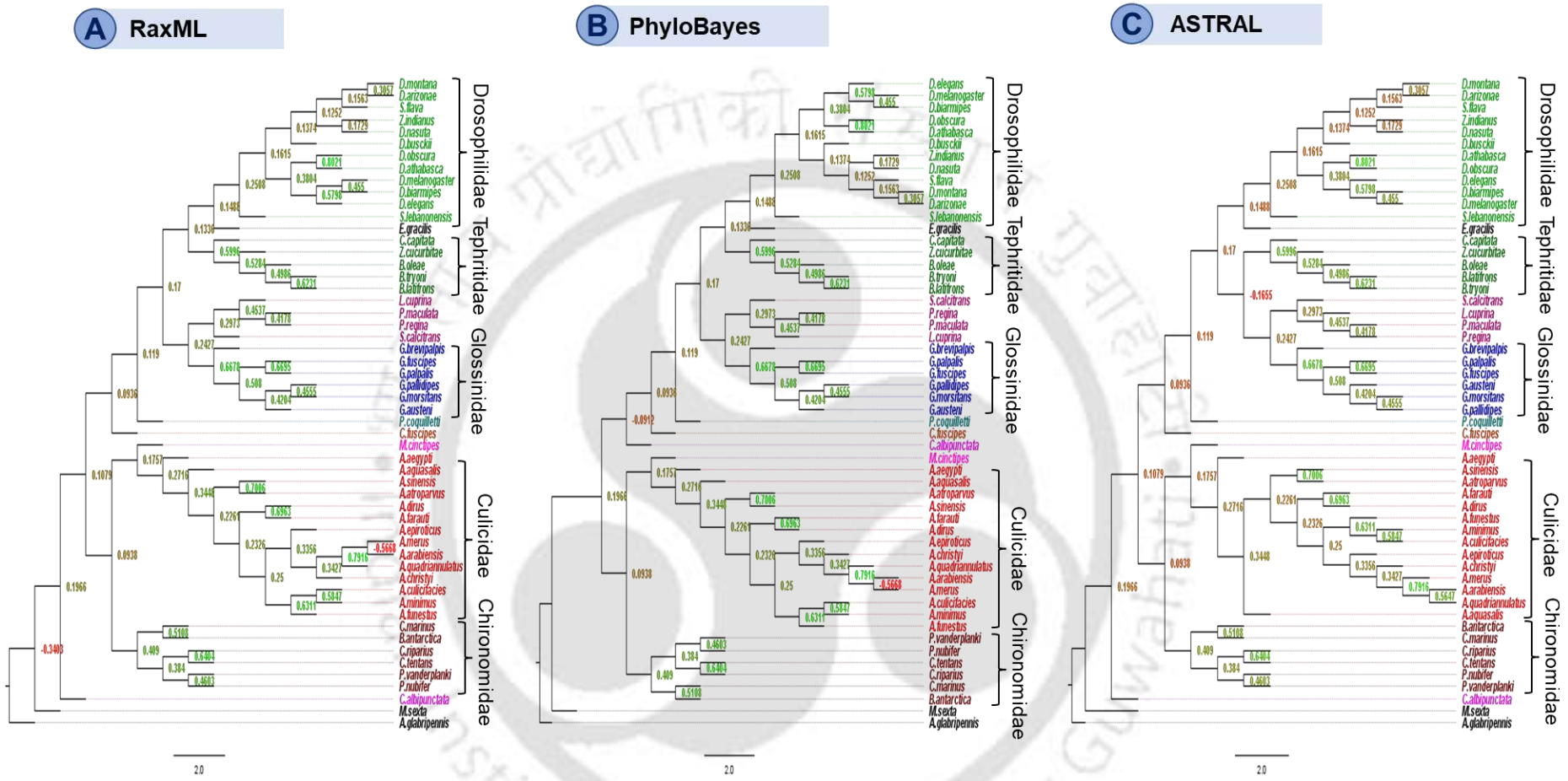


Figure 4.7: Representation of Internode Certainty All (ICA) score of the nodes in the species trees. The species trees are generated by RaxML(A), PhyloBayes (B), and ASTRAL (C). The numbers associated with each node is the ICA score and colour of numbers denotes the strength of internode certainty; green: high ICA (strong certainty) and red: low ICA (high conflict).

We calculated ICA (Internode Certainty All) scores on the species tree given the set of homolog trees (gene trees). The estimated ICA scores varied significantly across the nodes in all three represented tree topologies (Fig. 4.7). For the concatenation-based Diptera species tree RaxML and PhyloBayes results ICA values ranged from -0.5668 to 0.8021, while coalescence-based species tree ASTRAL results ICA values -0.1655 to 0.8021. The node connecting *C. fuscipes* with Brachycera displayed low ICA score, 0.0936 in all three species trees, RaxML, PhyloBayes and ASTRAL. At the node in PhyloBayes tree where *C. albipuctata* was placed as sister to *C. fuscipes* and other Brachycera lineages, the ICA was negative (-0.0912). On the other hand, *C. albipuctata* was found to be sister to all other Diptera in RaxML and the ASTRAL tree, with a score of 0.1966. The node linking between Nematocera and Brachycera also displayed relatively low ICA, 0.1079 in RaxML and ASTRAL species tree while PhyloBayes species tree showed relatively better ICA score (0.1966). The node connecting Glossinidae and Tephritidae where the cluster between Calliphoridae and Rhinophoridae located as sister to Glossinidae displayed negative ICA (-0.1655) in coalescence-based ASTRAL species tree whereas, in both concatenation-based species trees that node doesn't exist. The node linking Culicidae and Chironomidae also showed low ICA score, 0.0938. Overall, the results revealed that ICA values were lower along the backbone of the tree, while ICA values were higher in many of the nested clades (Fig. 4.7).

4.3.7 Species tree-wise substitution rate:

In this section, we used the M0 model to calculate omega (ω = dN/dS) and kappa (κ = Ts/Tv) values for 335 genes based on species trees, where only a single value of ω and κ value was obtained for each gene per species tree. The ω defines as the ratio between non-synonymous substitution rate (dN) and synonymous substitution rate (dS); while κ defines as the ratio between transition (Ts) and Transversion (Tv). The result shows that the ω and kappa both values dependent on the phylogenetic trees (Fig. 4.8). The outcome of ASTRAL, MP-EST and

ExaBayes trees show very low mean of ω (ASTRAL: 0.0222, MP-EST: 0.0272, and ExaBayes: 0.0242) and very high mean of κ (ASTRAL: 3.5832, MP-EST: 3.31, and ExaBayes: 3.8391). Whereas IQepro, IQeul, PhyloBayes and RaxML trees show very high mean of ω (IQepro: 0.1038, IQeul: 0.1072, PhyloBayes: 0.0784 and RaxML: 0.0812) and very low mean κ (IQepro: 1.3465, IQeul: 1.3375, PhyloBayes: 1.3708 and RaxML: 1.3528). This indicates that when trees (ASTRAL, MP-EST and ExaBayes) estimate low ω and high κ, which means dN is also low as compared to dS and Ts is high as compared to Tv. When trees (IQepro, IQeul, PhyloBayes and RaxML) estimates high ω and low κ, that means dN is also high as compared dS and Ts is also low as compared to Tv.

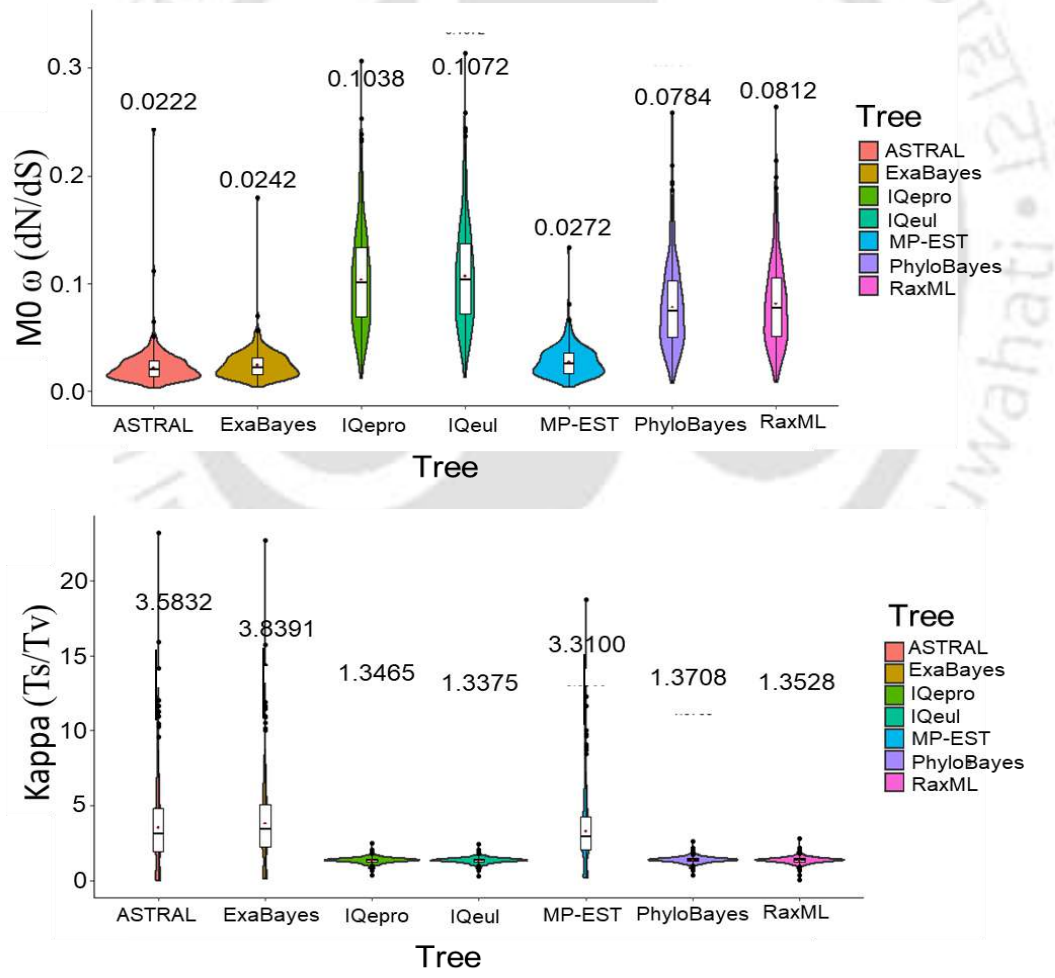


Figure 4.8: The substitution rate ω (dN/dS) (Top) and corresponding kappa (Ts/Tv) (Bottom) variation on different Phylogenetic trees using M0 model (one rate for a tree).

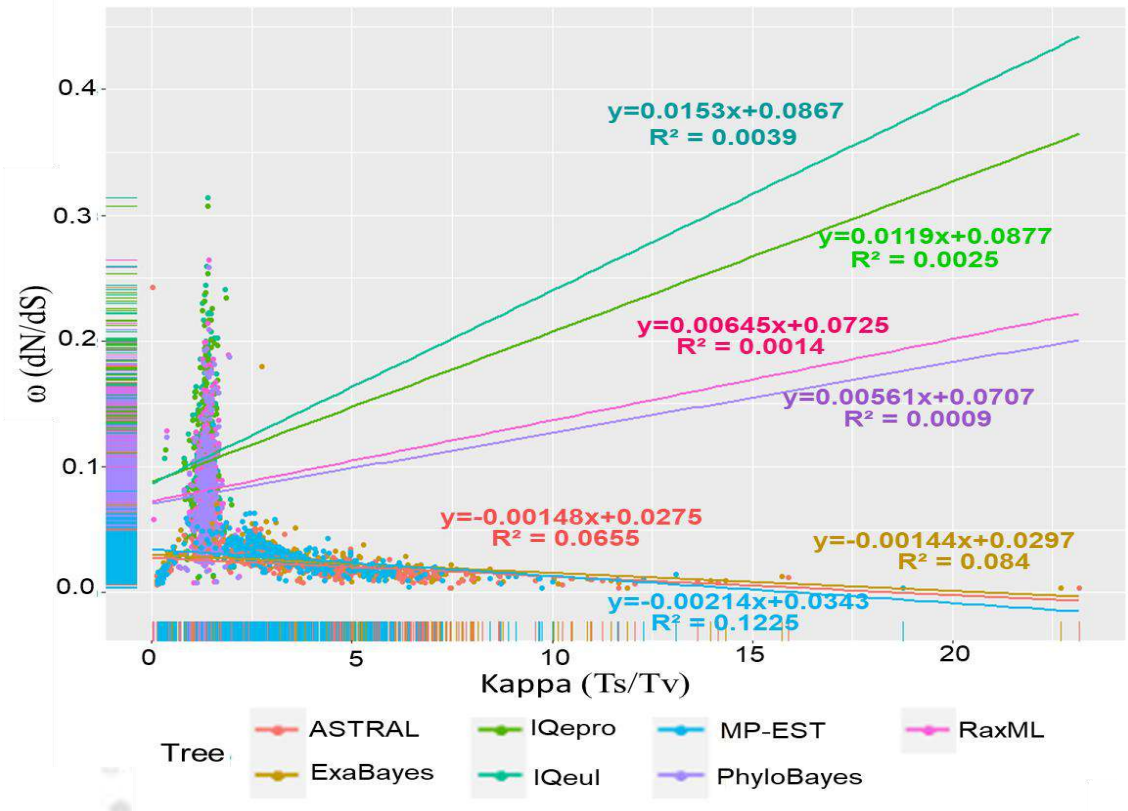


Figure 4.9: Scatter plot of ω and kappa values of 335 genes based on seven phylogenetic trees

Therefore, it suggests that estimation of substitution rate of any gene highly influenced by the choice of the reference tree. We further plotted scatter plot between ω and κ to understand the correlation between these two distinct substitution rates in the presence of a species tree. The plot between ω and κ display -ve slope for ASTRAL, MP-EST and ExaBayes species trees while IQpro, IQul, PhyloBayes and RaxML show +ve slope (Fig. 4.9). Although the square of correlation coefficient (R^2) is very low for IQpro, IQul, PhyloBayes and RaxML species trees than ASTRAL, MP-EST and ExaBayes.

4.3.8 Evolutionary Timescales of Diptera:

One of the most contentious issues in evolutionary biology is the evolutionary timeline of Diptera. Recent attempts to determine the Diptera timescale have mostly relied on

mitochondrial marker[8] with inadequate analysis using nuclear genes[97]. To unravel the evolutionary chronology of certain prominent Dipteran clades, we used 335 nuclear gene datasets based on previously created species trees and employed two distinct fossils calibration strategies (1 fossil calibration (1F) and 5 fossils calibration (5F)). Here, we report 95% credibility intervals for estimated divergence times for crown groups, as obtained through analysis with MCMCTree. The findings show that the estimated period of divergence depends greatly on the calibration of the fossils and the reference tree utilized in the analysis (Fig. 4.10, 4.11, 4.12).

According to the estimated chronology, Diptera diverged at ~290 mya estimated using 5F strategy, although 1F results showed Diptera diversification around ~247 mya in RaxML, ~264

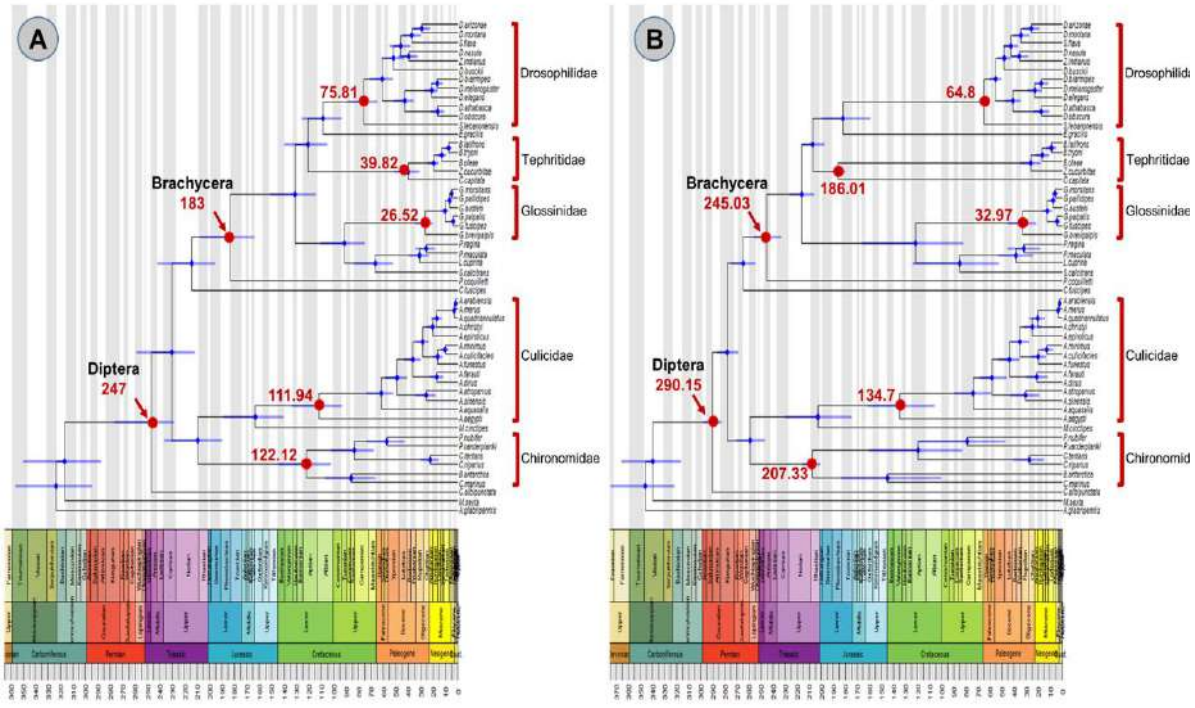


Figure 4.10: Chronogram Depicting the Evolutionary Timescale of 52 Diptera and Two Outgroups (Lepidoptera and Coleoptera) based on RaxML species tree. (A) Divergence times were estimated using Bayesian inference of 335 genes with 1 calibration point in MCMCTree. (B) Divergence times were estimated using Bayesian inference of 335 genes with 5 calibration points in MCMCTree. Horizontal bars represent 95% credibility intervals.

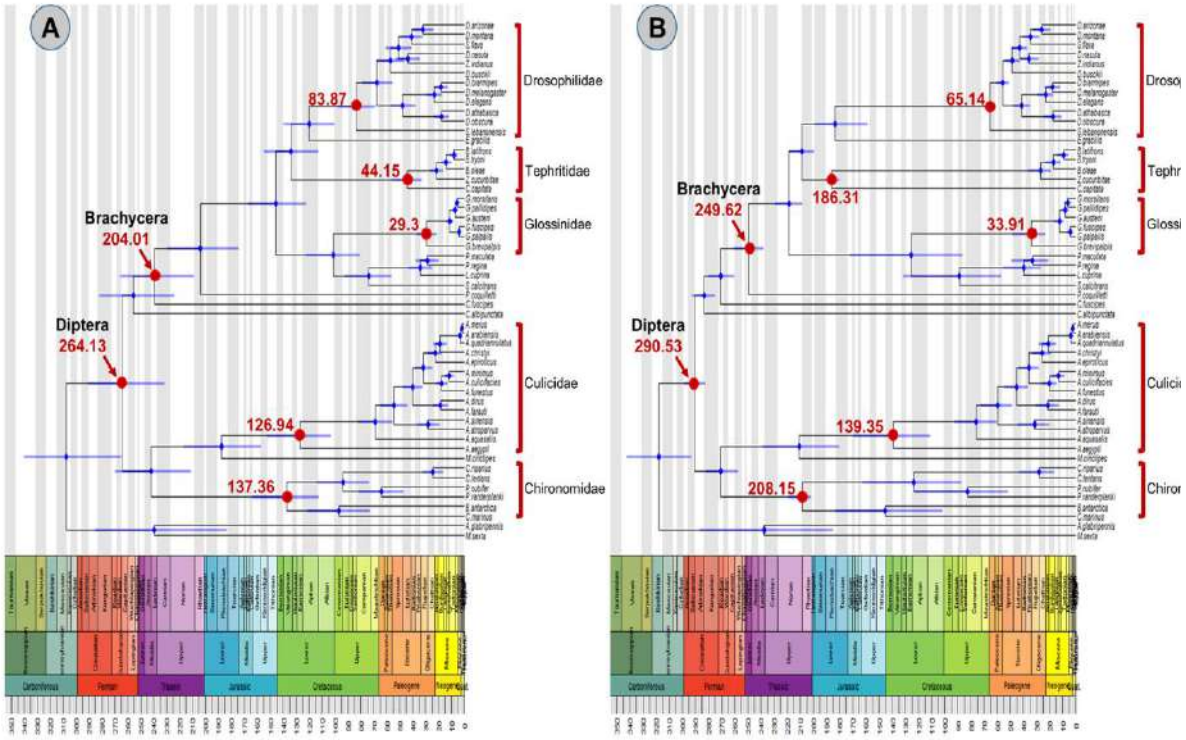


Figure 4.11: Chronogram Depicting the Evolutionary Timescale of 52 Diptera and Two Outgroups (Lepidoptera and Coleoptera) based on PhyloBayes species tree. (A) Divergence times were estimated using Bayesian inference of 335 genes with 1 calibration point in MCMCTree. (B) Divergence times were estimated using Bayesian inference of 335 genes with 5 calibration points in MCMCTree. Horizontal bars represent 95% credibility intervals.

mya in PhyloBayes, and 269 mya in ASTRAL. It indicates that split of Diptera lineage occurred during Late Permian to Late Triassic period. The diversification of Brachycera was inferred to begin at around ~244-250 mya through 5F, whereas 1F estimated ~183 mya in RaxML, ~204 mya in PhyloBayes, and ~197.5 mya in ASTRAL. Thus, 5F strategy suggests Brachycera diversified during Late Triassic period, in contrast 1F indicates that diversification of Brachycera happened during Late Jurassic to Early Triassic period. We inferred a Nematocera lineage, Chironomidae to have arisen in the Early Triassic (~207–208 mya) through 5F, while 1F strategy estimated Late Cretaceous (~122-137 mya) period for Chironomidae origin. The Culicidae divergence was dated between 134 mya and 139 mya using the 5F strategy, and 111 mya to 126 mya using the 1F strategy. Therefore, the Culicidae family is thought to have

originated around the Middle to Late Cretaceous period. We measured a Brachycera lineage, Tephritidae evolved in the Late Jurassic (~186 mya) by 5F strategy, whereas 1F strategy suggested a considerably later origination, around the Eocence epoch (~39-44 mya) of the Paleogene period. The Drosophilidae family evolved during the Late Paleocene epoch (~64-65.1 mya) of the Paleogene period, as estimated by 5F, and the Early Cretaceous period (~75-83.8 mya), as estimated by 1F. The Glossinidae family appeared around the Late Oligocene epoch (32-33.9 mya) via the 5F strategy and the Early to Late Oligocene epoch (26-29.3 mya) via the 1F strategy. According to our estimates, the Muscidae-Rhinophoridae- Calliphoridae lineage cluster arose in the Early Cretaceous (~78-88 mya) and (~65-74 mya) using 5F and 1F approach respectively. Therefore, with the exception of Drosophilidae, this comparison study reveals that estimation using 5 fossil calibration points suggested an earlier origin of Dipteran major clades than estimation using 1 fossil calibration point.

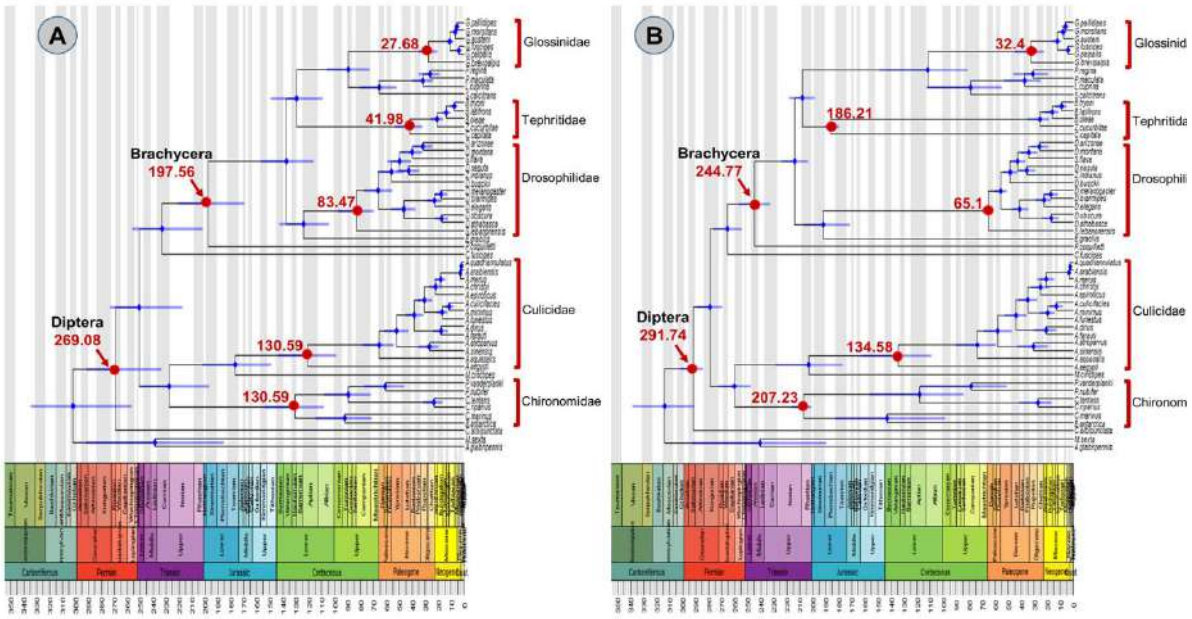


Figure 4.12: Chronogram Depicting the Evolutionary Timescale of 52 Diptera and Two Outgroups (Lepidoptera and Coleoptera) based on ASTRAL species tree. (A) Divergence times were estimated using Bayesian inference of 335 genes with 1 calibration point in MCMCTree. (B) Divergence times were estimated using Bayesian inference of 335 genes with 5 calibration points in MCMCTree. Horizontal bars represent 95% credibility intervals.

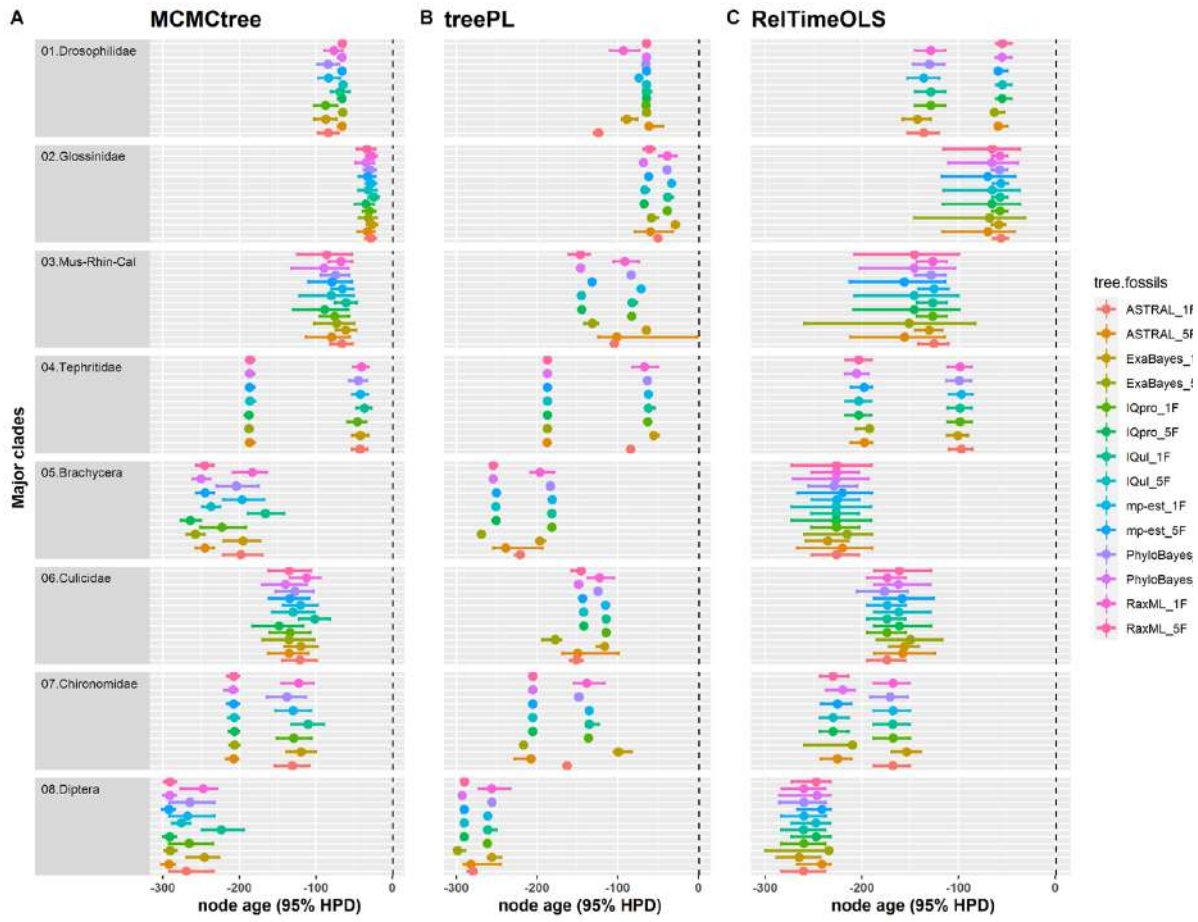


Figure 4.13: Comparative node age (95% HPD) of Dipteran major clades deduced from three different molecular dating methods. (A) MCMCtree, Bayesian method, (B) treePL, Penalized likelihood method, (C) RelTimeOLS, Ordinary least square method. Utilized various reference phylogenetic tree and fossil calibration shown in right side legend.

Furthermore, we examined the diversification of major clades of Diptera using two alternative molecular dating methods: Penalized likelihood estimate by treePL and the Ordinary least square approach by RelTimeOLS (Fig. 4.13).

4.3.9 Exploration of trees in the tree space:

As we saw in the preceding sections of this chapter, incongruent phylogenetic trees were inferred using different phylogenetic methods, which impedes a fundamental challenge in the study of evolution for a particular group of species[89]. While phylogenetic incongruence is typically seen as an annoyance, it can also signify real biological processes as well as relevant

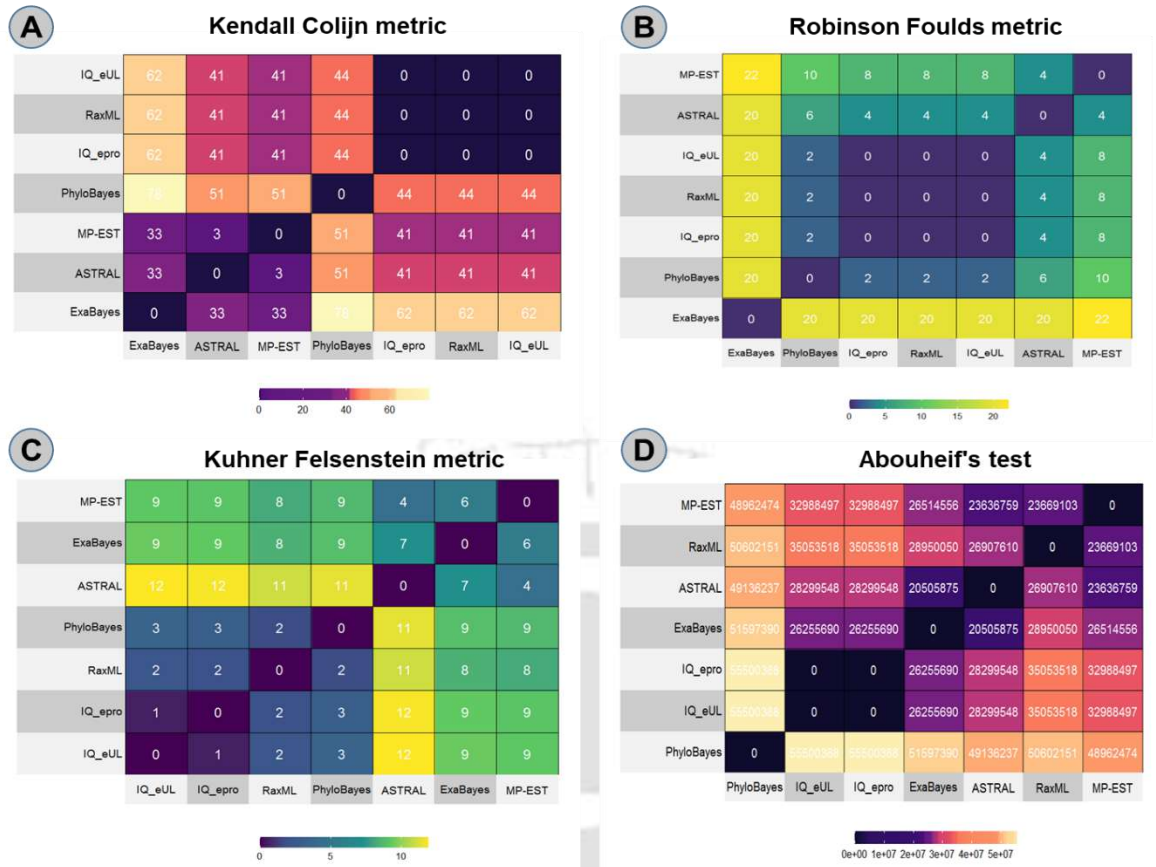


Figure 4.14: Heatmap of four distinct distance matrix ((A) Kendal Colijn metric, (B) Robinson Foulds metric, (C) Kuhner Felsenstein metric, and (D) Abouheif's metric) of seven species trees.

statistical uncertainty, both of which would yield valuable insights in evolutionary biology[35]. We investigated phylogenetic incongruence using phylogenetic tree landscapes, where tree metrics and multivariate analysis are combined to give low-dimensional representations of topological differences in a set of trees. It then led to the identification of related tree clusters and group-specific consensus phylogenies[35]. Using four separate measures Kendal Colijn metric, Robinson Foulds metric, Kuhner Felsenstein metric, and Abouheif's metric, we used TREESPACE to analyse potential similarity and inconsistencies in more detail. This analysis show that in Kendall Colijn metric and Robinson Foulds metric of the RaxML, IQepro and IQeul species trees have no difference in distance and Abouheif's test display that IQepro and IQeul trees don't have difference in distance (Fig. 4.14). Further when we floated the trees into

the treespace through multi-dimensional scaling (MDS) in two dimension and it exhibits that RaxML, IQepro and IQeul species trees overlapped in Kendall Colijn metric and Robinson Foulds metric treespace. Whereas, in Abouheif's treespace only IQepro and IQeul species trees overlap and coalscence-based species trees ASTRAL and MP-EST overlapped in only in Kendall Colijn metric tree space. We observed that RaxML and ExaBayes, as well as IQpro and IQul species trees, do not overlap in Kuhner Felsenstein metric tree space, but they are very close to each other (Fig. 4.15).

In addition to that, we also used individual gene tree clustering based on Robinson Foulds and Kendall-Colijn with two alternative number of clusters 5 and 10 distances, which revealed a diverse set of topologies (Fig. 4.16).

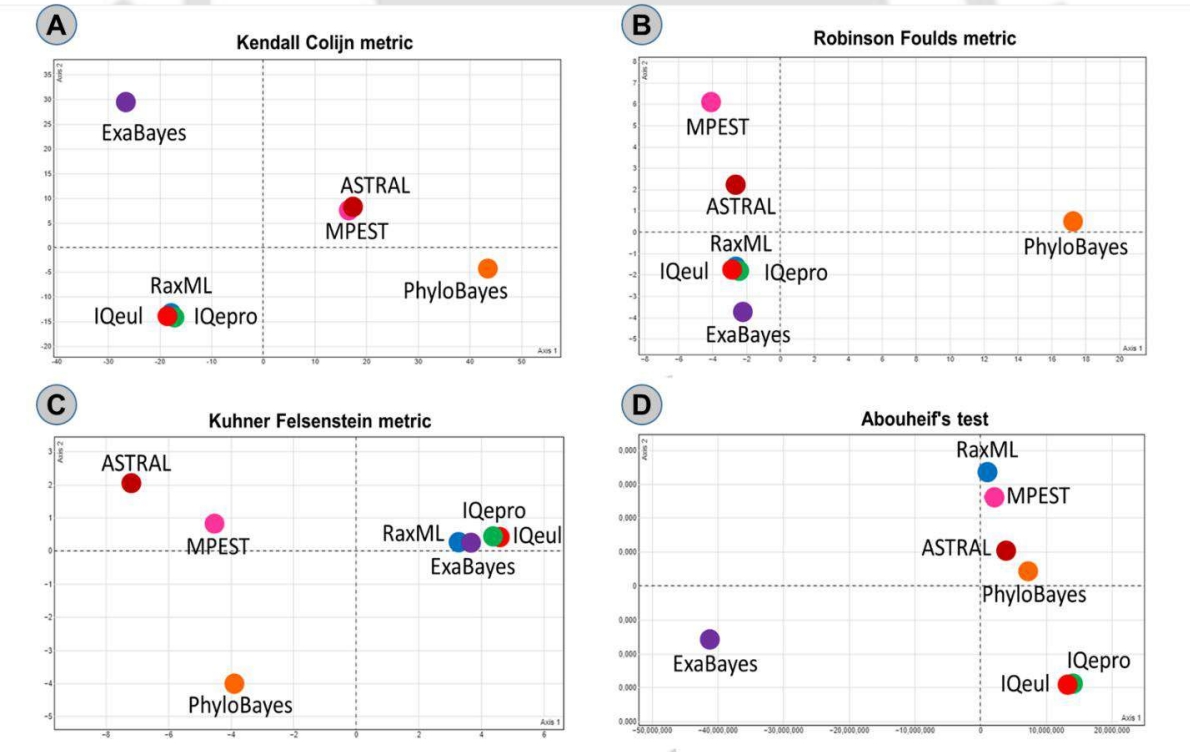


Figure 4.15: Position of seven species trees using multidimensional scaling (MDS) in two dimensions measured through four different metrics; ((A) Kendal Colijn metric, (B) Robinson Foulds metric, (C) Kuhner Felsenstein metric, and (D) Abouheif's metric

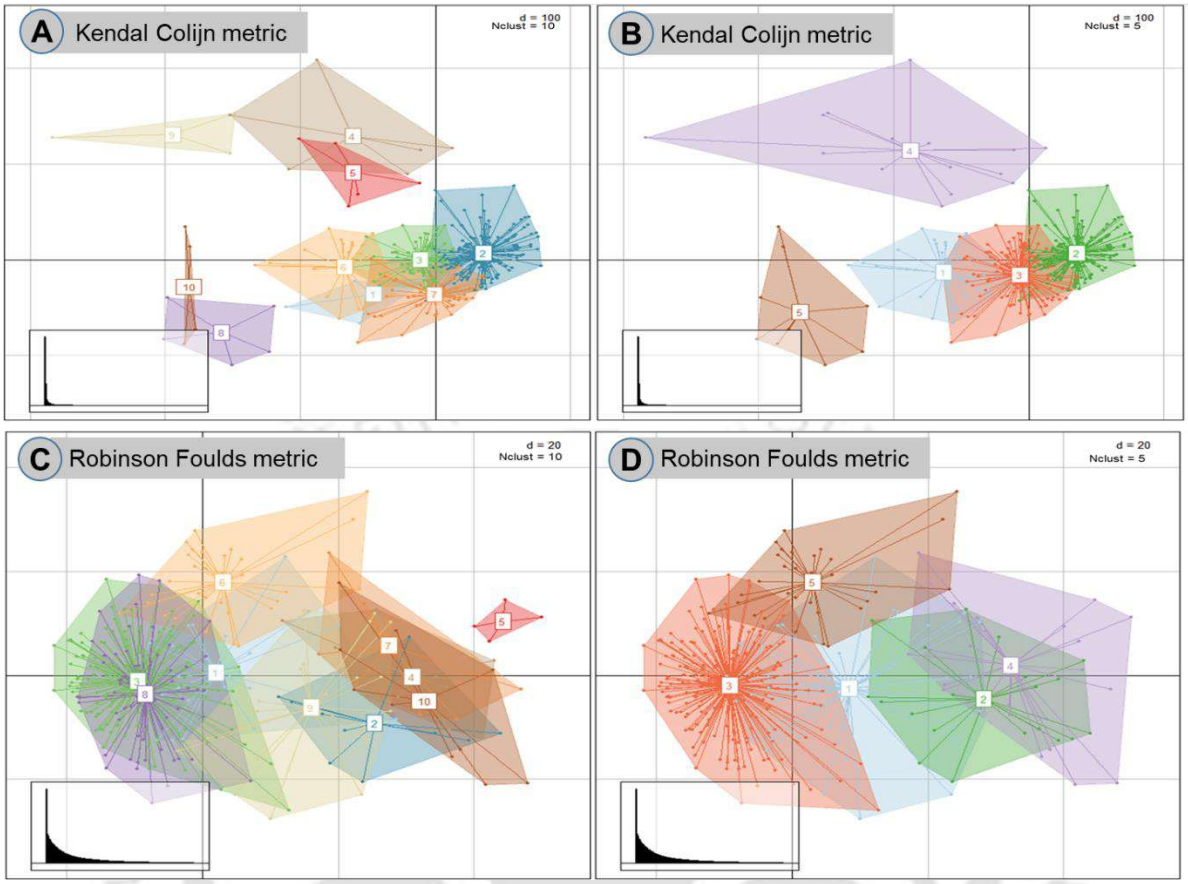


Figure 4.16: Gene trees cluster analysis of Metric Multidimensional Scaling (MDS), using (A) Kendal Colijn metric (number of clusters = 10), (B) Kendal Colijn metric (number of clusters = 5), (C) Robinson Foulds metric (number of clusters = 10), and (D) Robinson Foulds metric (number of clusters = 5).

4.4 Conclusion:

In this chapter we took advantage of vast amount of genomics data present in various public databases, especially the NCBI, to analyse and ascertain insights regarding the phylogenomic relationship of Diptera flies. Databases like NCBI receive genomic information from a multitude of sources at different times, and genomic data producing sequencing technologies are continuously improving, making it difficult to screen relevant data for research. In this study we used Benchmarking Universal Single-Copy Orthologs (BUSCO) to assess the completeness of the retrieved Diptera genome sequences, retain the quality genomes, and

identify the Single-Copy Orthologs for the analysis[98]. The BUSCO statistics display that, all of the retained Diptera genome (52) have more than 85% completeness, and 67% of the Diptera genomes have more than 95% completeness, according to OrthoDB's insecta_odb9 dataset[37]. The comparison of Genome, OGs, and CDS size reveals that Genome size varies substantially and simultaneously with the size of total OGs, whereas the size of the entire CDS remains consistent throughout the taxa analysed. The measurement of average nucleotide identity (ANI) of OGs reveals that closely related taxa (same family) have high ANI than the distantly related taxa (different family).

Since the primary focus this work is to resolve the Diptera phylogeny with large genome-scale datasets, yet, the relationships among Brachycera and Nematocera remained poorly resolved in our study, and the taxon sampling of was still rather limited. In our phylogenomic analyses, we recovered inconsistent topologies by using both coalescent- and concatenation-based approaches. This study demonstrates that the main source of problematic nodes of Diptera phylogeny are the unclear positioning of Psychodomorpha flies as well as the ambiguous relationship between Calliphoridae and Rhinophoridae. Increasing taxon sampling may result in more accurate inference or bootstrap support, but the availability of high-quality data is a major concern. As we noticed, high-quality genomic data is primarily concentrated on taxa containing model organisms (e.g., mosquitos, fruit flies), while genomic data for non-model organisms is limited. In light of this, we continued our investigation, focusing on the relationships between the major clades as well as other factors that may have been hampered by the spurious relationship. We calculated topological concordance and discordance between 335 gene trees and the species tree, showing low amount of topological concordance but a large degree of gene tree heterogeneity at deep internal branches, especially along the backbone of the Diptera phylogeny. Similarly, the analysis of internode certainty also results poor certainty at the backbone of the Diptera phylogeny. The concatenation approach implicitly ignores some

intricate evolutionary issues, such as gene tree conflict owing to ILS and hybridization, by assuming that all genes have the same or similar evolutionary histories. Although, in this study we have also found conflicting gene trees with coalescence-based analysis. This study also shows that estimation of ω = (dN/dS) and κ (Ts/Tv) influenced by the choice of reference tree used. We also found that, rise in GC content in any gene is associated with an increase in codon usage bias, an increase in the quantity of GC biased codons, an increase in gene expression, as well as an enhancement in translational efficiency.

The estimated chronology of major Diptera clade divergence utilizing 335 nuclear gene datasets demonstrates that it has been influenced by reference species trees and fossil calibration strategies. Our findings show that the Diptera lineage split occurred between the Late Permian and Late Triassic periods. Although it is noteworthy that 5 fossil calibration strategies show that Brachycera diversified during the Late Triassic period, one fossil calibration strategy suggests that Brachycera originated during the Late Jurassic to Early Triassic period. We found sharp contradiction in estimated divergence time of some of the key clades, such as the Chironomidae and Tephritidae families, using different fossil calibration methodologies. Our analysis reveals that the Culicidae family evolved during the Middle to Late Cretaceous period, the Drosophilidae family emerged during the Late Paleogene to Early Cretaceous period, the Glossinidae family appeared during the Late Paleogene period, and the Muscidae-Rhinophoridae-Calliphoridae lineage cluster arose during the Early Cretaceous period. Overall, our study estimates the earlier origin of some of the major Diptera clades by 5 fossil calibration points compared to estimates based on a single fossil calibration point.

In this study, we also looked at tree space for both species trees and gene trees to get a better understanding of the incongruence between inferred phylogeny from different methodologies and different genes. The multidimensional scaling of species trees and gene trees using various

metrics illustrates the placement and clustering of species trees and gene trees in two – dimensional space.

4.5 References:

1. Grimaldi, D. A. & Engel, M. S. *Evolution of the insects*. (Cambridge University Press, 2005).
2. Wiegmann, B. M. *et al.* Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* **108**, 5690–5695 (2011).
3. Skevington, J. H. & Dang, P. T. Exploring the diversity of flies (Diptera). *Biodiversity* **3**, 3–27 (2002).
4. Yeates, D. K. & Wiegmann, B. M. Congruence and controversy: toward a higher-level phylogeny of Diptera. *Annu. Rev. Entomol.* **44**, 397–428 (1999).
5. Abd-Algalil, Zambare, S. P. & Mashaly. First record of Chrysomya saffranea (Diptera: Calliphoridae) of forensic importance in India. *Trop. Biomed.* **33**, 102–108 (2016).
6. Bunchu, N. *et al.* Morphology and Developmental Rate of the Blow Fly, *Hemipyrellia ligurriens* (Diptera: Calliphoridae): Forensic Entomology Applications. *J. Parasitol. Res.* **2012**, 1–10 (2012).
7. Sinha, S. K. Sarcophagidae, Calliphoridae and Muscidae (Diptera) of the Sundarbans Biosphere Reserve, West Bengal, India. *Occas. Pap. - Rec. Zool. Surv. India* (2009).
8. Zhao, Z. *et al.* The Mitochondrial Genome of Elodia flavipalpis Aldrich (Diptera: Tachinidae) and the Evolutionary Timescale of Tachinid Flies. *PLoS One* **8**, 61814 (2013).
9. Wang, K. *et al.* The complete mitochondrial genome of the Atylotus miser (Diptera: Tabanomorpha: Tabanidae), with mitochondrial genome phylogeny of lower Brachycera (Orthorrhapha). *Gene* **586**, 184–196 (2016).
10. Narayanan Kutty, S. *et al.* Phylogenomic analysis of Calyptratae: resolving the phylogenetic relationships within a major radiation of Diptera. *Cladistics* **35**, 605–622 (2019).
11. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evol. Int. J. Org. Evol.* **63**, 1–19 (2009).
12. Nelson, L. A. *et al.* Beyond barcoding: A mitochondrial genomics approach to molecular phylogenetics and diagnostics of blowflies (Diptera: Calliphoridae). *Gene* **511**, 131–142 (2012).
13. Carolina, A. *et al.* Large-scale mitogenomics enables insights into Schizophora (Diptera) radiation and population diversity. *Sci. Rep.* **6**, 1–13 (2016).
14. Ballard, J. W. O. & Whitlock, M. C. The incomplete natural history of mitochondria. *Mol. Ecol.* **13**, 729–744 (2004).
15. Galtier, N., Benoit Nabholz, S. Glémin, and G. D. D. H. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol. Ecol.* **18**, 4541–4550 (2009).
16. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nat. 2008 4527188* **452**, 745–749 (2008).
17. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–31 (2014).
18. Smith, S. A. *et al.* Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* **480**, 364–367 (2011).
19. Yang, L. *et al.* Phylogenomic Insights into Deep Phylogeny of Angiosperms Based on Broad Nuclear Gene Sampling. *Plant Commun.* **1**, 100027 (2020).
20. Sharma, P. P. *et al.* Phylogenomic Interrogation of Arachnida Reveals Systemic Conflicts in Phylogenetic Signal. *Mol. Biol. Evol.* **31**, 2963–2984 (2014).

21. Chaudhary, R., Burleigh, J. G. & Fernández-Baca, D. Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol. Biol. 2013 81* **8**, 1–12 (2013).

22. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 1–15 (2015).

23. Galtier, N. & Daubin, V. Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 4023–4029 (2008).

24. Ané, C., Larget, B., Baum, D. A., Smith, S. D. & Rokas, A. Bayesian Estimation of Concordance among Gene Trees. *Mol. Biol. Evol.* **24**, 412–426 (2007).

25. Knowles, L. L. Estimating Species Trees: Methods of Phylogenetic Analysis When There Is Incongruence across Genes. *Syst. Biol.* **58**, 463–467 (2009).

26. Bayzid, M. S. & Warnow, T. Naive binning improves phylogenomic analyses. *Bioinformatics* **29**, 2277–2284 (2013).

27. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).

28. Boussau, B. *et al.* Genome-scale coestimation of species and gene trees. *Genome Res.* **23**, 323–330 (2013).

29. Salichos, L., Stamatakis, A. & Rokas, A. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).

30. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).

31. Krzemiński, W., Krzemińska, E. & Papier, F. Grauvogelia arzvilleriana sp.n. - the oldest Diptera species [Lower-Middle Triassic of France]. *Acta Zool. Cracoviensia* **37**, (1994).

32. Montagna, M. *et al.* Recalibration of the insect evolutionary time scale using Monte San Giorgio fossils suggests survival of key lineages through the End-Permian Extinction. *Proc. R. Soc. B* **286**, (2019).

33. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science (80-.).* **346**, 763–767 (2014).

34. Cerretti, P. *et al.* First fossil of an oestroid fly (Diptera: Calyptratae: Oestroidea) and the dating of oestroid divergences. *PLoS One* **12**, e0182101 (2017).

35. Jombart, T., Kendall, M., Almagro-Garcia, J. & Colijn, C. TREESPACE: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* **17**, 1385–1392 (2017).

36. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

37. Zdobnov, E. M. *et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45**, D744–D749 (2017).

38. Keller, O., Kollmar, M., Stanke, M., Waack, S. & Bateman, A. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).

39. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, 1002195 (2011).

40. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).

41. 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* **31**, 3210–3212 (2015).

42. Barter, R. L. & Yu, B. Superheat: An R Package for Creating Beautiful and Extendable Heatmaps for Visualizing Complex Data. *J. Comput. Graph. Stat.* **27**, 910–922 (2018).

43. Pritchard, L., Cock, P., Esen, Ö. & YT. widdowquinn/pyani: v0.2.8. (2019)

doi:10.5281/ZENODO.2584238.

44. Cutter, A. D., Wasmuth, J. D. & Blaxter, M. L. The evolution of biased codon and amino acid usage in nematode genomes. *Mol. Biol. Evol.* **23**, 2303–2315 (2006).

45. Zhang, Z. CUA: a Flexible and Comprehensive Codon Usage Analyzer. *bioRxiv* 022814 (2015) doi:10.1101/022814.

46. Xia, X. DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *J. Hered.* **108**, 431–437 (2017).

47. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

48. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).

49. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018).

50. Bayega, A. *et al.* De novo assembly of the olive fruit fly (Bactrocera oleae) genome with linked-reads and long-read technologies minimizes gaps and provides exceptional Y chromosome assembly. *BMC Genomics 2020 211* **21**, 1–21 (2020).

51. Gilchrist, A. S. *et al.* The draft genome of the pest tephritid fruit fly Bactrocera tryoni: Resources for the genomic analysis of hybridising species. *BMC Genomics* **15**, 1–17 (2014).

52. Papanicolaou, A. *et al.* The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biol.* **17**, 1–31 (2016).

53. Sanchez-Flores, A. *et al.* Genome evolution in three species of cactophilic drosophila. *G3 Genes, Genomes, Genet.* **6**, 3097–3105 (2016).

54. Bracewell, R., Chatla, K., Nalley, M. J. & Bachtrog, D. Dynamic turnover of centromeres drives karyotype evolution in drosophila. *Elife* **8**, (2019).

55. Chen, Z. X. *et al.* Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Res.* **24**, 1209–1223 (2014).

56. Zhou, Q. & Bachtrog, D. Ancestral Chromatin Configuration Constrains Chromatin Evolution on Differentiating Sex Chromosomes in Drosophila. *PLOS Genet.* **11**, e1005331 (2015).

57. Chakraborty, M. *et al.* Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nat. Genet.* (2017) doi:10.1038/s41588-017-0010-y.

58. Parker, D. J. *et al.* Inter and Intraspecific Genomic Divergence in Drosophila montana Shows Evidence for Cold Adaptation. *Genome Biol. Evol.* **10**, 2086–2101 (2018).

59. Mohanty, S. & Khanna, R. Genome-wide comparative analysis of four Indian Drosophila species. *Mol. Genet. Genomics* **292**, 1197–1208 (2017).

60. Nozawa, M., Onizuka, K., Fujimi, M., Ikeo, K. & Gojobori, T. Accelerated pseudogenization on the neo-X chromosome in Drosophila miranda. *Nat. Commun.* **7**, 1–9 (2016).

61. Vicoso, B. & Bachtrog, D. Numerous Transitions of Sex Chromosomes in Diptera. *PLOS Biol.* **13**, e1002078 (2015).

62. Gloss, A. D. *et al.* Evolution of herbivory remodels a Drosophila genome. *bioRxiv* **767160**, (2019).

63. Attardo, G. M. *et al.* Comparative genomic analysis of six Glossina genomes, vectors of African trypanosomes. *Genome Biol. 2019 201* **20**, 1–31 (2019).

64. Anstead, C. A. *et al.* Lucilia cuprina genome unlocks parasitic fly biology to underpin future interventions. *Nat. Commun. 2015 61* **6**, 1–11 (2015).

65. Andere, A. A., Platt, R. N., Ray, D. A. & Picard, C. J. Genome sequence of Phormia regina Meigen (Diptera: Calliphoridae): implications for medical, veterinary and forensic research. *BMC Genomics* **17**, 842 (2016).

66. Kraaijeveld, K., Neleman, P., Mariën, J., de Meijer, E. & Ellers, J. Genomic Resources for Goniozus legneri, Aleochara bilineata and Paykullia maculata, Representing Three Independent Origins of the Parasitoid Lifestyle in Insects. *G3 Genes|Genomes|Genetics* **9**, 987–991 (2019).

67. Dikow, R. B., Frandsen, P. B., Turcatel, M. & Dikow, T. Genomic and transcriptomic resources for assassin flies including the complete genome sequence of Proctacanthus coquilletti (Insecta: Diptera: Asilidae) and 16 representative transcriptomes. *PeerJ* **2017**, e2951 (2017).

68. Olafson, P. U. *et al.* The genome of the stable fly, Stomoxys calcitrans, reveals potential mechanisms underlying reproduction, host interactions, and novel targets for pest control. *BMC Biol.* **19**, 1–31 (2021).

69. Khanna, R. & Mohanty, S. Whole genome sequence resource of Indian Zaprionus indianus. *Mol. Ecol. Resour.* **17**, 557–564 (2017).

70. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.

71. Drake-Brockman, A., Neafsey, D. E. & Waterhouse, R. M. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science (80-.).* **347**, 1258522 (2015).

72. Ghurye, J. *et al.* A chromosome-scale assembly of the major African malaria vector Anopheles funestus. *Gigascience* **8**, 1–8 (2019).

73. Zhou, D. *et al.* Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. *BMC Genomics* **15**, 1–13 (2014).

74. Kelley, J. L. *et al.* Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nat. Commun.* **5**, 1–8 (2014).

75. Kutsenko, A. *et al.* The Chironomus tentans genome sequence and the organization of the Balbiani ring genes. *BMC Genomics* **15**, 1–12 (2014).

76. Kaiser, T. S. *et al.* The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature* **540**, (2016).

77. Gusev, O. *et al.* Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. *Nat. Commun.* **5**, 4784 (2014).

78. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. (2012) doi:10.1093/molbev/mss020.

79. Lanfear, R., Calcott, B., Kainer, D., Mayer, C. & Stamatakis, A. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* (2014) doi:10.1186/1471-2148-14-82.

80. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

81. 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. *Mol. Biol. Evol.* **32**, 268–274 (2015).

82. Aberer, A. J., Kobert, K. & Stamatakis, A. Exabayes: Massively parallel bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.* **31**, 2553–2556 (2014).

83. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).

84. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 15–30 (2018).

85. Liu, L., Yu, L. & Edwards, S. V. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* **10**, 25–27 (2010).

86. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

87. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

88. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).

89. Kendall, M. & Colijn, C. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Mol. Biol. Evol.* **33**, 2735–2743 (2016).

90. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468 (1994).

91. Pavoine, S., Ollier, S., Pontier, D. & Chessel, D. Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities. *Theor. Popul. Biol.* **73**, 79–91 (2008).

92. Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L. & Atkinson, P. W. The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics* **12**, 1–23 (2011).

93. Sharpl, P. M. & Li2, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1987 (1987).

94. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).

95. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).

96. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21 (1981).

97. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–7 (2014).

98. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

CHAPTER 5

Dietary Adaptation of Diptera

> "The survival or preservation of certain favoured words in the struggle for existence is natural selection."
>
> — Charles Darwin

TH-2821_146106003

CHAPTER 5

Classification of Diptera based on different dietary adaptation and the tempo and mode of their diversification

Abstract:

Diptera are diverse in nature present in almost all biogeographic realm and their ability to thrive on various food sources fortified them to survive in multiple standings of food chains in different morphological stages. For living successful life every species needs energy and mitochondria provides more than 95% of the energy of the cell. Mitochondrial ATP generation bioenergetic efficiency depends on variability of nutrient molecule and availability of oxygen. We hypothesize that dietary strategy plays a role in the occurrence of mitochondrial mutation. To put these hypotheses to the test, we conducted a comparative genomic investigation of 112 Diptera complete mitochondrial genomes to explore the impact of different feeding habits on mitochondrial evolution. We found that food habit of Diptera is significantly correlated with nucleotide substitution either nonsynonymous (dN) or synonymous (dS) or their ratio (dN/dS, ω). This study shows that mitochondrial protein coding genes of detritivores flies have significantly higher nucleotide substitution rate. Whereas, heamtophagy and carnivorous have lower nucleotide substitution rate. We show that net diversification rate (the cumulative effect of speciation and extinction), and transition rate differs significantly among living flies,

depending upon food habit. We also tested dS, dN, and their ratio (ω) as continuous traits of Diptera evolution and investigated their relevance in evolutionary diversification of Diptera. Furthermore, we find that lineages with different food habit evolve towards their optimal continuous molecular trait (substitution rates) value.

5.1 Introduction:

An appropriate mitochondrial activity, oxidative phosphorylation (OXPHOS) needs the synchronized interaction of dozens of mitochondrial genes and hundreds of nuclear genes. These interactions result in systems that turn the food we ingest and the oxygen we inhale into energy for our life as well as systems that control a variety of other cellular processes, making mitochondria a fascinating link between macroevolution and microevolution[1]. The mitochondrial DNA (mtDNA or mitogenome) strand is a closed circular DNA strand that encodes the essential proteins for the oxidative phosphorylation machinery: The NADH dehydrogenase or NADH ubiquinone oxidoreductase complex is comprised of seven subunits (ND: ND1, 2, 3, 4, 4L, 5, and 6), the cytochrome b subunit of the ubiquinol cytochrome c oxidoreductase or cytochrome bc1 complex (CytB), and three subunits of the cytochrome c oxidase (COX1, COX2, COX3), and two subunits of ATP synthase (ATPase: ATP6 and ATP8)[2]. Four of the five OXPHOS complexes' core components interact with additional nuclear DNA genome-encoded subunits[3]. In OXPHOS, reducing equivalents produced by the oxidation of nutrients like glucose are transported along a chain of molecules (the electron transport chain) with increasing standard reduction potentials. The generated free energy is converted into a proton gradient, which is utilized by ATP synthase to convert ADP (adenosine diphosphate) into ATP (adenosine triphosphate) [3].

The amount of ATP produced by mitochondria per nutrient molecule is referred to as mitochondrial bioenergetic efficiency[4]. Changes in dietary nutrients can effect mitochondrial bioenergetics during biogeographic, demographic, cultural, or seasonal transitions through

influencing electron transport system (ETS) function[5,6]. Different combinations of major energy-producing micronutrients (e.g., protein and carbohydrates) in fly diets have been shown to influence metabolism, behaviour, and biochemistry[5–7]. In addition, mutations in mitochondrial complex-I influence Drosophila larval growth as dietary nutrients are absorbed through this complex during immaturity[6]. According to various studies, mutations in mitochondrial protein-coding genes involved in OXPHOS (generates up to 95 % in eukaryotic cells) may have a direct influence on metabolic performance[8–11]. Due to such necessity of this biochemical pathway, various empirical studies have shown that assessing selection forces acting on mtDNA can reveal essential insight on the adaptive evolution of the mitochondrial genome[12–15].

In general, two types of mutations are observed in protein coding genes of any species: synonymous and non-synonymous. Although synonymous mutations do not change the translated amino acid, selection on certain synonymous substitutions induces bias in most genomes, which can further influence gene expression[16–18]. Non-synonymous mutations, on the other hand, alter the amino acid, that could result in phenotypic and morphological variations[19–21]. Since natural selection usually operates at the protein level, synonymous and nonsynonymous mutations are driven by various selective forces and fixed at distinct rates[22]. Hence, comparing a protein's synonymous and nonsynonymous substitution rates can reveal the magnitude and direction of natural selection acting on it[23,24]. Whereas, the ω ratio quantifies the degree and direction of selection on amino acid changes, with values of $\omega < 1$, $=1$, and > 1 signifying negative purifying selection, neutral evolution, and positive selection, respectively[22].

Diptera are diverse in nature present in almost all biogeographic realm and their ability to thrive on various food sources fortified them to survive in multiple standings of food chains in different climatic niches and in different morphological stages[25,26]. Overall larval Diptera

exhibit nearly every sort of feeding habit that includes herbivores, carnivores, detritivores and hematophagy in adult phase[26,27]. For living successful life every species needs energy and mitochondria being the power house of the cell different factors namely climate, temperature, nutrition plays key role for alteration of mitochondrial functions mutation rate[14].

Although some particular Dipteran fly lineages have been thoroughly investigated in terms of diet adaptation, few studies have investigated the impact of diet on large-scale macroevolutionary aspects. Here we present a quantitative macroevolutionary assessment of the tempos of lineage diversification and trophic transition across extant Diptera. Despite the fact that nutrition is intimately linked to the evolution of flies, there is no evidence on how trophic strategy (i.e., herbivory or carnivory) influences transition rates and drives speciation and extinction rate[27].

As far our understanding synonymous substitution and codon usage bias influences gene expression, while nonsynonymous substitution effects protein structure and function. Therefore, both parameters might have substantial biological significance on modulating phenotypic and morphological traits. So, the synonymous and nonsynonymous substitution rate can be cast as continuous character for attempting to relate a change in some aspect of the biology of the lineage and developing an idea about how evolution works.

In this work, we have attempted to respond to the aforementioned notion in a systematic manner. We collected dietary information of 112 Diptera flies (like the number of taxa from Chapter 3). The synonymous, nonsynonymous substitution rate and ω ratio was estimated for thirteen mitochondrial protein coding genes individually as well as in combination. We investigated whether flies' dietary habits were correlated to the inferred continuous molecular traits synonymous substitution rate (dS), non-synonymous substitution rate (dN), and ω ratio. We also investigated whether there was a transition between species with distinct feeding habits. Further, we looked at how discrete binary dietary habits (herbivore vs carnivore,

detritivore vs non-detritivore, and haematophagy vs non-haematophagy) influenced diversification rates. Whether or not species with varied food habits evolve in a deterministic way, and if so, whether they merge around a single or multiple optimum values of molecular traits (dS, dN, and ω). In addition, we modelled the evolution of continuous molecular traits (dS, dN, and ω) to understand the importance of molecular traits in the diversification process.

5.2 Materials and Method:

5.2.1 Phylogeny construction and substitution rate estimation:

Raw mitochondrial genomes (112 Diptera species) were downloaded from the National Center for Biotechnology Information database. All protein-coding genes were aligned through MAFFT algorithm guided by amino acid using TranslatorX, and each multiple nucleotide sequence alignment for the 13 mtDNA concatenated and were used to reconstruct a Diptera phylogenetic tree using RaxML by constraining few taxa into their desired family. The CodeML program implemented in the PAML package[22] was used to compute the nonsynonymous (dN) and synonymous (dS) substitution rates along each branch of the tree. Model 1 was conducted, which allows the overall substitution rate and the ratio of dN/dS changes to have branch-specific values. The ω (dN/dS), dN and dS values associated with the external branches were used in the subsequent analyses as we focused only on the rate of accumulation of slightly deleterious mutations (dN/dS) between modern species and their most recent reconstructed ancestors. Further, we calculated distance of tips to the root using adephylo package of R and renamed the values as DRω, DRdN and DRdS to get total substitution rate of a taxon from the root. All statistical analysis was performed by using R statistical package.

5.2.2 Data Collection and Dietary Categorization:

We constructed a database of diets of Diptera species from published accounts of primary research reporting data obtained through literature search. We recorded complete descriptions

of diet from the sources; these descriptions were then converted to discrete character codings for the presence or absence of four food types in the diet: we tested whether Herbivore (based on mature leaves, stems, fruits, and bark), Carnivore (based on animal, insect body part), Detritivore (decomposing plant and animal parts as well as faces), Haematophagy (blood sucking).

5.2.3 Discrete Trait-Dependent Diversification:

We compared the fit of models with equal rates (ER), symmetric (SYM) forward/reverse rates, meristic (MER) transitions that occur in a stepwise manner, and all rates different (ARD) using the fitdiscrete function in the R package geiger[28]. To account for polymorphism, we used fitpolyMk function in the R package phytools as the feeding habit of flies is not entirely distinguishable[29]. Here, also we compared the fit of models with ER, SYM, and ARD. To calculate ancestral state probabilities, we employed the AIC-best model. As a result of this, we constructed marginal ancestral states for the best model, as defined by the lowest AIC.

To assess the effect of different Dietary habit on diversification rates, we used the BiSSE (binary state), MuSSE (multiple state) implemented in the diversitree package in R and HiSSE (hidden binary state) implemented in hisse package in R[30–32]. We fitted diverse range of models and evaluated them based on log-likelihoods. We performed a likelihood ratio test and computed Akaike information criterion (AIC) weights to identify the best models, and we conducted MCMC estimations of those models for 10000 generations for BiSSE and MuSSE. To test whether diversification rate heterogeneity is associated with shifts in Dietary habit or changes in other unmeasured traits, we used a model with four states that describes the joint evolution of Dietary habit as well as an unobserved character with hidden states A and B. We fit 24 different models to the 3 sets Dietary data of Diptera (Herbivore and Carnivore, Detritivore and Non-detritivore, Haematophagy and Non-haematophagy). Four of these models corresponded to BiSSE models that either removed or constrained parameters, 16

corresponded to various HiSSE models that assumed a hidden state associated with and four corresponded to various forms of our trait- both the observed states, independent models (i.e., CID-2, CID-4).

5.2.4 Continuous Trait-Dependent Diversification and Disparity analysis:

Here we consider nonsynonymous (dN), synonymous (dS) substitution rates and their ratio ω (dN/dS) as continuous trait. As synonymous substitution in a gene does not change resultant amino acid but it has significant role in mRNA transcription, protein translation whereas, non-synonymous substitution directly alters the subsequent amino acid as consequence protein structure and protein folding and the ω value more than 1 signifies positive adaptive mutation. Diversification of traits (dS, dN and ω) was inspected by trait disparity-through-time (DTT) analysis using the R package 'geiger'[33]. The null model of trait evolution is Brownian motion (BM), a stochastic evolution model of constant variance, generated by averaging 1000 BM iterations. The disparity index (DI) quantified the disparity from the reconstructed evolution of trait from extant species and the median trait values under Brownian evolution simulations. Whereby a positive DI signifies a higher disparity than Brownian expectation, on the other hand a negative DI represents less disparity than Brownian expectation[34,35].

5.2.5 Phylogenetic ANOVA:

We used phylogenetic ANOVA to test for differences in molecular traits (dS, dN and ω) between feeding habit type in binary states (herbivore vs carnivore, detritivore vs non-detritivore, and haematophagy vs non-haematophagy). We performed a phylogenetic ANOVA in the R package geiger v 2.0.7 and phytools v 0.7.70 using the function aov.phylo and phylANOVA respectively. We ran 1,000 simulations using a Brownian motion model of evolution for the phylogenetic ANOVA.

5.2.6 Mode of molecular trait evolution:

Evolutionary models are designed to infer the several potential dynamics that shape phenotypic evolution and to test various assumptions about the mode of evolution. The assessment of continuous molecular traits (dS, dN and ω) evolution without any pre-defined classification of behavior was carried out using fitContinuous function in the R package 'geiger'[33]. We measured the fit of three evolutionary models to our phenotypic variables: Brownian motion (BM, diffusive drift), Ornstein-Uhlenbeck (OU, bounded evolution around a single phenotypic optimum), and Early-Burst (EB, exponential declining of evolutionary rates). The OU model offers the most fundamental mathematical explanation for an evolutionary process incorporating selection. It should be noted that BM is a special instance of OU and that OU is distinguished from BM by the presence of a defined optimum[36]. It can be represented by this equation:

$$dX(t) = \alpha\,[\theta - X(t)]\,dt + \sigma\,dB(t) \qquad (1)$$

The change in quantitative trait X along the branch of a phylogenetic tree that is divided into deterministic and stochastic components. The first component can be described as the influence of selection on the character, while the second can be explained as the result of random drift and other unmodeled forces. Equation (1) shows the amount of change in character *X* over a brief period; precisely, *dX* is the small change in character *X* during the small interval from time t to time t + dt. The term *dB(t)* stands for "white noise," which signifies that the random variables *dB(t)* is independent and uniformly distributed normal random variables, each having a mean of 0 and a variation of *dt*. The parameter α quantifies the selection intensity and when α = 0, the deterministic element of the OU model is eliminated, and equation (1) reduces to the typical BM model of pure drift. Higher values of α suggest increased levels of selection, resulting in trait distributions that are more closely distributed around their optima. BM has only two parameters: the evolutionary rate, σ^2, and root state of the trait, θ. While OU includes

additional parameters, $\theta_{1,.,n}$, reflecting the optimal state for each of the n food habit regimes modelled.

We evaluated the fit of a Brownian motion (BM) model of molecular traits (dS, dN, and ω) evolution in Diptera with Ornstein–Uhlenbeck (OU) models, which allow for different optimal molecular traits, based on pre-defined categorization (feeding habit). As $\alpha \rightarrow 0$, OU degenerates to BM. The fit of models with a single optimum size for all Diptera flies (OU1) and distinct optima for each feeding habit type in binary states (herbivore vs carnivore, detritivore vs non-detritivore, and haematophagy vs non-haematophagy—OU2) was compared. We would expect OU2 to best fit our data if food habit choice impacts molecular traits (dS, dN, and ω). If habitat has no effect on molecular traits, BM or OU1 should suit better. We fit all models using OUCH 2.17 [36] in R v. 4.0.2.

To evaluate whether possible ecological misclassifications impacted the outcomes of the OUCH analysis, we used R package 'SURFACE' and 'ℓ1ou'[37,38] to capture evolutionary shifts in molecular traits (dS, dN and ω) optima without a pre-defined classification of behavior. SURFACE employs the stepwise Akaike Information Criterion to first locate regime shifts on a tree and then determine if the shifts are toward converging regimes[37]. ℓ1ou detects shifts in trait evolution using a model-selection and evaluation approach based on the least absolute shrinkage and selection operator (LASSO) method. ℓ1ou also employs a phylogenetic Bayesian information criterion (pBIC), that considers the phylogenetic association between species as well as the complexities of predicting an unknown number of shifts at unknown positions in the phylogeny[38].

5.2.7 Modelling of continuous trait-dependent diversification:

In order to determine whether the rate at which lineages diversify is dependent on that lineage's continuous trait (dS, dN and ω) we ran trait-dependent diversification analysis using; a tip rate correlation technique termed inverse of equal-splits with simulated null model (ES-sim) and

TB-pgls[39,40]. ES-sim measures the tip-specific speciation rate and simulates a null distribution under a given model of trait evolution to test for significance. We ran ES-sim for both Brownian and OU null distributions by running 10,000 simulations, using the 'essim' and 'tbpgls' code in R[40]. For ES-sim Spearmen's correlation was used to determine a significant monotonic (i.e., linear, and sigmoidal) trait-dependent diversification relationship of extant species. The maximum likelihood approach Quantitative State Speciation and Extinction (QuaSSE) was also used for modeling of trait evolution and diversification using the R package 'diversitree' [31,41]. The phylogenetic tree and molecular trait data (dS, dN and ω) were used to model linear-, sigmoidal-, and modal-diversification for both stochastic trait evolution and directional trait evolution using a birth-death process[41]. The QuaSSE analysis computes maximum likelihood for model selection, from which the AIC was derived to determine model goodness-of-fit.

5.3 Result and discussion:

We reconstructed a maximum likelihood phylogenetic tree of flies based on 112 complete Diptera using thirteen protein coding genes from mitochondrial genomes. This tree was used for various downstream analysis in this chapter.

5.3.1 Dipteran traits and nucleotide substitution rates:

The ratio of the rates of nonsynonymous (change in amino acid) substitution over synonymous (silent) substitutions (dN/dS) substitutions is widely used to calculate the degree of selection. We estimated the dN/dS values associated with terminal branches to evaluate the degree of selection during each species' most recent divergence to see whether the mtDNA of different flies faced distinct selective pressures. We also estimated the rate of non-synonymous (dN) and synonymous (dS) mutation to establish if purifying or positive mutation was involved.

To measure the difference in dN/dS ratios of flies with different features, we classified our samples into three categories including 1) food habit (herbivore, carnivore, detritivore, and

haematophagy); 2) life-style (endo-parasite and ecto-parasite); 3) climate (equatorial, arid, warm temperate, cold temperate, and polar).

Synonymous substitution: This study shows statistics of synonymous substitution (dS) of various traits (Table 5.1) of 13 combined genes. Detritivores exhibit higher mean dS (1.6058) and hematophagy has lower mean dS (0.630842) than the species of other food habit. The median dS statistics show similar trend in which detritivore have higher median dS (1.6497) and Hematophagy have lower median dS (0.5094) with high significance level ($p= 2.5e^{-05}$). Distance to root dS (DRdS) also show higher mean DRdS for detritivores (3.4762) and low in carnivores (2.6250) with insignificant p value. The classification according to life style, mean dS and DRdS of endo-parasite (0.8635, 2.9458) is higher than ecto-parasite (0.6085, 2.5948) with statistically insignificant. The classification with climate habitat shows that mean of dS and DRdS in Equatorial (0.6186, 2.8236) and polar (0.8142, 2.5046) species have low compare to Warm Temperate species (1.0386, 3.0018) with marginal significance level (p ~.09).

Non synonymous substitution: The statistics (Table 5.2) of mean nonsynonymous substitution (dN) of various traits shows that species with detritivore food habit have higher mean dN (0.0420) and haematophagy have lower mean dN (0.0091). Similarly, the median dN also show that detritivore has higher median dN (0.0437) and haematophagy have lower median dN (0.0034) with significancy level ($p=2.5e^{-06}$). Whereas, DRdN of detritivores is lower (0.1322) than carnivores (0.1582) with marginal significance level (p = 0.032). The classification according to lifestyle shows high dN for endo parasite (0.0155) (insignificant, p=0.28) whereas, DRdN of endo parasite also shows higher (0.1639) value with marginal significance level (p=0.041). Climate wise classification shows higher dN of species in warm temperate region (0.0219) and lower in Equatorial (0.0112) with marginal significance level (p=0.069). DRdN of climate classification is not significant and agreeable with null hypothesis (p=0.92).

Table 5.1: The Values of dS (synonymous substitution) and DRdS of Traits for Each Subgroup Classified

Food habit	Mean dS	Med dS	Mean DRdS	Med DRdS	Life-style	Mean dS	Med dS	Mean DRdS	Med DRdS	Climate	Mean dS	Med dS	Mean DRdS	Med DRdS
Herbivores	1.0330	0.6381	3.2490	2.6173	Ecto-parasite	0.6086	0.5127	2.5948	2.4418	Equatorial	0.6187	0.5091	2.8236	2.5127
Carnivores	0.7574	0.5413	2.6250	2.5073	Endo-parasite	0.8635	0.5619	2.9459	2.5127	Arid	0.9779	0.6822	2.9482	2.7357
Detritivores	1.6058	1.6497	3.4762	2.4186						Warm Temperate	1.0387	0.6517	3.0018	2.5256
Hematophagy	0.6308	0.5094	2.7817	2.3769						Cold Temperate	0.9444	0.6085	2.9858	2.3920
										Polar	0.8143	0.6052	2.5047	2.4576

Table 5.2: The Values of dN (non-synonymous substitution) and DRdN of Traits for Each Subgroup Classified

Food habit	Mean dN	Med dN	Mean DRdN	Med DRdN	Life-style	Mean dN	Med dN	Mean DRdN	Med DRdN	Climate	Mean dN	Med dN	Mean DRdN	Med DRdN
Herbivores	0.0211	0.0081	0.1444	0.1515	Ecto-parasite	0.0095	0.0037	0.1463	0.1536	Equatorial	0.0112	0.0034	0.1477	0.1540
Carnivores	0.0146	0.0042	0.1582	0.1569	Endo-parasite	0.0155	0.0045	0.1639	0.1569	Arid	0.0211	0.0097	0.1531	0.1605
Detritivores	0.0420	0.0437	0.1322	0.1383						Warm Temperate	0.0219	0.0074	0.1498	0.1554
Hematophagy	0.0091	0.0034	0.1403	0.1536						Cold Temperate	0.0198	0.0081	0.1506	0.1541
										Polar	0.0162	0.0070	0.1533	0.1535

Table 5.3: The Values of ω (dN/dS) and DRω of Traits for Each Subgroup Classified

Food habit	Mean ω	Med ω	Mean DRω	Med DRω	Life-style	Mean ω	Med ω	Mean DRω	Med DRω	Climate	Mean ω	Med ω	Mean DRω	Med DRω
Herbivores	0.0154	0.0138	218.6069	1.5256	Ecto-parasite	0.0172	0.0095	1.5005	1.6131	Equatorial	0.0131	0.0108	189.1420	1.6093
Carnivores	0.0219	0.0112	15.2597	1.6192	Endo-parasite	0.0208	0.0119	171.4631	1.6246	Arid	0.0167	0.0149	127.5029	1.5345
Detritivores	0.0248	0.0257	196.9434	1.5961						Warm Temperate	0.0197	0.0135	108.0194	1.6077
Hematophagy	0.0168	0.0080	204.5029	1.6312						Cold Temperate	0.0197	0.0140	96.1417	1.6157
										Polar	0.0163	0.0141	1.4174	1.5822

The ω ratio (dN/dS): The ω statistics of different traits described in this section (Table 5.3). Detritivores have higher mean ω (0.0248) and herbivore have lower mean ω (0.0154. However, the lowest median ω observed for hematophagy (0.0080) and highest in detritivores (0.0257) with significant level high ($p=6.5e^{-06}$). Whereas, DRω of herbivores have high (218.6069) and carnivores have low (15.2596) with significant level (p = 0.0054). Mean ω of endo-parasite have higher (0.0207) than the ecto-parasite (0.0172) but p value is insignificant (p = 0.26). DRω of endo-parasite (171.4631) also higher than ecto parasite (1.5005) with significant level (p=.0039). Climatic classification of ω exhibit higher value for both temperate region (~.0196) and low for Equatorial (0.0130) with insignificant p-value (p = 0.24). DRω of Equatorial show higher (189.142) and polar species show lower (1.4174) with insignificant p-value (p = 0.36).

Our analysis shows that only trait food habit of Diptera flies that is statistically significant with the mean dS, dN, and ω. To investigate the skew of the data, we determined the median value of dS, dN, and ω for each subgroup. Detritivore has greater median for all the molecular traits (dS, dN, and ω) than the mean value, whereas other dietary habit traits have lower median value (Table 5.1, Table 5.2, Table 5.3, Fig. 5.1). This outcome suggests that unlike detritivore, species with other food habits have all substitution rates that are skewed towards lower value. The mean ω of herbivore have lowest value 0.0154, but median ω of herbivore is higher than that of carnivore and haematophagy, which suggest that ω of carnivore and haematophagy skewed towards lower value. The median ω of carnivore and haematophagy is about 48% and 52% lower than the mean ω value, respectively. The median dS of carnivore and haematophagy is around 28% and 19% lower than mean dS value. Whereas, median dN of carnivore and haematophagy is around 71% and 62% lower than mean dN. This implies that the drop in median dN from mean dN was greater than the decrease in median dS from mean dS, which resulted in a fall in the median ω value of carnivore and haematophagy. Therefore, this analysis indicates that species with carnivore and haematophagy food habit have undergone stronger

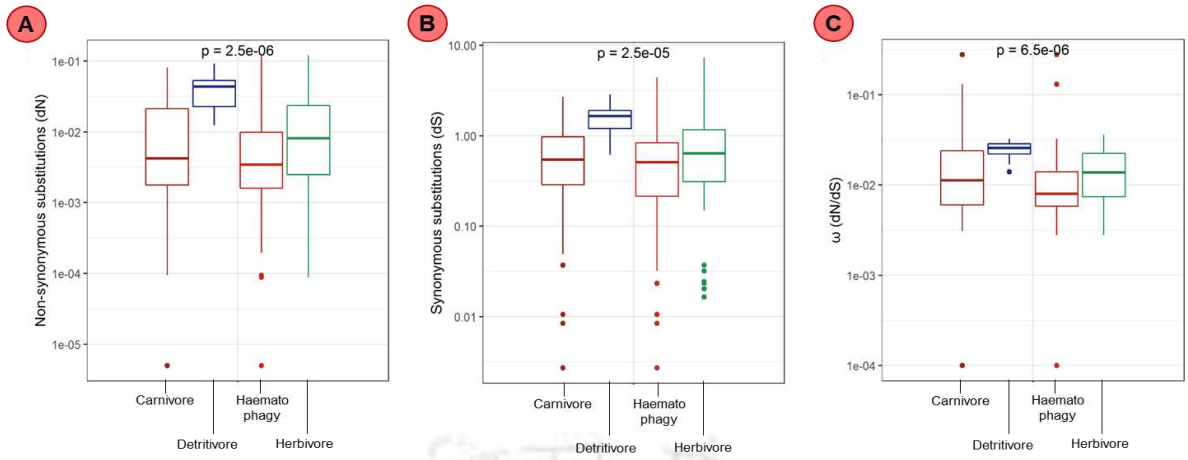


Figure 5.1: Comparisons of for molecular traits (dS, dN and ω) calculated from concatenated mitochondrial protein coding genes among groups in semi-log graph (y-axis is in log scale). (A) Non-synonymous substitution rate (dN) comparisons among carnivore, detritivore, haematophagy and herbivore; (B) Synonymous substitution rate (dS) comparisons among carnivore, detritivore, haematophagy and herbivore; (C) dN/dS (ω) comparisons among carnivore, detritivore, haematophagy and herbivore.

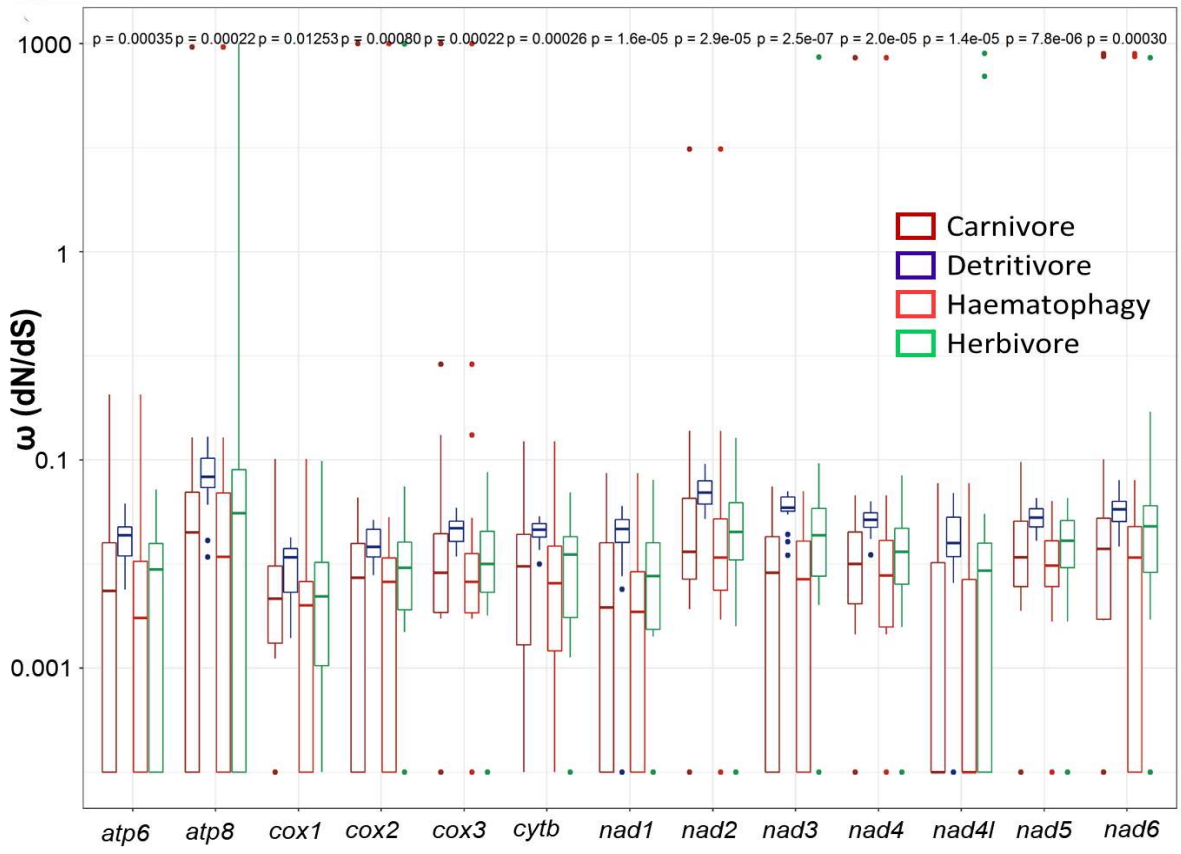


Figure 5.2: Comparisons of the dN/dS (ω) ratios for the 13 mitochondrial protein–coding genes between the carnivore, detritivore, haematophagy and herbivore in a semi-log graph (y-axis is in log scale).

selective constraint as their nonsynonymous substitution rate more skewed towards lower value than synonymous substitution rate. Whereas, median ω of herbivore around 10% lower than mean ω, and median ω of detritivore about 3% higher than mean ω value. The median dS of a detritivore rise by 2.7% compared to the mean dS, while the median dN increases by roughly 3.9% compared to the mean dN. Hence, this observation indicates that in detritivore species the nonsynonymous substitution rate more skewed towards higher value than the synonymous substitution rate that led to elevated ω value. A similar pattern can also be seen when comparing individual mitochondrial genes, as shown in Figure 5.2.

5.3.2 Sequence of discrete-trait (dietary habits) evolution:

We began by estimating transition among several binary traits such as herbivore and carnivore, detritivore, and non-detritivore, and haematophagy and non-haematophagy. We also considered multiple traits such as herbivore, carnivore, and detritivore, as well as polymorphic traits namely carnivore + detritivore and herbivore + detritivore. The best supported model was chosen based on the smallest Akaike information criterion (AIC) value.

In this analysis, a model in which all rates differ (ARD) among states was preferred over other models namely equal rates ER model, stepwise transition MER model and SYM model. In addition, multiple tree transformations were performed and evaluated based on the lowest AIC value to determine the best fit. The ARD model with EB (early burst) transformation fit well for herbivore and carnivore traits (ARD-EB: AIC = 89.99), in which herbivore to carnivore transition is prohibited and carnivore to herbivore transition rate is ~0.0035 (Table 5.4.2, Fig. 5.3 A). The ARD model with lambda transformation fit well for traits such as detritivore and non-detritivore (ARD-lambda: AIC = 70.098), as well as haematophagy and non-haematophagy (ARD-lambda: AIC = 78.551) (Table 5.5.2, 5.6.2). The rate of transition from detritivore to non-detritivore (~0.1174) is higher than the opposite transition rate (~0.004) (Fig.

5.3 B). Whereas, the transition rate from haematophagy to non- haematophagy (~0.00301) is lower than the opposite transition rate (~0.02625) (Fig. 5.3 C).

When multiple and polymorphic traits are taken into account, the ARD model similarly fits better than other models. The multiple traits dataset supported lambda transformation (ARD-lambda: AIC = 106.681) than without or other transformations (Table 5.7.2). Whereas, the polymorphic traits dataset does not show variation in fit after transformation (ARD: AIC = 138.95). No transition has been detected between herbivore and carnivore through both these approaches. When considering multiple states, the transition from herbivore to detritivore is prohibited but detritivore to herbivore transition is present. The rate of transition from detritivore to carnivore is higher than the opposite transition (Fig. 5.4 A). When considering polymorphic states, transition from herbivore to detritivore + herbivore to detritivore prohibited (Fig. 5.4 B). In addition, we have done stochastic character mapping on the tree based on aforementioned best models depicted in Figure 5.3 D-F (binary states), Figure 5.4 C (multiple states), and Figure 5.4 D (polymorphic states).

5.3.3 Discrete Trait-Dependent Diversification:

BiSSE model: Speciation and extinction in BiSSE (binary state speciation and extinction) follow a birth–death process, with the rate of speciation and extinction varying with a binary state that evolves using a continuous-time Markov process[31]. We fitted 7 different model and analysed 3 different scenarios under BISSE model namely, a) Herbivore and Carnivore b) Detritivore and Non-detritivore c) Haematophagy and Non-haematophagy. In the first case, equal extinction rate of herbivore and carnivore regarded as best fitted model and herbivore (net.div1 = 0.26) showed higher net diversification rate than carnivore (net.div0 = 0.19) (Table 5.8). In the second scenario, the best fitted model shows an equal rate of speciation and extinction for both non-detritivore and detritivore lineages, implying that the net diversification rate of both detritivore and non-detritivore lineages is equivalent (net.div = 0.224). Although,

Table 5.4.1: Model comparison using two traits (herbivore and carnivore)

Model	params	lnL	AIC	AICc
ARD	6	-39.92	91.833	92.633
ER	1	-46.91	95.812	95.849
SYM	3	-43.31	92.611	92.833
MER	2	-61.21	126.43	126.54

Table 5.4.2: ARD model with different transformation (herbivore and carnivore)

transform	params	lnL	AIC	AICc
EB	7	-38	89.99	91.067
lambda	7	-39.44	92.883	93.96
kappa	7	-38.91	91.829	92.906
delta	7	-38.76	91.52	92.597

Table 5.5.1: Model comparison using two traits (detritivore and non-detritivore)

Model	params	lnL	AIC	AICc
ARD	2	-33.17	70.339	70.449
ER	1	-37.74	77.486	77.523
SYM	1	-37.74	77.486	77.523
MER	1	-37.74	77.486	77.523

Table 5.5.2: ARD model with different trans- formation (detritivore and non-detritivore)

transform	params	lnL	AIC	AICc
EB	3	-32.34	70.677	70.899
kappa	3	-32.53	71.058	71.28
lambda	3	-32.05	70.098	70.32
delta	3	-32.11	70.22	70.443

Table 5.6.1: Model comparison using two traits (haematophagy and non-haematophagy)

Model	params	lnL	AIC	AICc
ARD	2	-40.23	84.465	84.575
ER	1	-41.56	85.116	85.152
SYM	1	-41.56	85.116	85.152
MER	1	-41.56	85.116	85.152

Table 5.6.2: ARD model with different transformation (haematophagy and non-haematophagy)

transform	params	lnL	AIC	AICc
EB	3	-39.39	84.774	84.996
lambda	3	-36.28	78.551	78.773
kappa	3	-38.82	83.648	83.87
delta	3	-39.48	84.956	85.178

sTable 5.7.1: Model comparison using three traits (herbivore, carnivore, and detritivore)

Model	params	lnL	AIC	AICc
ARD	6	-49.0816	110.1633	110.9633
ER	1	-58.4287	118.8574	118.8937
MER	2	-56.5897	117.1794	117.2895
SYM	3	-54.9879	115.9758	116.1981

Table 5.7.2: ARD model with different transformation (herbivore, carnivore, and detritivore)

transform	params	lnL	AIC	AICc
EB	7	-113.99	241.989	243.066
lambda	7	-46.34	106.681	107.757
kappa	7	-48.574	111.148	112.225
delta	7	-48.206	110.412	111.489

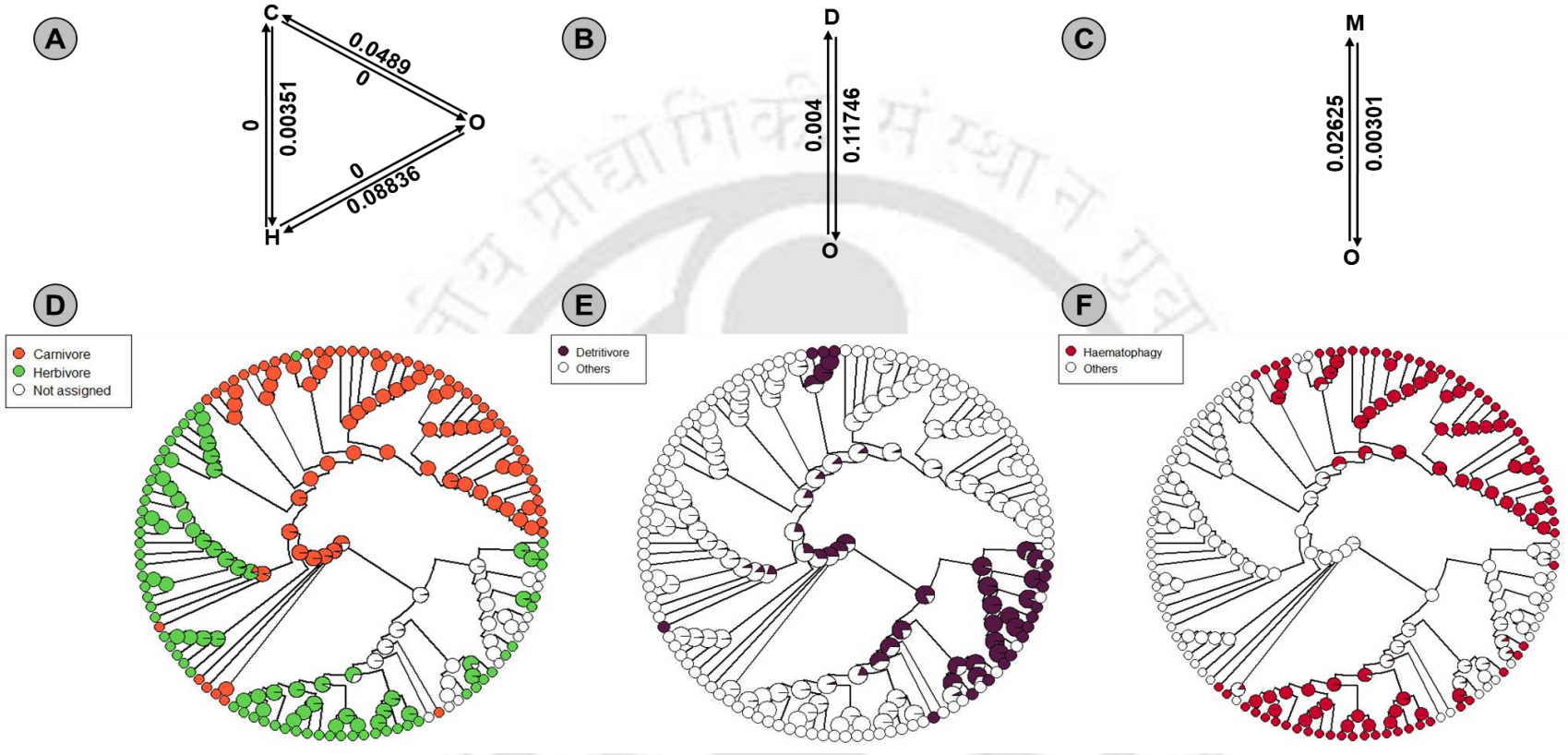


Figure 5.3: Macroevolutionary dynamics of food habit of Diptera considering two states. (A-C) The best fitted ARD model describing transition between states, C: Carnivore, H: Herbivore, D: Detritivore, M: Haematophagy, O: others; zero signifies no transition between two states. (D-F) 1000 Stochastic character mapping into the tree with the best fitted model.

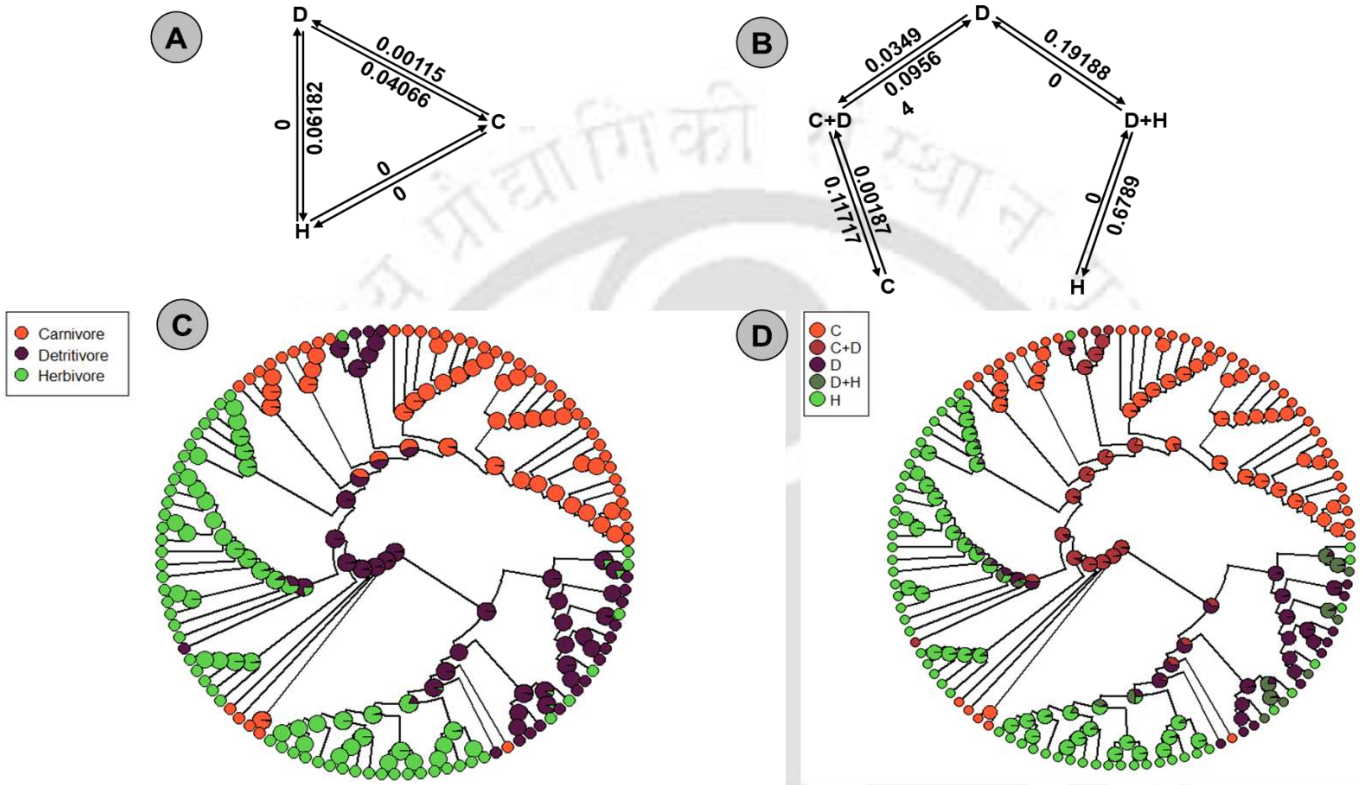


Figure 5.4: Macroevolutionary dynamics of food habit of Diptera considering multiple (A,C) and polymorphic states (B,D). (A) The best fitted ARD model describing transition between states, C: Carnivore, H: Herbivore, D: Detritivore; zero signifies no transition between two states. (B) The best fitted ARD model describing transition between states, C+D: polymorphic states between Carnivore and Detritivore, D+H: polymorphic states between Detritivore and Herbivore. (C) 1000 Stochastic character mapping of multiple states into the tree with the best fitted model. (D) 1000 Stochastic character mapping of polymorphic states into the tree with the best fitted model.

Although the second-best model which have <2 ΔAIC, is the equal extinction rate model and detritivore have higher diversification rate than non-detritivore lineages (Table 5.9). In the third scenario, the best fitted model had an equal extinction rate of haematophagy and non-haematophagy lineages, and the net diversification rate of haematophagy (net.div1 = 0.26) lineages was higher than that of non-haematophagy lineages (net.div0 = 0.19) (Table 5.10).

Further we performed Bayesian analysis of best fitted models of a) Herbivore and Carnivore and b) Haematophagy and Non-haematophagy with MCMC chain runs for 10000 generations in each case. The posterior density plot in Figure 5.5 is consistent with earlier likelihood analyses, indicating that herbivore and haematophagy lineages have a greater net diversification rate.

HiSSE model: The HiSSE (Hidden State Speciation and Extinction) model implies that "hidden" states are associated to each observed state in the model and have possibly different diversification dynamics and transition rates than the observed states in independence[32]. We compared the fit of 24 models of diversification in the HiSSE framework to our trait data coded as binary, similar to BiSSE model. Furthermore, there is another trait linked to hidden states A and B that is not visible[32]. The results indicate that the CID-2 model fit better in the first scenario of herbivore (1) and carnivore (0) than other alternative models with low and equal diversification rates (net.div = 2.06×10^{-9} lineage/myr) of states without hidden states (0A and 1A). Whereas, states (0B and 1B) in which hidden state is present have equal and higher diversification rates (0.24 lineage/myr).

In second case of detritivores (1) and non-detritivores (0), the best model suggests character-dependent diversification, which corresponds to four separate net turnover rates, and equal rate of transition between states was allowed. The result suggests highest diversification rate of state 1A (net.div = 0.351) followed by 0B (net.div = 0.273), 1B (net.div = 0.185), and state 0A

Table 5.8: BiSSE Model Fitting for Herbivore (1) and Carnivore (0), best model based on ΔAIC denoted in different colour.

Model	Df	lnLik	AIC	ΔAIC	net.div0	net.div1	q01	q10
full	6	-295.25	602.5	3.68	0.203946	0.19869	0.00647	0.02036
Equal_speciation	5	-294.59	599.18	0.36	0.152647	0.251628	0.01703	0.00631
Equal_extinction	5	-294.41	598.82	0	0.190049	0.26291	0.02182	0.00559
Equal_speciation_extinction	4	-295.88	599.76	0.94	0.224464	0.224464	0.00581	0.02397
Equal_transition	5	-295.02	600.04	1.22	0.14854	0.263678	0.01355	0.01355
No_transition1_0	5	-296.11	602.22	3.4	-3.6937	0.270775	0.02476	0.00000
No_transition0_1	5	-299.77	609.55	10.73	0.196618	0.210526	0.00000	0.03140

Table 5.9: BiSSE Model Fitting for Detritivore (1) and Non-detritivore (0); with the best model based on ΔAIC denoted in different colour.

Model	Df	lnLik	AIC	ΔAIC	net.div0	net.div1	q01	q10
full	6	-305.62	623.24	3.55	0.205603	0.257524	0.00739	0.14021
Equal_speciation	5	-305.75	621.49	1.8	0.203227	0.237997	0.00705	0.14170
Equal_extinction	5	-305.64	621.28	1.59	0.210876	0.250573	0.00759	0.13507
Equal_speciation_extinction	4	-305.85	619.69	0	0.224472	0.224472	0.00736	0.13130
Equal_transition	5	-310	629.99	10.3	0.225966	0.082303	0.04079	0.04079
No_transition1_0	5	-314.6	639.19	19.5	0.23785	-0.09084	0.05027	0.00000
No_transition0_1	5	-306.4	622.81	3.12	0.196675	0.250904	0.00000	0.14829

Table 5.10: BiSSE Model Fitting for Haematophagy (1) and Non-haematophagy (0); best model based on ΔAIC denoted in different colour.

Model	Df	lnLik	AIC	ΔAIC	net.div0	net.div1	q01	q10
full	6	-313.04	638.09	2	0.195446	0.269772	0.03411	0.00753
Equal_speciation	5	-313.72	637.44	1.35	0.173533	0.247696	0.02877	0.00903
Equal_extinction	5	-313.04	636.09	0	0.195402	0.26979	0.03411	0.00758
Equal_speciation_extinction	4	-314.34	636.68	0.59	0.224469	0.224469	0.03483	0.00771
Equal_transition	5	-314.01	638.02	1.93	0.166329	0.271377	0.02480	0.02480
No_transition1_0	5	-313.67	637.34	1.25	0.196509	0.273046	0.03678	0.00000
No_transition0_1	5	-319.36	648.71	12.62	0.141538	0.256121	0.00000	0.06697

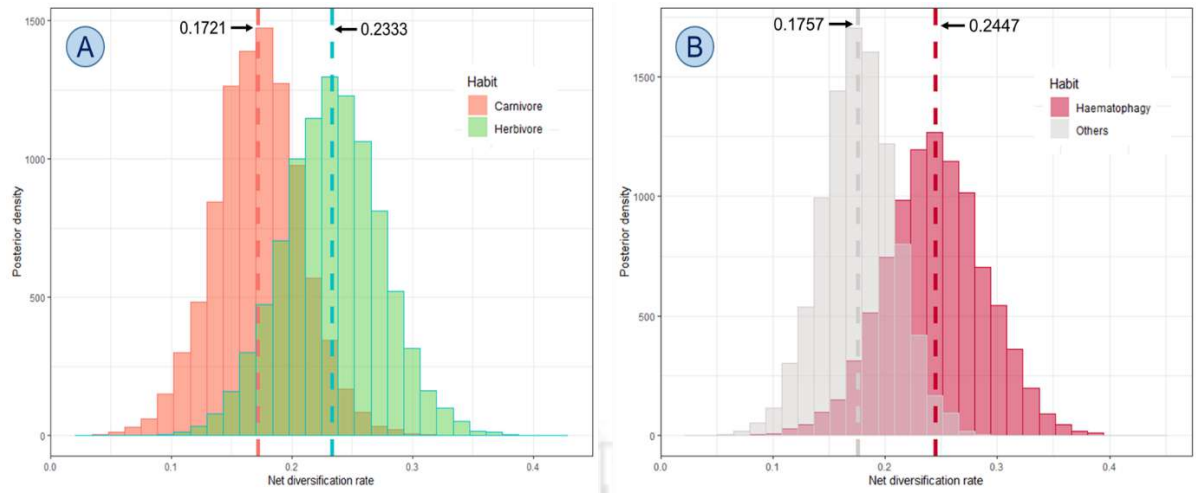


Figure 5.5: Posterior density plot of net diversification rate from best (equal extinction rate) BiSSE model of two different binary states Carnivore vs Herbivore (A) and Haematophagy vs others (B). The mean value of net diversification rate denoted by the vertical dashed line.

had the lowest diversification rate (net.div = 2.06×10^{-9}). This implies that trait detritivore without hidden state has highest diversification rate while trait non-detritivore without hidden state has lowest diversification rate.

The best model in the third scenario of haematophagy (1) and non-haematophagy (0) implies character-dependent diversification, in which the rate of diversification was identical for two states, 0A and 1A (net.div = 0.252), and transition between states was allowed except for the 0B and 1B states, which were prohibited. State 1B has the highest diversification rate (net.div = 0.276), while state 0B has the lowest diversification rate (net.div = 2.06×10^{-9}). These analyses displayed that, apart from the first case of herbivore and carnivore, hisse model supported four states models with varying diversification rates in the other two situations. It was previously reported that when diversification rates are highly heterogeneous across the phylogeny, BiSSE often rejects the null model in favour of a trait-dependent diversification model, even if diversification is not trait-dependent[30,40].

Table 5.11: HiSSE Model Fitting for Herbivore and Carnivore, with the best model based on Δ AIC and Akaike weights (aic.w) denoted in different colour.

Model	logL	AIC	np	delta.aic	aic.w
BiSSE: all free	-276.162	564.3248	6	4.608736	0.020637
BiSSE: ε0=ε1	-276.264	562.527	5	2.810989	0.050701
BiSSE: q's equal	-276.876	563.7528	5	4.036722	0.027469
BiSSE: ε0=ε1, q's equal	-277.374	562.7477	4	3.031645	0.045405
CID-2: q's equal	-275.858	561.716	5	1.999998	0.076053
CID-2: ε's, q's equal	**-275.858**	**559.716**	**4**	**0**	**0.206734**
CID-4: q's equal	-275.352	568.7043	8	8.988275	0.00231
CID-4: ε's equal, q's equal	-275.356	562.7112	5	2.995142	0.046241
HiSSE: q's equal	-275.082	568.1633	9	8.447294	0.003028
HiSSE: ε's equal, q's equal	-275.042	562.0841	6	2.368081	0.063269
HiSSE: τ0A=τ1A=τ0B, ε0A=ε1A=ε0B, q's equal	-277.212	564.425	5	4.708945	0.019628
HiSSE: ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-277.616	563.2323	4	3.516208	0.035635
HiSSE: q0B1B=0, q1B0B=0, all other q's equal	-273.726	565.4517	9	5.735669	0.011747
HiSSE: τ0A=τ1A=τ0B, ε's equal, q's equal	-275.065	562.1296	6	2.413599	0.061845
HiSSE: τ0A=τ1A=τ0B, ε0A=ε1A=ε0B, q0B1B=0,q1B0B=0, all other q's equal	-277.594	565.1886	5	5.472511	0.013399
HiSSE: τ0A=τ1A=τ0B, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-277.61	563.2209	4	3.504824	0.035838
HiSSE: τ0A=τ0B, ε0A=ε0B, q's equal	-276.876	567.7528	7	8.036722	0.003718
HiSSE: τ0A=τ0B, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-277.374	564.7477	5	5.031642	0.016703
HiSSE: τ0A=τ0B, ε0A=ε0B, q0B1B=0, q1B0B=0, all other q's equal	-276.852	567.7045	7	7.988462	0.003808
HiSSE: τ0A=τ0B, ε's equal, q's equal	-277.309	564.6188	5	4.902791	0.017815
HiSSE: τ0A=τ1A, ε0A=ε1A, q0B1B=0, q1B0B=0, all other q's equal	-275.038	564.0756	7	4.359512	0.023375
HiSSE: τ0A=τ1A, ε's equal, q's equal	-275.043	560.0858	5	0.369716	0.171842
HiSSE: τ0A=τ1A, ε0A=ε1A, q's equal	-276.456	566.9121	7	7.196016	0.00566
HiSSE: τ0A=τ1A, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-276.575	563.1496	5	3.433557	0.037139

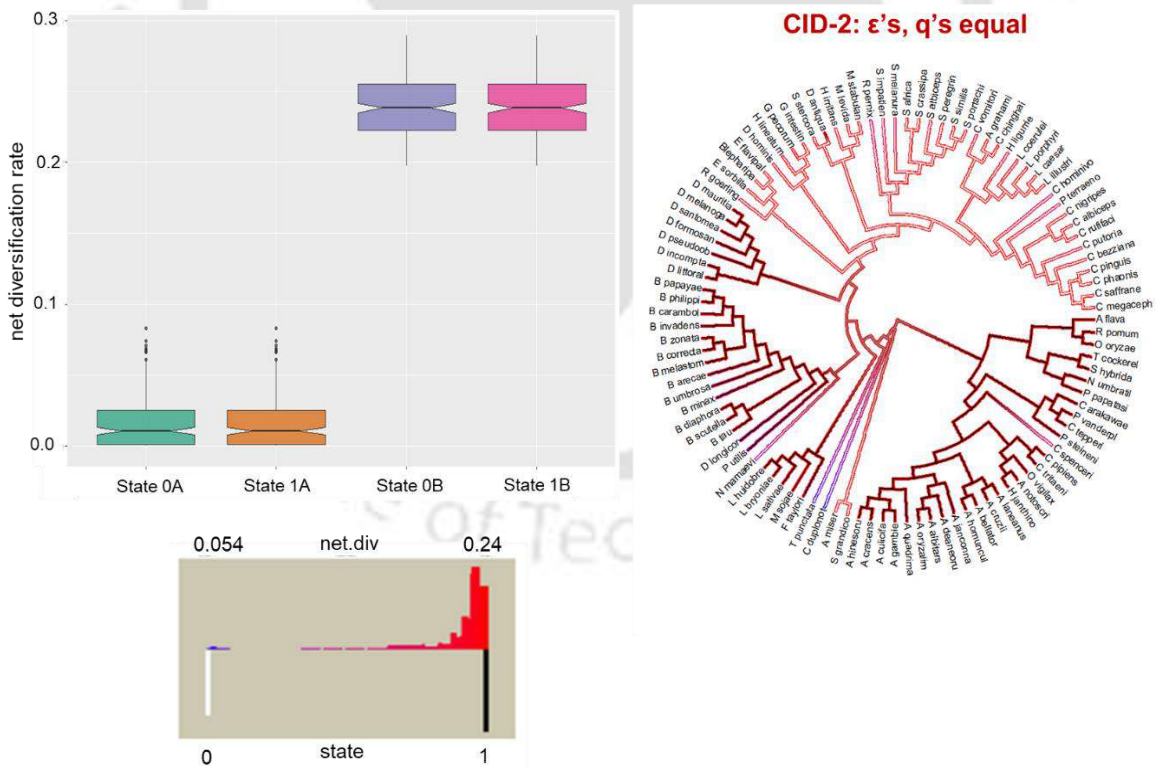


Figure 5.6: Character reconstructions of states and net diversification rates under best HiSSE model (CID-2: ε's, q's equal) for trait Herbivore (1) and Carnivore (0).

Table 5.12: HiSSE Model Fitting for Detritivores and Non-Detritivores, with the best model based on Δ AIC and Akaike weights (wi) denoted in different colour.

Model	logL	AIC	np	delta.aic	aic.w
BiSSE: all free	-305.619	623.2384	6	2.53167	0.101938
BiSSE: ε0=ε1	-305.64	621.2802	5	0.573457	0.271367
BiSSE: q's equal	-309.995	629.9897	5	9.282988	0.003486
BiSSE: ε0=ε1, q's equal	-311.758	631.5154	4	10.80869	0.001626
CID-2: q's equal	-309.044	628.0886	5	7.381907	0.009018
CID-2: ε's, q's equal	-309.044	626.0886	4	5.381907	0.024514
CID-4: q's equal	-308.232	634.4644	8	13.75766	0.000372
CID-4: ε's equal, q's equal	-308.232	628.4644	5	7.757657	0.007474
HiSSE: q's equal	-306.797	631.5946	9	10.88782	0.001563
HiSSE: ε's equal, q's equal	-308.799	629.5982	6	8.891456	0.00424
HiSSE: τ0A=τ1A=τ0B, ε0A=ε1A=ε0B, q's equal	-308.233	626.4663	5	5.75961	0.020295
HiSSE: ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-311.872	631.7432	4	11.03643	0.001451
HiSSE: q0B1B=0, q1B0B=0, all other q's equal					
HiSSE: τ0A=τ1A=τ0B, ε's equal, q's equal	**-304.353**	**620.7067**	**6**	**0**	**0.361478**
HiSSE: τ0A=τ1A=τ0B, ε0A=ε1A=ε0B, q0B1B=0,q1B0B=0, all other q's equal	-307.645	625.2905	5	4.583792	0.036536
HiSSE: τ0A=τ1A=τ0B, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-308.224	624.4471	4	3.740334	0.055703
HiSSE: τ0A=τ0B, ε0A=ε0B, q's equal	-309.995	633.9897	7	13.28299	0.000472
HiSSE: τ0A=τ0B, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-311.758	633.5154	5	12.80869	0.000598
HiSSE: τ0A=τ0B, ε0A=ε0B, q0B1B=0, q1B0B=0, all other q's equal	-306.782	627.5632	7	6.856459	0.011728
HiSSE: τ0A=τ0B, ε's equal, q's equal	-307.531	625.0624	5	4.355649	0.040951
HiSSE: τ0A=τ1A, ε0A=ε1A, q0B1B=0, q1B0B=0, all other q's equal	-308.212	630.4234	7	9.716666	0.002806
HiSSE: τ0A=τ1A, ε's equal, q's equal	-308.799	627.5982	5	6.891463	0.011524
HiSSE: τ0A=τ1A, ε0A=ε1A, q's equal	-307.325	628.6502	7	7.943467	0.006811
HiSSE: τ0A=τ1A, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-308.063	626.1268	5	5.420072	0.024051

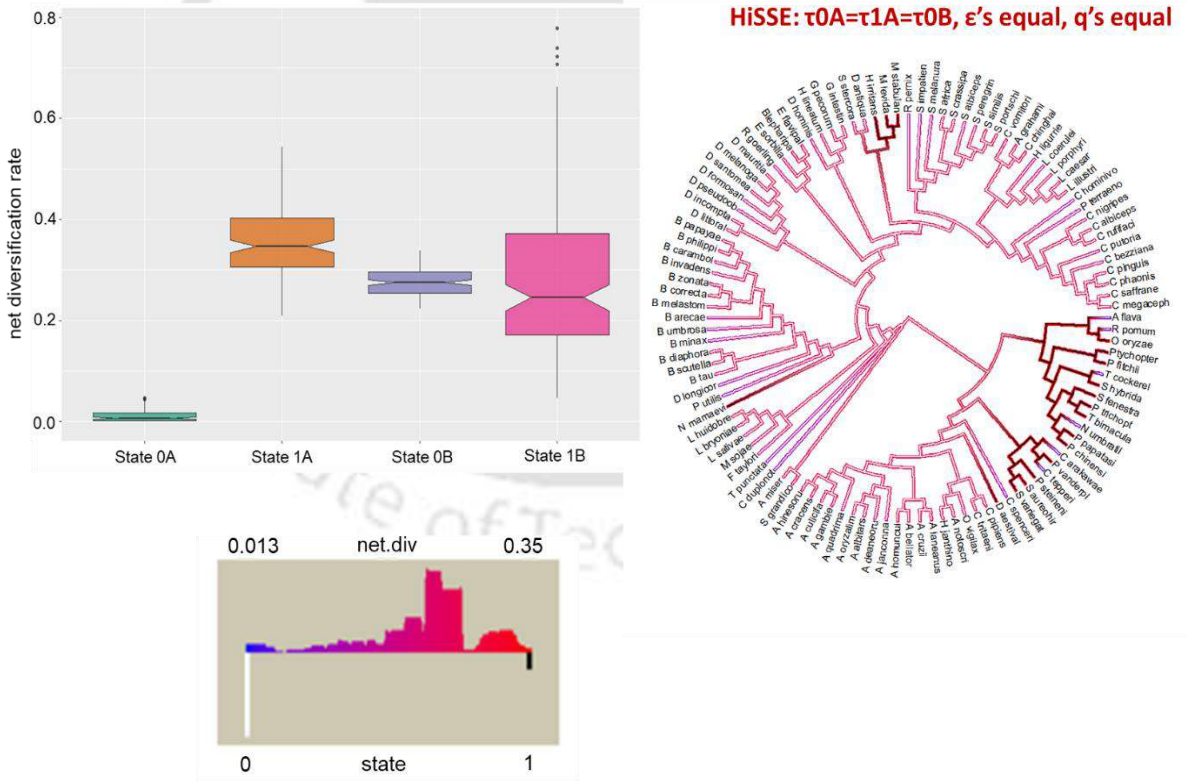


Figure 5.7: Character reconstructions of states and net diversification rates under best HiSSE model (τ0A=τ1A=τ0B, ε's equal, q's equal) for trait Detritivores (1) and Non-Detritivores (0).

Table 5.13: HiSSE Model Fitting for Haematophagy and Non-haematophagy, with the best model based on Δ AIC and Akaike weights (wi) denoted in different colour.

Model	logL	AIC	np	delta.aic	aic.w
BiSSE: all free	-313.044	638.0883	6	5.492962	0.012125
BiSSE: ε0=ε1	-313.044	636.0883	5	3.492962	0.032958
BiSSE: q's equal	-314.01	638.0198	5	5.424417	0.012547
BiSSE: ε0=ε1, q's equal	-314.222	636.4447	4	3.849393	0.027578
CID-2: q's equal	-312.45	634.9002	5	2.304828	0.059699
CID-2: ε's, q's equal	-312.45	632.9002	4	0.304828	0.162279
CID-4: q's equal	-311.688	641.3764	8	8.781049	0.002342
CID-4: ε's equal, q's equal	-312.48	636.96	5	4.364629	0.021315
HiSSE: q's equal	-311.453	640.9056	9	8.310218	0.002964
HiSSE: ε's equal, q's equal	-311.622	635.2433	6	2.647962	0.050287
HiSSE: τ0A=τ1A=τ0B, ε0A=ε1A=ε0B, q's equal	-314.391	638.7821	5	6.186792	0.008571
HiSSE: ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-314.6	637.2008	4	4.605496	0.018897
HiSSE: q0B1B=0, q1B0B=0, all other q's equal	-309.851	637.7029	9	5.107598	0.014701
HiSSE: τ0A=τ1A=τ0B, ε's equal, q's equal	-311.276	634.5526	6	1.957248	0.07103
HiSSE: τ0A=τ1A=τ0B, ε0A=ε1A=ε0B, q0B1B=0,q1B0B=0, all other q's equal	-313.204	636.4076	5	3.812276	0.028095
HiSSE: τ0A=τ1A=τ0B, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-313.466	634.9319	4	2.336583	0.058759
HiSSE: τ0A=τ0B, ε0A=ε0B, q's equal	-313.66	641.3195	7	8.724128	0.00241
HiSSE: τ0A=τ0B, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	-313.914	637.8284	5	5.233032	0.013808
HiSSE: τ0A=τ0B, ε0A=ε0B, q0B1B=0, q1B0B=0, all other q's equal	-312.954	639.9073	7	7.311929	0.004883
HiSSE: τ0A=τ0B, ε's equal, q's equal	-313.294	636.5878	5	3.992427	0.025675
HiSSE: τ0A=τ1A, ε0A=ε1A, q0B1B=0, q1B0B=0, all other q's equal	-311.42	636.84	7	4.244714	0.022632
HiSSE: τ0A=τ1A, ε's equal, q's equal	-311.622	633.2433	5	0.647959	0.136695
HiSSE: τ0A=τ1A, ε0A=ε1A, q's equal	-311.507	637.0133	7	4.418002	0.020754
HiSSE: τ0A=τ1A, ε's equal, q0B1B=0, q1B0B=0, all other q's equal	**-311.298**	**632.5953**	**5**	**0**	**0.188997**

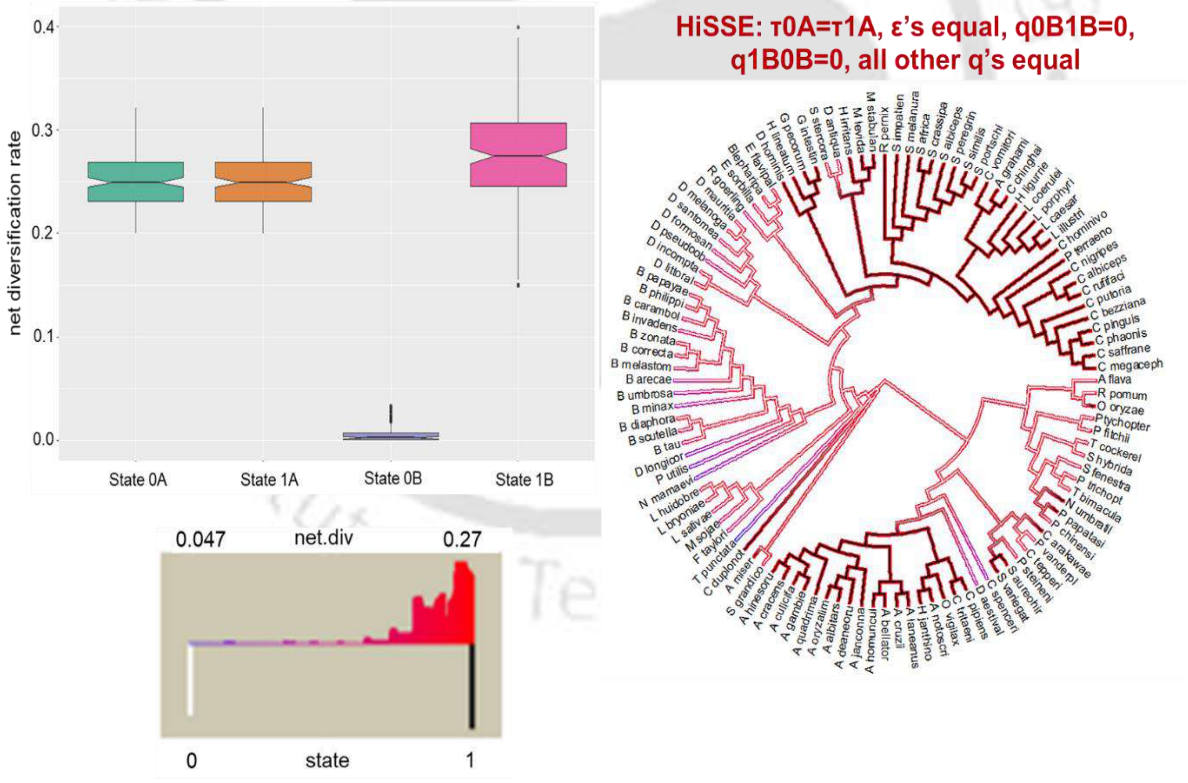


Figure 5.8: Character reconstructions of states and net diversification rates under best HiSSE model (τ0A=τ1A, ε's equal, q0B1B=0, q1B0B=0, all other q's equal) for trait Haematophagy (1) and Non-haematophagy (0).

Table 5.14: MuSSE Model Fitting for Herbivore (H), Carnivore (C), Herbivore and Carnivore (HC), others (0); best model based on ΔAIC denoted in different colour.

Model	Df	lnLik	AIC	ΔAIC	ChiSq	Pr(>|Chi|)
minimal	6	-331.55	675.11	2.38		
vary.lambda	8	-331.32	678.65	5.92	0.4615	0.7939
int.lambda	9	-327.37	672.73	0	8.3718	0.0389 *
vary.mu	8	-332.04	680.08	7.35	-0.9699	1.0000
vary.lam.mu	10	-330.88	681.77	9.04	1.3390	0.8547

Table 5.15: MuSSE Model Fitting for Herbivore (H), Carnivore (C), Detritivore (D), Herbivore and Carnivore (HC), Herbivore and Detritivore (HD), Carnivore and Detritivore (CD), others (0); best model based on ΔAIC denoted in different colour.

Modelc	Df	lnLik	AIC	ΔAIC	ChiSq	Pr(>|Chi|)
minimal	8	-363.29	742.58	0		
vary.lambda	11	-360.67	743.34	0.76	5.2478	0.15453
int.lambda	14	-357.48	742.97	0.39	11.6173	0.07107
vary.mu	11	-362.14	746.27	3.69	2.3104	0.51052
vary.lam.mu	14	-360.68	749.35	6.77	5.2339	0.51419

Table 5.16: MuSSE Model Fitting for Herbivore (H), Carnivore (C), Detritivore (D), Haematophagy (M), others (0); best model based on ΔAIC denoted in different colour.

Model	Df	lnLik	AIC	ΔAIC	ChiSq	Pr(>|Chi|)
minimal	10	-403.52	827.05	3.04		
vary.lambda	14	-398.01	824.01	0	11.0366	0.02616 *
int.lambda	20	-396.45	832.89	8.88	14.1549	0.16604
vary.mu	14	-403.37	834.75	10.74	0.3043	0.98954
vary.lam.mu	18	-398.9	833.79	9.78	9.2585	0.32096

MuSSE model: For MuSSE analysis, we fitted 5 different models and created 3 different scenarios namely, a) Herbivore (H), Carnivore (C), Herbivore and Carnivore (HC), others (0) b) Herbivore (H), Carnivore (C), Detritivore (D), Herbivore and Carnivore (HC), Herbivore and Detritivore (HD), Carnivore and Detritivore (CD), others (0) c) Herbivore (H), Carnivore (C), Detritivore (D), Haematophagy (M), others (0). The interactions between speciation of different traits were found as the best fitted model for the first scenario, while varying speciation rates were identified as the best fitted model for the third case. Whereas, the second scenario supported minimal model with constant speciation and extinction rate between different traits.

5.3.4 Evolution pattern of quantitative molecular traits:

Correlation between food habit and molecular traits through Phylogenetic ANOVA: The results of the Phylogenetic ANOVA revealed that only molecular trait non-synonymous substitution rate (dN) was statistically significant with detritivore vs non-detritivore and haematophagy vs non-haematophagy (result from geiger discussed here). The detritivore lineages have statistically higher dN than non-detritivore lineages regardless of phylogenetic placement ($F = 13.385$, $p = 0.09$ (given phylogeny)). The haematophagy lineages have statistically lower dN than non- haematophagy lineages regardless of phylogenetic placement ($F = 20.823$, $p = 0.07$ (given phylogeny)). Without concerning of phylogeny along with dN the synonymous substitution rate (dS) also displayed significant support. Higher substitution rate for detritivores (dS: $F = 7.2062$, $p = 0.008$ (without phylogeny); dN: $F = 13.385$, $p = 0.0003$ (without phylogeny)). Higher substitution rate for herbivores than carnivores (dS: $F = 2.807$, $p = 0.06$ (without phylogeny); dN: $F = 3.4875$, $p = 0.03$ (without phylogeny)). Lower substitution rate for haematophagy than non-haematophagy (dS: $F = 10.839$, $p = 0.001$ (without phylogeny); dN: $F = 20.823$, $p = 0.00001$ (without phylogeny)).

Disparity through time (DTT) plots of molecular traits: The disparity through time (DTT) approach introduced by Harmon et al. (2003) for the understanding of trait disparity within and between clade[34]. The DTT plots of quantitative molecular traits (dS, dN and ω) show inconsistent with the BM null model throughout most of the lineage's history (Fig. 5.9). The Disparity Index (DI) is calculated by adding the deviations of the empirical DTT curve from the median of the null model simulations. The positive value of the DI (dS: 0.244, dN: 0.507, ω: 0.325) indicates that a clade contains a large amount of variation, and that clades substantially overlap in trait space but contain significant proportions of the total variation of the group[43]. However, the differences between these profiles were not statistically significant in any of the cases (all p-values >0.05). The DTT curve of dS show more or less consistent over the null model. Whereas, dN shows that the observed DTT curves deviate significantly from BM expectations beyond the 95% confidence interval towards the base, with little fluctuation in recent periods, indicating that disparity mainly accumulated early in history[44]. The ω displays fairly consistent DTT curves with the BM null model near the base but substantial variance in the recent time, implying a rapid acceleration in trait diversification in recent history[44].

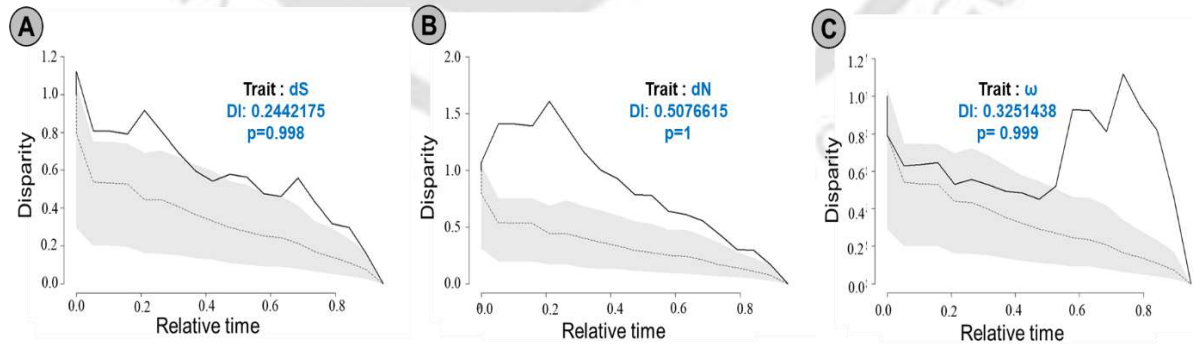


Figure 5.9: The disparity-through-time (DTT) for molecular traits (solid line: dS (A), dN (B), ω (C)), against a median of 1,000 simulations of the null model of Brownian evolution (dotted line), 95% confidence intervals (grey region). The positive values indicating that disparity is greater than expected.

5.3.5 Mode of molecular-traits evolution:

The geiger analyses: The assessment of molecular trait evolution by fitting three continuous trait evolution models without a pre-defined classification of behavior indicates that the OU model had the best fit in continuous trait analysis (AICw = 0.97 (OU), AICw = 0.02 (BM) for dS; AICw = 0.88 (OU), AICw = 0.08 (BM) for dN and AICw = 1 (OU), AICw = $1.1e^{-10}$ (BM) for ω). Although, the estimated strength of the selection (α = 0.094 for dS, α = 0.077 for dN

Table 5.17: Continuous trait evolution analysis of molecular trait synonymous substitution rate (dS) using Brownian motion, Ornstein-Uhlenbeck, and Early-Burst models; the best model based on ΔAIC denoted in different colour.

Model	free parameters	Model Parameters	root state (z0)	sig.sq	lnL	AIC	ΔAIC	w	
Brownian motion (BM)	2		1.66865	0.144826	-133.581118	271.1622	7.627048	0.021424	
Ornstein-Uhlenbeck (OU)	3	alpha = 0.094363	1.390502	0.2062514	-128.767594	263.5352	0	0.970696	**
Early-Burst (EB)	3	a = -0.000001	1.668651	0.1448285	-133.58117	273.1623	9.627152	0.007881	

Table 5.18: Continuous trait evolution analysis of molecular trait non-synonymous substitution rate (dN) using Brownian motion, Ornstein-Uhlenbeck, and Early-Burst models; the best model based on ΔAIC denoted in different colour.

Model	free parameters	Model Parameters	root state (z0)	sig.sq	lnL	AIC	ΔAIC	w	
Brownian motion (BM)	2		0.039958	9.95E-05	274.246539	-544.4931	4.679103	0.085147	
Ornstein-Uhlenbeck (OU)	3	alpha = 0.077973	0.033671	1.36E-04	277.58609	-549.1722	0	0.883531	*
Early-Burst (EB)	3	a = -0.000001	0.039958	9.96E-05	274.246495	-542.493	6.67919	0.031322	

Table 5.19: Continuous trait evolution analysis of molecular trait dN/dS (ω ratio) using Brownian motion, Ornstein-Uhlenbeck, and Early-Burst models; the best model based on ΔAIC denoted in different colour.

Model	free parameters	Model Parameters	root state (z0)	sig.sq	lnL	AIC	ΔAIC	w	
Brownian motion (BM)	2		0.021343	0.000230759	227.166028	-450.3321	45.85983	1.10E-10	
Ornstein-Uhlenbeck (OU)	3	alpha = 0.374288	0.01871	0.000541969	251.095945	-496.1919	0	1.00E+00	**
Early-Burst (EB)	3	a =-0.000001	0.021343	0.000230763	227.165943	-448.3319	47.86	4.05E-11	

and $\alpha = 0.374$ for ω) is quite low for dS and dN, and intensity of the random fluctuation ($\sigma^2 = 0.206$ (OU), $\sigma^2 = 0.144$ (BM) for dS; $\sigma^2 = 1.36e^{-4}$ (OU), $\sigma^2 = 9.95e^{-5}$ (BM) for dN and $\sigma^2 = 0.00054$ (OU), $\sigma^2 = 0.00023$ (BM) for ω) were higher than BM (Table 5.17, 5.18, 5.19). The fit of OU over BM shows the presence of a stabilising selection pressure around one or more adaptive optima, but it does not entirely rule out a random walk mode of trait evolution. Stabilizing selection under an OU model would drive species to overlap in the trait-space around these molecular traits (dS, dN, ω) fitness optima.

The OUCH analyses: We already demonstrated that molecular traits (dS, dN and ω) vary according to the diverse dietary habits exist in Diptera flies. Herein, the main goal is to determine whether the molecular traits converged around one or more fitness optima for various food habits. To investigate the adaptive evolution of quantitative molecular traits, we implemented and compared the two Hansen models (OU1 and OU2) in conjunction with the Brownian motion (BM) model. The best-fitting model, according to AICc and AICc weight, was the OU1 (7 out of 9) and OU2 (2 out of 9), which was much better than BM. This implies that adaptation of molecular traits based on food habits is not the result of pure drift, but that the major role of selection must be present. The ouch analysis of dS for herbivore vs carnivore lineages reveals the best-fit model of evolution is OU1 (OU1: AICc = 286, AICc-weight = 0.886; OU2: AICc = 290, AICc-weight = 0.112), suggests strong selection ($\alpha = 2.335$) and high drift ($\sigma = 2.314$) with a single optimum trait ($\theta = 1.324$) for both lineages, while the second-best model OU2 has ΔAICc = 4 and lower AICc-weight showing little higher θ for herbivore lineages ($\theta_H = 1.298$, $\theta_C = 1.239$). The analysis of dN shows the best-fit model of evolution is OU1 (OU1: AICc = -527, AICc-weight = 0.859; OU2: AICc = -524, AICc-weight = 0.139), suggests strong selection ($\alpha = 2.427$) and low drift ($\sigma = -0.061$) with a single optimum trait ($\theta = 0.029$) for both lineages, while the second-best model OU2 has ΔAICc < 4 and lower AICc-weight showing little higher θ for herbivore lineages ($\theta_H = 0.027$, $\theta_C = 0.025$). The analysis

with ω also supported OU1 (OU1: AICc = -496, AICc-weight = 0.867; OU2: AICc = -492, AICc-weight = 0.133), suggests very strong selection (α = 7.102) and low drift (σ = -0.101) with a single optimum trait (θ = 0.018) for both lineages, in contrast the second-best model, OU2 had ΔAICc = 4 and lower AICc-weight showing lower θ for herbivore lineages (θ_H = 0.015, θ_C = 0.02) (Table 5.20). The ouch analysis of dS for detritivore vs non-detritivore lineages finds the best-fit model of evolution is OU1 (OU1: AICc = 286, AICc-weight = 0.658; OU2: AICc = 287, AICc-weight = 0.341), suggests strong selection (α = 2.335) and high drift (σ = 2.314) with a single optimum trait (θ = 1.324) for both lineages, while second-best model supported higher optimal trait value for detritivore lineages (θ_D = 1.664, θ_N = 1.133). The analysis of dN shows the best-fit model of evolution is OU2 (OU2: AICc = -528, AICc-weight = 0.542; OU1: AICc = -527, AICc-weight = 0.457) supporting higher optimal trait value of detritivore lineages (θ_D = 0.043, θ_N = 0.021) with strong selection (α = 3) and low drift (σ = -0.064), while the second-best model OU1 supported a single optimal trait value (θ = 0.029) for both lineages. The analysis with ω supported OU1 (OU1: AICc = -496, AICc-weight = 0.683; OU2: AICc = -495, AICc-weight = 0.317) as best fitted model suggests very strong selection (α = 7.102) and low drift (σ = -0.101) with a single optimal trait value (θ = 0.018) and second-best model OU2 display higher optimal trait for detritivore lineages (θ_D = 0.024, θ_N = 0.017) (Table 5.21). A similar analysis of dS for haematophagy vs non-haematophagy lineages reveals that the best-fit model of evolution is OU2 (OU2: AICc = 279, AICc-weight = 0.975; OU1: AICc = - 286, AICc-weight = 0.0251), indicating strong selection (α = 2.98) and high drift (σ = 2.363), where the model supported two different optimal trait value (θ) for the both lineages with lower θ for haematophagy lineages (θ_M = 0.373, θ_N = 1.665). The analysis of dN shows the best-fit model of evolution is OU1 (OU1: AICc = 286, AICc-weight = 0.658; OU2: AICc = 287, AICc-weight = 0.341), indicates strong selection (α = 2.427) and low drift (σ = -0.061) with a single optimal trait value for both lineages (θ = 0.029), while the second-best model, OU2 display lower θ for haematophagy lineages (θ_M = 0.00067, θ_N = 0.04). The analysis with

Table 5.20: Comparison of model fits and parameter values for molecular traits (dS, dN and ω) between herbivore (H) and carnivore (C) species with the best model based on ΔAIC denoted in different colour.

	Model	k	loglik	AICc	ΔAICc	weights	α	σ	θ	θ_C	θ_H
dS	BM	2	-147	298	12	0.002		1.868643	1.681812		
	OU1	3	-140	286	0	0.886	2.335025	2.314611	1.324023		
	OU2	5	-140	290	4	0.112	2.465814	2.343232		1.239756	1.298135
dN	BM	2	259	-514	13	0.001		0.049798	0.039356		
	OU1	3	267	-527	0	0.859	2.42781	-0.0619	0.029477		
	OU2	5	267	-524	3	0.139	2.686459	-0.06331		0.025807	0.027571
ω	BM	2	227	-450	46	0		0.066213	0.021052		
	OU1	3	251	-496	0	0.867	7.102822	-0.10138	0.01862		
	OU2	5	251	-492	4	0.133	7.308523	-0.1023		0.020095	0.015844

Table 5.21: Comparison of model fits and parameter values for molecular traits (dS, dN and ω) between detritivore (D) and non-detritivore (N) species with the best model based on ΔAIC denoted in different colour.

	Model	k	loglik	AICc	ΔAICc	weights	α	σ	θ	θ_D	θ_N
dS	BM	2	-147	298	12	0.002		1.868643	1.681812		
	OU1	3	-140	286	0	0.658	2.335025	2.314611	1.324023		
	OU2	4	-140	287	1	0.341	2.655599	2.381575		1.664672	1.133496
dN	BM	2	259	-514	14	0.001		0.049798	0.039356		
	OU1	3	267	-527	1	0.457	2.42781	-0.0619	0.029477		
	OU2	4	268	-528	0	0.542	3.009462	-0.06477		0.043937	0.021848
ω	BM	2	227	-450	46	0		0.066213	0.021052		
	OU1	3	251	-496	0	0.683	7.102822	-0.10138	0.01862		
	OU2	4	251	-495	1	0.317	7.242945	-0.10191		0.024321	0.017332

Table 5.22: Comparison of model fits and parameter values for molecular traits (dS, dN and ω) between haematophagy (M) and non-haematophagy (N) species with the best model based on ΔAIC denoted in different colour.

	Model	k	loglik	AICc	ΔAICc	weights	α	σ	θ	θ_M	θ_N
dS	BM	2	-147	298	19	0		1.868643	1.681812		
	OU1	3	-140	286	7	0.025	2.335025	2.314611	1.324023		
	OU2	4	-135	279	0	0.975	2.980274	2.363792		0.373854	1.665368
dN	BM	2	-147	298	12	0.002		0.049798	0.039356		
	OU1	3	-140	286	0	0.658	2.42781	-0.0619	0.029477		
	OU2	4	-140	287	1	0.341	3.512371	-0.06352		0.000676	0.040557
ω	BM	2	227	-450	46	0		0.066213	0.021052		
	OU1	3	251	-496	0	0.647	7.102822	-0.10138	0.01862		
	OU2	4	252	-495	1	0.353	7.158557	-0.10128		0.015202	0.021711

ω supported OU1 (OU1: AICc = -496, AICc-weight = 0.647; OU2: AICc = -495, AICc-weight = 0.353), suggests very strong selection (α = 7.102) and low drift (σ = -0.101) with a single optimal trait value (θ = 0.018) for both lineages, while the second-best model, OU2 display lower θ for haematophagy lineages (θ_M = 0.015, θ_N = 0.021) (Table 5.22).

Analysis of continuous trait evolution models using both geiger and OUCH suggests that dS had substantially higher random drift with the best performing one or multiple optima models than dN or ω. As Ornstein-Uhlenbeck (OU) processes have a drift component, raising the parameter for assessing the strength of the random fluctuation also broadens the distribution of final states in OU processes[36]. dN, on the other hand, exhibited significantly reduced driftwith all of the best performing models implying a concentrated distribution of final state. While, ω shown substantially stronger selection strength than dN and dS with all of the best performing models, this indicates a more faster approach to the optimum value as well as a narrower dispersion of phenotypes around the optimum[36].

Outcome from ℓ1ou and SURFACE analyses: The ℓ1ou and SURFACE was used for the detection of multiple adaptive shifts and optima in trait evolution. The ℓ1ou and SURFACE analyses also identified multiple instances of convergence across lineages with adaptive peaks between clades with similar trait phenomenon.

We used two methods to determine the number of model shifts in ℓ1ou (and to assess the robustness of any identified convergent regimes) because the method used to infer trait optima on trees can have a significant impact on results: the widely used Akaike information criterion (AICc) and the more conservative Bayesian information criterion (pBIC)[38]. Using the pBIC approach, in the ℓ1ou analysis the best model (pBIC = 177.0112) recovered 6 adaptive shifts in dS throughout the phylogeny, while best model of AICc (AICc = 95.83) detected 15 adaptive shifts. Using the pBIC technique, 14 adaptive shifts were observed (pBIC = -587.4), whereas the best model of AICc method (AICc = -740.184) identified 24 adaptive shifts for the

molecular trait dN across the phylogeny. The ℓ1ou analysis found 4 adaptive shifts in ω across phylogeny using pBIC method (pBIC = -733.787) while the best model of AICc (AICc = -830.076) detected 16 adaptive shifts.

Table 5.23: ℓ1ou convergence parameters from the best selected regime estimated for the molecular traits (dS, dN and ω)

trait	k	logLik	pBIC	AICc	α	σ^2	γ
dS	6	-42.51	177.0112	-	1.00E-07	0.5412	2.71E+06
dN	14	3.96E+02	-587.404	-	1.06E-07	2.15E-04	1.02E+03
ω	4	402.46	-733.787	-	1.64	2.88E-04	8.76E-05
dS	15	-0.5328	-	95.83482	0.5734907	0.297435	0.25932
dN	24	4.65E+02	-	-740.184	7.01E-01	7.50E-05	5.34E-05
ω	16	4.67E+02	-	-830.076	1.32E+01	3.78E-04	1.44E-05

k: Number of shifts, pBIC: Phylogenetic Bayesian information criterion, α: Adaptation rate, σ^2: Variance, γ: Stationary variance.

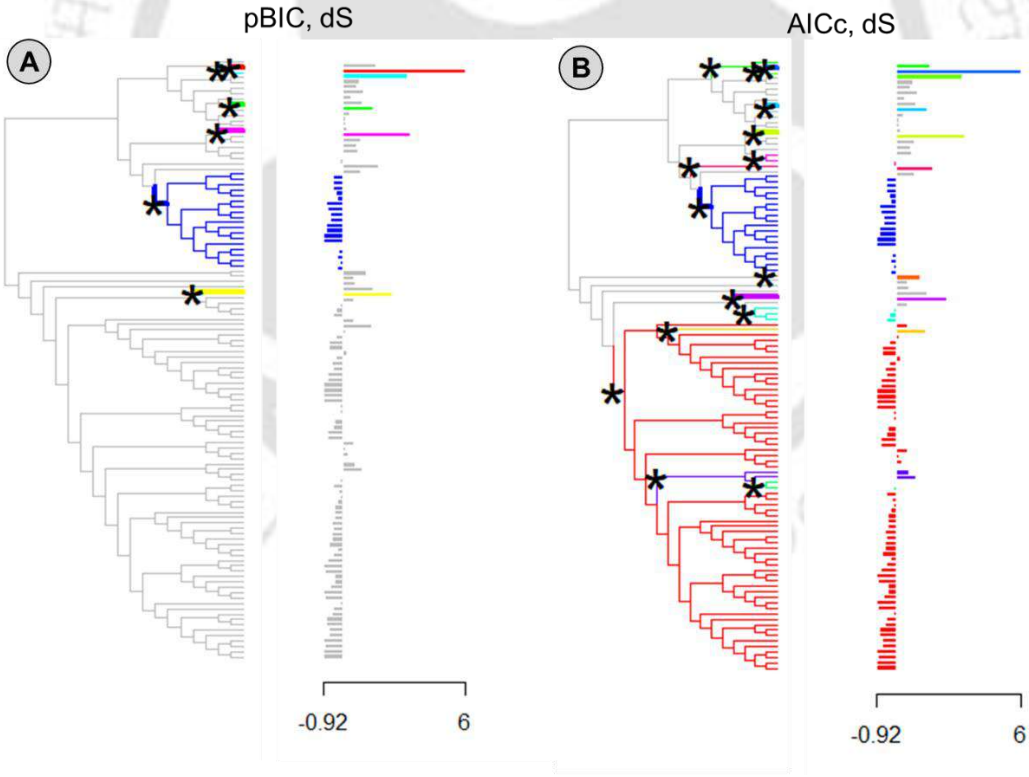


Figure 5.10: Evolutionary shifts (asterisks) in synonymous substitution rate (dS) across Diptera phylogeny using pBIC (A) and AICc (B) methods in ℓ1ou. Coloured branches represent convergent adaptive peaks and gray or black branches represent non-convergent regimes. The bar graphs showing at the right side of phylogenetic tree the individual trait used for analysis.

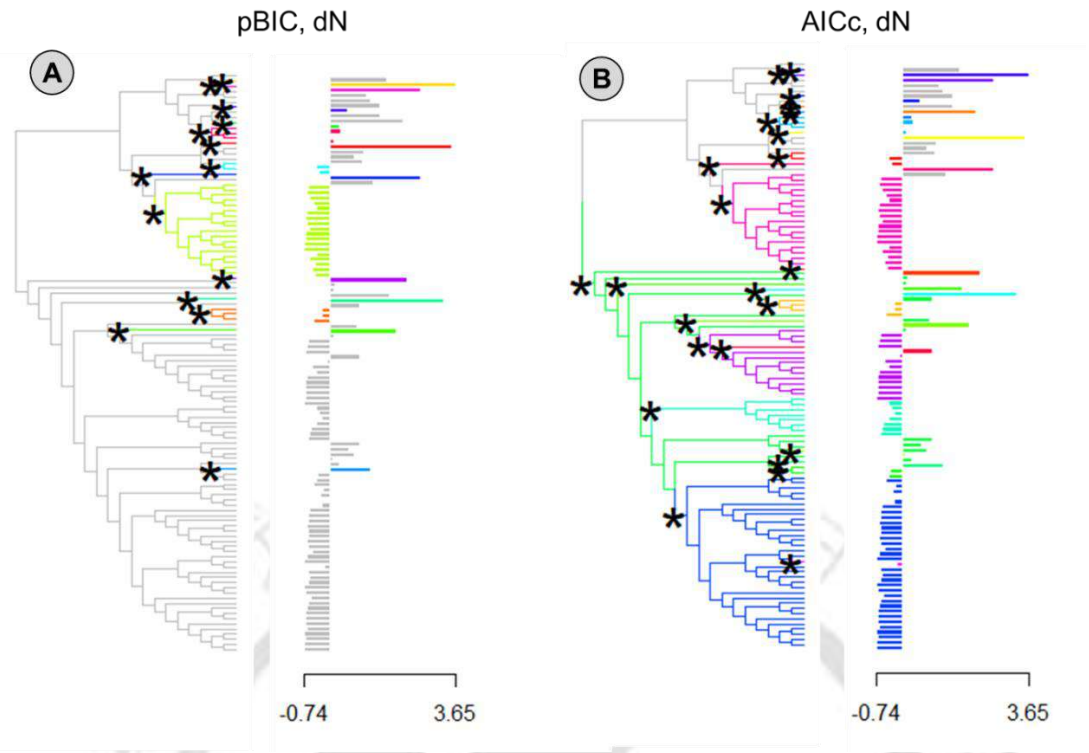


Figure 5.11: Evolutionary shifts (asterisks) in non-synonymous substitution rate (dN) across Diptera phylogeny using pBIC (A) and AICc (B) methods in ℓ1ou. Coloured branches represent convergent adaptive peaks and gray or black branches represent non-convergent regimes. The bar graphs showing at the right side of phylogenetic tree the individual trait used for analysis.

Table 5.24: SURFACE convergence parameters from the best selected regime estimated for the molecular traits (dS, dN and ω)

trait	AICc	k	k′	Δk	c	k′conv	k′_nonconv	c/k	σ^2	α
dS	52.1	20	10	10	16	6	4	0.80	0.010384	0.022074
dN	-791.6	24	14	10	19	9	5	0.79	4.83E-06	0.057591
ω	-880.2	19	7	12	16	4	3	0.84	0.001052	45.68318

k: Number of regime shifts, k′: Number of distinct regime, Δk: Reduction in complexity of the adaptive landscape when accounting for convergence, c: Number of shifts that are towards convergent regimes occupied by multiple lineages, k′conv: Number of convergent regimes reached by multiple shifts, k′_nonconv: Number of nonconvergent regimes, c/k: Proportion of regime shifts that evolved towards convergent regimes, σ^2: Rate of stochastic evolution (one parameter per trait), α: Rate of adaptation to optima (one parameter per trait).

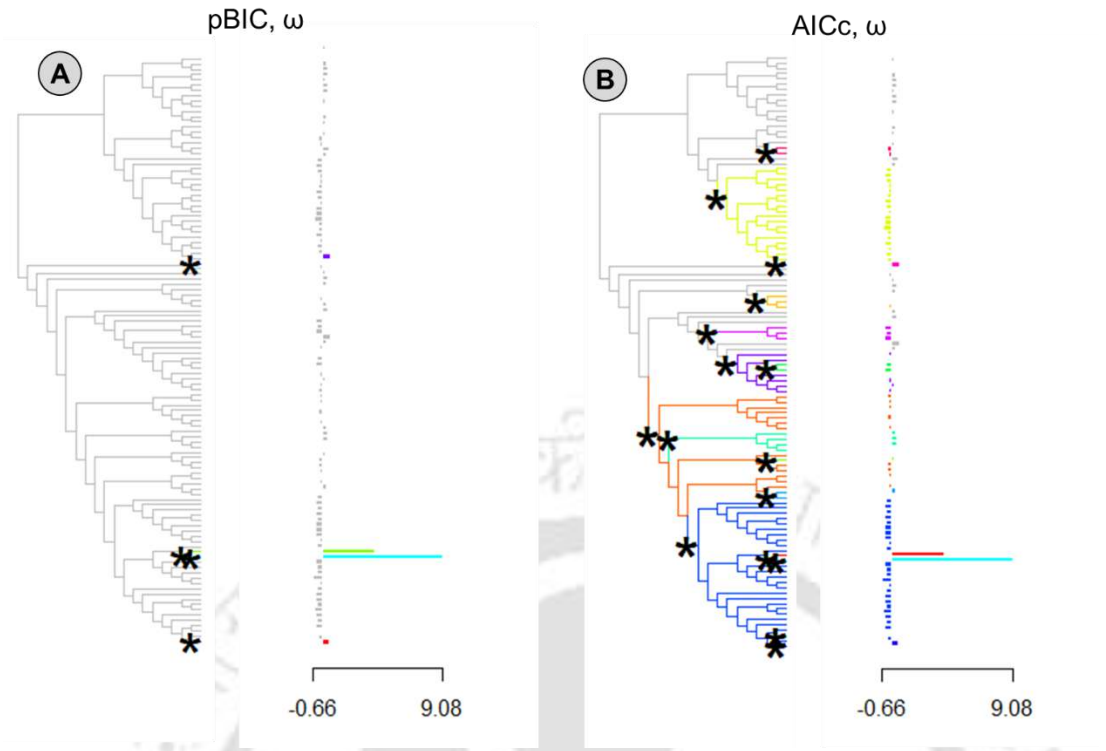


Figure 5.12: Evolutionary shifts (asterisks) in ω ratio across Diptera phylogeny using pBIC (A) and AICc (B) methods in ℓ1ou. Coloured branches represent convergent adaptive peaks and gray or black branches represent non-convergent regimes. The bar graphs showing at the right side of phylogenetic tree the individual trait used for analysis.

The primary use of SURFACE is to build a macroevolutionary adaptive landscape for a clade given only a phylogenetic tree and measurements of one or more continuous traits for each member species. This is accomplished by employing stepwise AIC algorithms to fit a series of 'Hansen' models with two distinct phases applying ouch functions. New selective regimes are added to the model in the forward phase, and in the backward phase, many regimes are 'collapsed' into convergent regimes recognized separately by various lineages. The backward phase in SURFACE identified 11 regimes in total for the dS molecular trait. The best SURFACE model selected (AICc = 52.1) was favored over a model with the same number of regime shifts (k = 20) but without any convergent regime (ΔAICc > 38.6). The final model included 20 regime shifts, 10 distinct regimes (Δk = 10), 16 convergent shifts (c), 6 convergent regimes (k′conv = c- Δk) and 4 non-convergent regimes (k′nonconv = k-c) (Table 5.23). The

10 adaptive optima (θ, one per regime per trait) SURFACE detected for dS (Fig. 5.10 C). A total of 11 regimes were observed for the dN molecular trait, and the best model (AICc = -791.6) was chosen over a model with the same number of regime shifts (k = 24) but without any convergent regime (ΔAICc > 51.7). The final model comprised of 24 regime shifts, 14 distinct regimes (Δk = 10), 19 convergent shifts, 9 convergent regimes and 5 non-convergent regimes (Table 5.23). SURFACE detected 14 adaptive optima (θ) for dN (Fig. 5.11 C). For the molecular trait ω, a total of 13 regimes were observed in the backward phase, and the best model (AICc = -880.2) was chosen over a model with the same number of regime shifts (k =

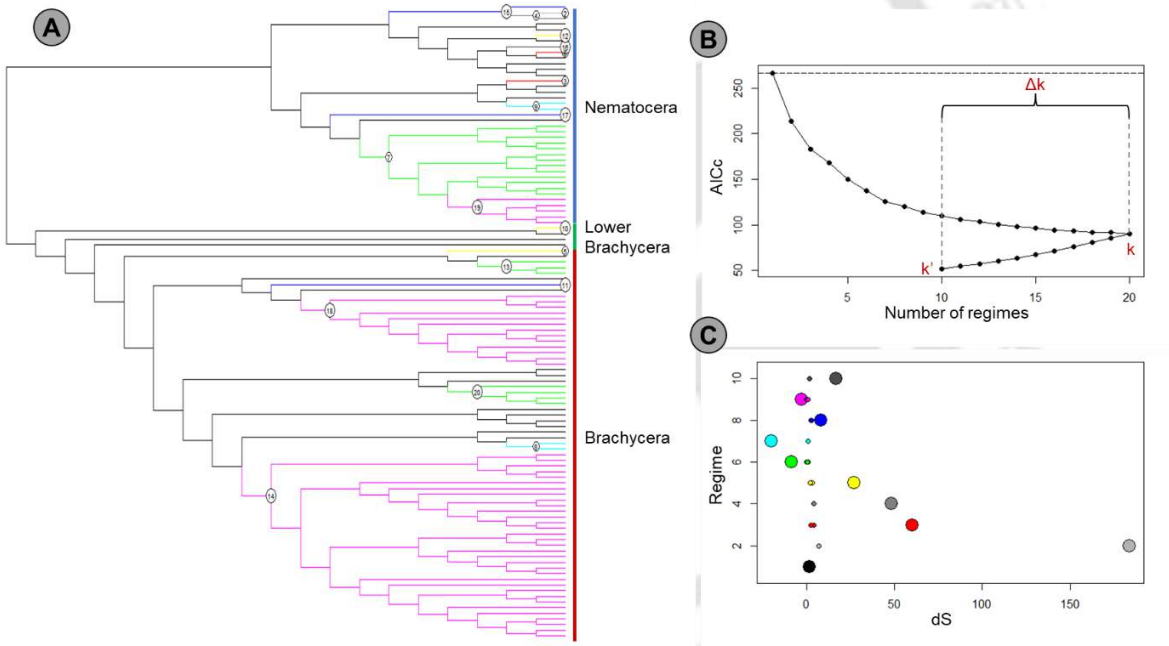


Figure 5.13: Results of a SURFACE analysis of Diptera flies with molecular trait synonymous substitution rate (dS). (A) Phylogenetic tree, with surfaceTreePlot used to paint convergent (coloured) and nonconvergent (greyscale) regimes onto the branches. (B) Change in AICc during the forward and backward phases of the analysis. (C) trait values for each species (small circles) and estimated optima (large circles), with regime colours matching those in the tree.

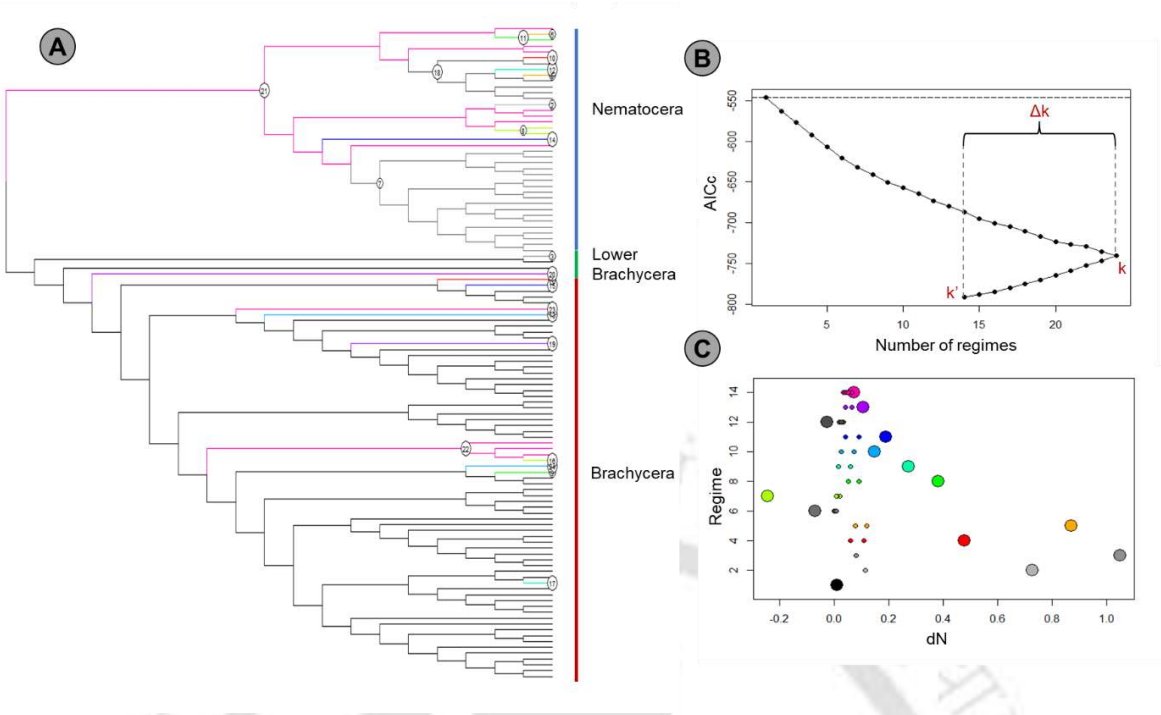


Figure 5.14: Results of a SURFACE analysis of Diptera flies with molecular trait non-synonymous substitution rate (dN). (A) Phylogenetic tree, with surfaceTreePlot used to paint convergent (coloured) and nonconvergent (greyscale) regimes onto the branches. (B) Change in AICc during the forward and backward phases of the analysis. (C) trait values for each species (small circles) and estimated optima (large circles), with regime colours matching those in the tree.

19) but no convergent regime (ΔAICc > 44.3). The final model included 19 regime shifts, 7 distinct regimes (Δk = 12), 16 convergent shifts, 4 convergent regimes and 3 non-convergent regimes (Table 5.23). The 7 adaptive optima (θ) SURFACE detected for molecular trait ω (Fig. 5.12 C). Similar molecular traits across various species are evidence of a stabilizing selection pressure, proposed in the OU model of trait evolution.

5.3.6 Continuous trait-dependent diversification:

ES-sim test: According to continuous trait evolution study, OU is the best-fit model for Diptera molecular traits (dS, dN, and ω), with a relatively strong pull for ω and a small pull for dS and dN (Table 5.17, 5.18, 5.19); hence, the OU null model for ES-sim is the best to test whether the lineage diversification rate is related to molecular trait evolution. This analysis able to detect sufficient trait-dependent correlations in the dS and dN traits using ES-sim with both

OU (p = 0.11 for dS and p = 0.08 for dN) or Brownian motion (p = 0.002 for dS and p = 0.001 for dN). The Spearman's rho was 0.2 for dS and 0.22 for dN with OU, indicating that diversification of Diptera satisfactorily corelated with molecular traits dS and dN (Table 5.25). On the other hand, however, the OU null model unable to detect relation with ω (p = 0.98), with least correlation between speciation rate and trait (Spearman's rho = -0.00177). Under a Brownian null model, diversification also showed insignificant corelation with the trait (Spearman's rho = -0.03637, p = 0.8) (Table 5.25). Thus, OU null model showed diversification is not dependent on ω when testing for a monotonic relationship.

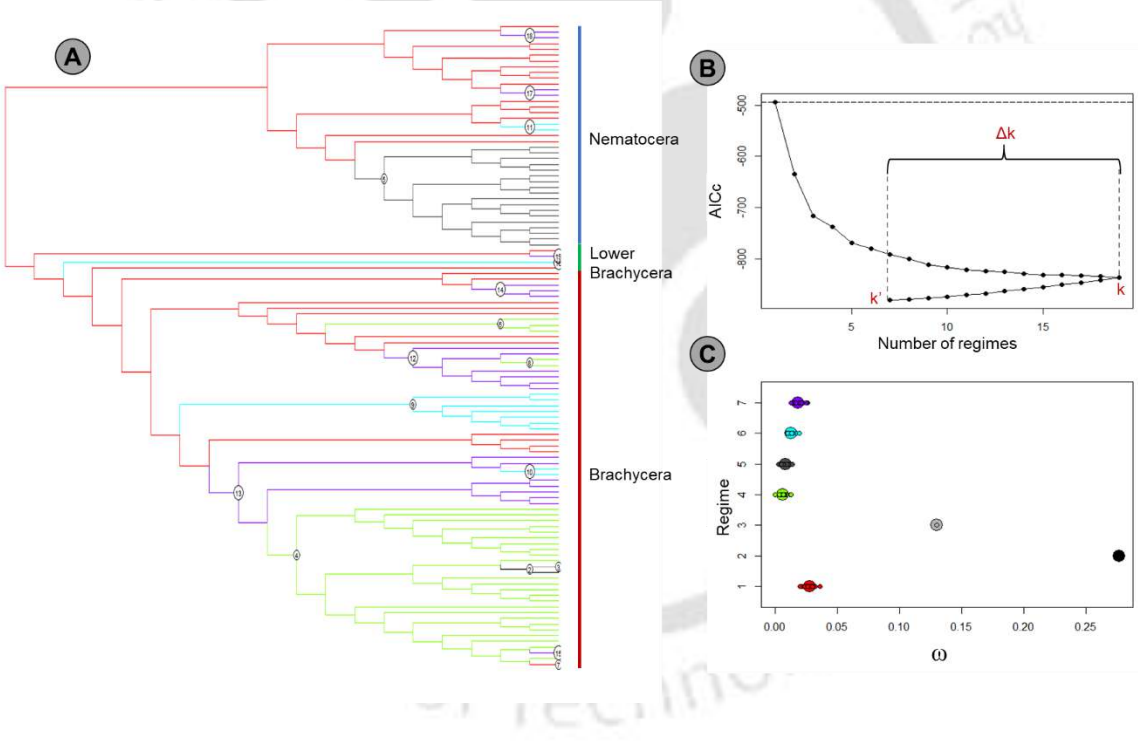


Figure 5.15: Results of a SURFACE analysis of Diptera flies with molecular trait ω. (A) Phylogenetic tree, with surfaceTreePlot used to paint convergent (coloured) and nonconvergent (greyscale) regimes onto the branches. (B) Change in AICc during the forward and backward phases of the analysis. (C) trait values for each species (small circles) and estimated optima (large circles), with regime colours matching those in the tree.

Table 5.25: Trait- dependent diversification tests through ES-sim and TB-pgls examined in this study

Method	parameter	dS	dN	ω
TB-pgls	Slope	-0.08239	-3.59211	1.012482
	P Value	0.026876	0.01112	0.281686
ES-sim Brownian	ρ (rho)	-0.38623	-0.42275	-0.03637
	P Value (spearman)	0.0024	0.0012	0.80532
ES-sim OU	ρ (rho)	0.201499	0.226435	-0.00177
	P Value (spearman)	0.118	0.0832	0.9856

Table 5.26: Quantitative State Speciation and Extinction (QuaSSE) analysis for molecular traits (dS, dN and ω) on the basis of diversification (Linear, Sigmoidal and Modal models). The best model based on AICw shown in different colour.

trait		Model	Df	lnLik	AIC	ΔAIC	AICw	χ^2	p
dS		Constant	3	-445.26	896.52	7.82	0.020041		
	Diffusional	Linear	4	-443.64	895.28	6.58	0.037254	3.2377	0.071964
		Sigmoidal	6	-438.35	888.7	0	1	13.8208	0.003159
		Modal	6	-444.54	901.09	12.39	0.00204	1.4315	0.69817
	Directional	Linear (φ)	5	-443.63	897.26	8.56	0.013843	3.2554	0.196378
		Sigmoidal (φ)	7	-438.23	890.47	1.77	0.412714	14.0504	0.007136
		Modal (φ)	7	-444.54	903.08	14.38	0.000754	1.4372	0.83771
dN		Constant	3	-29.279	64.558	0	1		
	Diffusional	Linear	4	-29.022	66.044	1.486	0.475685	0.5136	0.4736
		Sigmoidal	6	-28.981	69.962	5.404	0.067071	0.596	0.8973
		Modal	6	-29.206	70.411	5.853	0.053584	0.1466	0.9857
	Directional	Linear (φ)	5	-29.006	68.012	3.454	0.177817	0.5463	0.761
		Sigmoidal (φ)	7	-25.969	65.937	1.379	0.501827	6.6205	0.1574
		Modal (φ)	7	-28.978	71.956	7.398	0.024748	0.6016	0.9629
ω		Constant	3	-14.931	35.861	0	1		
	Diffusional	Linear	4	-14.365	36.73	0.869	0.647588	1.1306	0.2877
		Sigmoidal	6	-14.061	40.122	4.261	0.118778	1.7387	0.6284
		Modal	6	-13.583	39.165	3.304	0.191666	2.6956	0.441
	Directional	Linear (φ)	5	-13.413	36.826	0.965	0.617238	3.0349	0.2193
		Sigmoidal (φ)	7	-13.592	41.183	5.322	0.069878	2.6775	0.6132
		Modal (φ)	7	-13.101	40.203	4.342	0.114063	3.6578	0.4543

The models were run without and then with a directional function (indicated here by phi, φ). p value to test significant difference to a model of constant speciation and extinction. Delta AIC (ΔAIC) calculated by comparing model to the best-fit, lowest AIC, model.

QuaSSE analyses: The trait-dependent diversification analyses using QuaSSE selected the sigmoidal model of diversification under stochastic trait evolution as the best support for dS as a trait (AIC = 888.7). The subsequent alternative model is directional sigmoidal (AICw = 0.412; χ^2 test, $p < 0.05$). This implies that the rate of synonymous substitution has a considerable influence on the rate of diversification in the "S" shape curve manner. Whereas, dN and ω traits support constant null model provided a significantly better fit than any of the rate-variable models, as measured by the lowest AIC value (AIC = 64.558 for dN and AIC = 35.861 for ω) (Table 5.26). However, the directional sigmoidal (AICw = 0.5) and diffusional linear (AICw = 0.475) models are the subsequent alternative models for dN (ΔAIC < 2) not supported by χ^2 test ($p > 0.05$). Similarly, the diffusional linear (AICw = 0.647) and directional linear (AICw = 0.617) models are successive alternative models for ω (ΔAIC < 1) not supported by χ^2 test ($p > 0.05$) (Table 5.26). It indicates that neither dN nor ω are strongly related to the diversification rates.

5.3.7 Evolution of molecular traits in response to food habit:

This chapter focuses on the establishment of a link between different feeding habits and nucleotide substitution rate, which is portrayed as a continuous molecular trait of Diptera flies. Our findings reveal that molecular traits differ between species based on feeding habits, but are generally higher in detritivores and lower in haematophagy lineages. The trend of mean values for all three molecular traits is detritivore > herbivore > carnivore > haematophagy. So, here we observed that detritivores have higher substitution rates be it synonymous or non-synonymous than that of other lineages with different food habit. Consequently, the dN/dS ratio is likewise quite high in detritivore species, indicating that a detritivore fly has acquired nonsynonymous nucleotide substitutions more often than others. Thus, relatively stronger positive selection occurred in species those who pursued detritivore trophic habit and on the other hand haematophagy lineages gone through comparatively stronger purifying selection.

Detritivores species actually have smaller populations than other species, and since negative selection is more potent in species with larger populations, it is likely that in species with smaller effective population sizes, relatively detrimental nonsynonymous mutations would accumulate more quickly[45]. Therefore, the lineage with different food habit of flies have gone through different selection pressure for their survival. This study appears to reveal that Diptera flies diversified selection on mitochondrial genes to adapt for surviving on different food resources. The genesis of species is a widely debated issue, with many experts agreeing that species emerge through population adaptation in contrasting habitats[46,47]. For instance, the well-known research on Darwin's finches revealed that climatic change affected food resources, which eventually contributed to the survival and evolution of certain traits[46,48]. In this scenario, we believe it is highly possible that the molecular traits (dS, dN, and ω) have evolved in response to those dietary preferences of the flies.

We used this variability to analyze the tempo of trophic evolution across Diptera flies and the impact of diet on speciation and extinction. Dietary variation is a common characteristic among various species groups, and transitions between feeding strategies are major influence on the uneven patterns of lineage diversity[49]. Our findings show that transition rates between feeding habits vary, for example, herbivore to carnivore transition is prohibited, but carnivore to herbivore transition is tolerated; detritivore to non-detritivore transition rate is higher than the opposite transition rate; and haematophagy to non-haematophagy transition rate is lower than the opposite direction transition. Our results also imply that the diversification rates of species with different food preferences varies; for instance, herbivores have a higher diversification rate than carnivores, and haematophagy has a higher diversification rate than non-haematophagy. Even though our analysis reveals the same rate of diversification for detritivores and non-detritivores, this could be due to an inadequate model. This research also explores "hidden" states in the model that potentially have distinct diversification rates,

such as, a different rate of diversification on the two hidden states and a similar rate of diversification on the observed herbivore and carnivore traits; four separate diversification rates were allowed for the traits detritivore and non-detritivore including two hidden states; for the traits haematophagy and non-haematophagy, three distinct diversification rates were allowed with identical diversification rate between 0A and 1A.

The disparity of molecular traits was not entirely consistent with variance expected under BM evolution for Diptera history. Phylogenetic conservatism across the range of molecular traits has led in an OU process, which drives species to overlap in the trait-space around molecular traits (dS, dN, and ω) towards certain adaptive optima. The OUCH analysis confirms that the macroevolutionary processes underlying these patterns and suggest that higher theta value for herbivore over carnivore, detritivore over non-detritivore and non-haematophagy over haematophagy. This analyses also suggest that the molecular trait dS have higher drift which led to broader distribution final state of optima; and dN, on the other hand, have lower drift which implies a concentrated distribution of final state; ω have stronger selection this led to a faster approach to the optimum value as well as a narrower dispersion of phenotypes around the optimum. Furthermore, univariate ℓ1ou and SURFACE analyses using molecular traits (dS, dN, and ω) also indicate that there are several instances of convergence across lineages with adaptive peaks between clades with similar trait phenomenon. The quantitative traits can influence the processes of lineage diversification over macroevolutionary time spans[40]. Our findings show that dS and dN have positive correlation with diversification rate but ω unable to show any correlation with diversification rate through ES-sim analysis. According to the QuaSSE model, diversification rate is a sigmoidal function of dS, however neither dN nor ω are correlated to diversification rates.

5.4 Conclusion:

This entire study provides a persuasive theoretical framework and functional explanation for flies' feeding habits and their relationship with the rate of nucleotide substitution. Although the determinants of fly diversity are undoubtedly complex, diet has had a significant impact on Dipteran evolution, which allowed to quantify through substitution rates. The model we present can be combined with larger taxon sampling, thorough character coding, and data from the fossil record to deliver further insights and a more appropriate conclusion.

5.5 References:

1. Rand, D. M., Mossman, J. A., Zhu, L., Biancani, L. M. & Ge, J. Y. Mitonuclear epistasis, genotype-by-environment interactions, and personalized genomics of complex traits in Drosophila. *IUBMB Life* **70**, 1275–1288 (2018).
2. da Fonseca, R. R., Johnson, W. E., O'Brien, S. J., Ramos, M. J. & Antunes, A. The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics* **9**, 1–22 (2008).
3. Saraste, M. Oxidative Phosphorylation at the fin de siècle. *Science (80-.).* **283**, 1488–1493 (1999).
4. Liesa, M. & Shirihai, O. S. Mitochondrial dynamics in the regulation of nutrient utilization and energy expenditure. *Cell Metab.* **17**, 491–506 (2013).
5. Aw, W. C. *et al.* Genotype to phenotype: Diet-by-mitochondrial DNA haplotype interactions drive metabolic flexibility and organismal fitness. *PLoS Genet.* **14**, e1007735 (2018).
6. O. Ballard, J. W. & Youngson, N. A. Review: Can diet influence the selective advantage of mitochondrial DNA haplotypes? *Biosci. Rep.* **35**, 1–17 (2015).
7. Kwang, P. L. *et al.* Lifespan and reproduction in Drosophila: New insights from nutritional geometry. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2498–2503 (2008).
8. Galtier, N., Jobson, R. W., Nabholz, B., Glé min, S. & Blier, P. U. Molecular evolution Mitochondrial whims: metabolic rate, longevity and the rate of molecular evolution. *Biol. Lett.* **5**, 413–416 (2009).
9. Allio, R., Donega, S., Galtier, N. & Nabholz, B. Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Mol. Biol. Evol.* **34**, 2762–2772 (2017).
10. Arnqvist, G. *et al.* Genetic architecture of metabolic rate: environment specific epistasis between mitochondrial and nuclear genes in an insect. *Evolution (N. Y).* **64**, 3354–3363 (2010).
11. Salminen, T. S. & Vale, P. F. Drosophila as a Model System to Investigate the Effects of Mitochondrial Variation on Innate Immunity. *Front. Immunol.* **11**, 521 (2020).
12. Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V. & Wallace, D. C. Effects of Purifying and Adaptive Selection on Regional Variation in Human mtDNA. *Science (80-.).* **303**, 223–226 (2004).
13. Shen, Y. Y., Shi, P., Sun, Y. B. & Zhang, Y. P. Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. *Genome Res.* **19**, 1760–1765 (2009).
14. Sun, Y.-B., Shen, Y.-Y., Irwin, D. M. & Zhang, Y.-P. Evaluating the Roles of Energetic Functional Constraints on Teleost Mitochondrial-Encoded Protein Evolution. *Mol. Biol. Evol.* **28**, 39–44 (2011).
15. Yang, Y., Xu, S., Xu, J., Guo, Y. & Yang, G. Adaptive Evolution of Mitochondrial Energy Metabolism Genes Associated with Increased Energy Demand in Flying Insects. *PLoS One* **9**, e99120 (2014).

16. Zhou, Z. *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci.* **113**, E6117–E6125 (2016).

17. Song, H., Gao, H., Liu, J., Tian, P. & Nan, Z. Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in Arachis duranensis and Arachis ipaënsis orthologs. *Sci. Rep.* **7**, 14853 (2017).

18. Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E. & Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30**, 308–321 (2014).

19. Chattopadhyay, E. *et al.* Genome-wide mitochondrial DNA sequence variations and lower expression of OXPHOS genes predict mitochondrial dysfunction in oral cancer tissue. *Tumor Biol.* **37**, (2016).

20. Powder, K. E., Cousin, H., McLinden, G. P. & Craig Albertson, R. A Nonsynonymous Mutation in the Transcriptional Regulator lbh Is Associated with Cichlid Craniofacial Adaptation and Neural Crest Cell Development. *Mol. Biol. Evol.* **31**, 3113–3124 (2014).

21. Hassan, S. S., Choudhury, P. P. & Roy, B. SARS-CoV2 envelope protein: non-synonymous mutations and its consequences. *Genomics* **112**, 3890–3892 (2020).

22. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

23. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).

24. Miyata, T. & Yasunaga, T. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol. 1980 161* **16**, 23–36 (1980).

25. Bertone, Matthew A., and B. M. W. *True fies (Diptera). The timetree of life* (2009).

26. Skevington, J. H. & Dang, P. T. Exploring the diversity of flies (Diptera). *Biodiversity* **3**, 3–27 (2002).

27. Wiegmann, B. M. *et al.* Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* **108**, 5690–5695 (2011).

28. Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**, 129–131 (2008).

29. Revell, L. J. phytools : an R package for phylogenetic comparative biology (and other things). 217–223 (2012) doi:10.1111/j.2041-210X.2011.00169.x.

30. Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a Binary Character's Effect on Speciation and Extinction. *Syst. Biol.* **56**, 701–710 (2007).

31. FitzJohn, R. G. <tt>Diversitree</tt> : comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* **3**, 1084–1092 (2012).

32. Beaulieu, J. M. & O'Meara, B. C. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. *Syst. Biol.* **65**, 583–601 (2016).

33. Pennell, M. W. *et al.* geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216–2218 (2014).

34. Harmon, Luke J., James A. Schulte, Allan Larson, and J. B. L. Tempo and Mode of Evotutionary Radiation in Iguanian Lizards. *Science (80-.).* **301**, 961–964 (2003).

35. Slater, G. J., Price, S. A., Santini, F. & Alfaro, M. E. Diversity versus disparity and the radiation of modern cetaceans. *Proc. R. Soc. B Biol. Sci.* **277**, 3097–3104 (2010).

36. Butler, M. A. & King, A. A. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am. Nat.* **164**, 683–695 (2004).

37. Ingram, T. & Mahler, D. L. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods Ecol. Evol.* **4**, 416–425 (2013).

38. Khabbazian, M., Kriebel, R., Rohe, K. & Ané, C. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. *Methods Ecol. Evol.* **7**, 811–824 (2016).

39. Gomes, A. C. R., Sorenson, M. D. & Cardoso, G. C. Speciation is associated with changing ornamentation rather than stronger sexual selection. *Evolution (N. Y).* **70**, 2823–2838 (2016).

40. Harvey, M. G. & Rabosky, D. L. Continuous traits and speciation rates: Alternatives to state-dependent diversification models. *Methods Ecol. Evol.* **9**, 984–993 (2018).

41. FitzJohn, R. G. Quantitative Traits and Diversification. *Syst. Biol.* **59**, 619–633 (2010).

42. Rabosky, D. L. & Goldberg, E. E. Model Inadequacy and Mistaken Inferences of Trait-Dependent Speciation. *Syst. Biol.* **64**, 340–355 (2015).

43. Murrell, D. J. A global envelope test to detect non-random bursts of trait evolution. *Methods Ecol. Evol.* **9**, 1739–1748 (2018).

44. García-Navas, V., Noguerales, V., Cordero, P. J. & Ortego, J. Phenotypic disparity in Iberian short-horned grasshoppers (Acrididae): the role of ecology and phylogeny. *BMC Evol. Biol.* **17**, 1–14 (2017).

45. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713 (1962).

46. Lamichhaney, S. *et al.* A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science (80-.).* **352**, 470–474 (2016).

47. Arnegard, M. E. *et al.* Genetics of ecological divergence during speciation. *Nature* **511**, 307–311 (2014).

48. Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).

49. Price, S. A., Hopkins, S. S. B., Smith, K. K. & Roth, V. L. Tempo of trophic evolution and its impact on mammalian diversification. *Proc. Natl. Acad. Sci.* **109**, 7008–7012 (2012).

.

CHAPTER 6

Dynamics of Speciation-Extinction

"Extinction is the rule. Survival is the exception."

— Carl Sagan

TH-2821_146106003

CHAPTER 6

Deciphering major diversification time through molecular dating and dynamics of speciation-extinction events of Diptera

Abstract:

Diptera is the most diverse insect order, with about <150,000 extant species classified into three major subgroups based on morphology: Nematocera, lower Brachycera, and Schizophora. Estimating phylogenetic divergence periods is a challenging task that yields conflicting results in numerous clades across the tree of life. This discrepancy has been linked to a number of factors, including the quantity of the molecular data, nucleotide substitution rate heterogeneity, the reference phylogenetic tree, the calibration technique, the branching prior, the molecular clock model, and the molecular dating method. To understand the ecological and evolutionary processes that determine biodiversity, time-dependent rates of speciation and extinction must be estimated using dated phylogenetic trees of extant species (timetrees) and the reasons for variance must be assessed. In this work, we first investigated time estimation variations by adopting distinct nucleotide substitution rates (similar, different), different branching prior processes (yule and birth-death), various software, and different fossil calibration strategies. We inspected the diversification patterns and rate shift on the time calibrated trees, also allows for estimating potential speciation and extinction events. We explored whether biotic or abiotic factors better explain Diptera diversification dynamics by determining whether speciation and extinction were mediated by diversity-dependence (niche filling) and specialized traits or by

large-scale continuous changes in extrinsic factors such as climate or geology. We also examined potential speciation and extinction events using only fossil data. Our findings reveal that different dating approaches have a substantial influence on estimating the chronology of major Diptera lineages as well as subsequent downstream macroevolutionary analyses. We found considerable variation in diversification rates linked with distinct Diptera traits, as well as diversification that is dependent on diversity. According to palaeo-environment diversification models, high extinction rates occurred during periods of global warming and low speciation rates occurred during periods of elevated sea level. Also, this study found that Diptera witnessed mass-extinction. This analysis supports the relevance of a combination of factors rather than a single explanation in predicting lineage diversification.

6.1 Introduction:

Insects have diversified over the last 400 million years (Myr), and they were among the first creatures to inhabit land after the development of plants, acquiring a variety of features including as flight, complete metamorphosis, and advanced sociality[1,2]. Diptera is one of four mega-diverse insect orders, accounting for over 150,000 reported species on Earth[3]. This vast fly diversity is conventionally divided into two main suborders: the lower Diptera (Nematocera), mosquito-like flies with long antennae, and Brachycera, stout and fast-moving flies with short antennae[3]. Different research suggests Diptera flies came into existence either before catastrophic End-Permian Extinction (EPE) shown by Montagna, Matteo, et al.[4] or after EPE described by Misof et al.[5]. The study of Wiegmann et al. not only supports a post-EPE origin for Diptera but also suggests episodic radiations for important clades, such as Brachycera's emergence after the end of the Triassic event (at 201 Ma) and Schizophora's origin after the end of the Cretaceous event (at 66 Ma)[3]. But a dedicated investigation on Diptera by Zhao Z. et al. has shown pre-EPE origin of Diptera[6]. Simpsonian theory on adaptive radiation can explain this type of quick burst following any catastrophic events, in which a rapidly

reproducing lineage expands into vast ecological diversity as a result of increased resource availability and limited competition[7,8]. The emergence of vacant niches in the wake of mass extinction events is a major reason for ecological opportunity because previously inaccessible resources become available as a result of key evolutionary innovations, colonisation of new regions, or the removal of contenders due to external habitat depauperation mechanisms[7,9].

Evolutionary biologists have long sought to understand the essential variables that govern the evolutionary dynamics of these explosive radiations, which are thought to have occurred early in a lineage's history[10–12]. The majority of hypotheses attribute early rapid radiations to either biotic or abiotic factors[13]. The biotic factors described by the Red Queen (RQ) model of Van Valen (1973), which derives from Darwin and Wallace, include interactions among species, species ecology, and life-history features[10,14]. Whereas, the Court Jester (CJ) model, based on paleontological evidence, claims that historical abiotic factors such as sudden changes in climate, oceanography, or geological tectonic events drive speciation and extinction rates, typically acting clade-wide across lineages over longer time scales [10,15,16].

The recent advancement of probabilistic models in the field of diversification dynamics has enabled statistical assessment of the relative contributions of abiotic and biotic processes influencing diversity patterns[17–19]. One form of model estimates clade-dependent diversification rates and finds variations in diversification rates between clades that can be explained by key innovations[20–22], or by diversity-dependence and niche filling (i.e., diversification decreasing as the number of species increases)[23,24]. The second type of model involves identifying statistical relationships between diversification rates and changes in species traits (trait-dependent diversification models)[25,26]. The third form model posits continuous variation in diversification rates through time that is based on a paleoenvironmental parameter and analyzes whether abiotic environmental changes can impact diversification rates (e.g., paleotemperature)[27]. Finally, episodic birth–death models investigate tree-wide rate shifts

that occur concurrently throughout all lineages in a tree, such as a mass extinction event that wipes out a group of lineages at a specific time in history[19,28,29].

Explosive diversification episodes with irregular intervals of change have resulted in unequal patterns of species richness across the tree of life[30–32]. Recent developments in the modelling of the evolutionary processes now enable such inferences to be drawn from phylogenetic trees, despite the fact that fast radiations have been inferred in multiple clades[33]. However, the majority of species that have ever inhabited have left little trace in the fossil record, hence extant timetrees are the primary source of knowledge on their historical diversification dynamics[34]. As a result, measuring rates of speciation and extinction across time is a difficult task in evolutionary biology, with some researchers arguing that the lack of extinct taxa in these analyses may have obscured signatures of early rapid diversification[34,35]. Although fossil-based methods for inferring diversification dynamics have recently become prominent in the evolutionary biology[36,37]. Moreover, there is still a considerable disagreement between paleontological and molecular age estimates, or ages obtained from different molecular datasets, in flowering plants[38,39], mammals[40–42], insects[4,43], and other lineages spanning the Tree of Life[44,45]. The conflicting molecular age estimations are explained by a number of reasons, including paleontological calibration models[46–48], nucleotide substitution rate heterogeneity[49,50], and the suitability of the molecular clock model used[51,52].

The three regions of rapid radiation of Diptera correlate to the earliest fly lineages Nematocera (including mosquitos), lower Brachycera (including horse flies), and higher flies, schizophoran Cyclorrhapha (including Drosophila and house flies)[3]. The early aquatic lineages of extant lower Diptera from the Permian period spread rapidly in the Triassic[53]. Later, in the early Jurassic, mostly terrestrial lineages of lower Brachycera radiated, shortly after the recently proposed genesis of angiosperms; many of these lineages are flower visitors with long proboscides for nectar feeding[54,55]. The cyclorrhaphan clade Schizophora was responsible for

the largest insect radiation (together with macro-lepidopteran moths) after the K-Pg boundary in the early Tertiary (65–40 Ma), after a long gap between the earliest known Cyclorrhapha (Opetiala, 127 Ma), which coincided with the development of the ptilinal sac, an improved escape mechanism for the fly from its puparium[2,3,56]. Apart from its vigorous food habit and prolong history of explosive diversification, the sheer number of species in the Diptera order, including Nematocera (over 50,000 species in 40 families), Brachycera (over 100,000 species in 117 families), and Schizophora (over 60,000 species in 87 families), makes it an excellent choice for studying adaptive radiation patterns[3,53,57]. Several earlier studies have found links between the emergence of ecological opportunity and rapid bursts of diversity during three distinct time periods in Diptera evolution[3,53]. The occurrence of ecological opportunity and rapid diversification of Diptera led us to hypothesize that adaptive radiation was the fundamental process that encouraged the ecological— and hence phenotypic diversity— observed within subsequent clades.

Herein, we estimate a robust time-calibrated phylogeny of 112 extant species through Yule and Birth-Death branching prior process employing partitioned and non-partitioned molecular data and using two different fossil calibration (3 fossils and 9 fossils) strategy. These estimates are further compared to explore the role played by abiotic (CJ) and biotic (RQ) factors, in the origin and fate of biological radiations, including paleotemperature, past atmospheric oxygen level and sea level (abiotic), relative diversity of angiosperms, distinct morphological trait (biotic). We also used fossil-based methods in the program divDyn and PyRate to assess lineage diversification. If species interactions are the primary drivers of evolution, diversification rates should exhibit diversity-dependent dynamics or be dependent on ecological traits. Diversification rates, for example, might be expected to decrease as a function of diversity or to increase as a function of ecological opportunity. If, on the other hand, evolution is purely caused by changes in the physical environment, clade-wide responses to abrupt abiotic

instabilities should dominate macroevolutionary dynamics. For example, diversification rates may shift following major climatic changes that wiped out certain lineages while preferring the radiation of others. These hypotheses are tested using trait-dependent, time-dependent, environment-dependent, and episodic birth–death models, and the fit (explanatory power) of these models is compared using maximum likelihood (ML) and Bayesian inference (BI).

6.2 Materials and Method:

6.2.1 DNA sequence data collection:

The Gblock mitochondrial sequence alignment were acquired from the Chapter 3. The 3rd codon position was removed for escaping different associated issues such as substitution saturation, among site composition heterogeneity, sequence divergence heterogeneity. Size of alignment became 6012 bp of 116 taxa.

6.2.2 Fossil Calibration and Divergence Time estimation:

Calibrations are of main importance in divergence dating analysis because it's not possible to estimate absolute ages from molecular data alone. Observed genetic divergence is the product of 2 components (the substitution rate and time elapsed) that cannot be separated without additional, independent information. Such data can come in 2 main forms. The first are calibrations that convey temporal information about nodes in the evolutionary tree. Second form of calibrating information is a known substitution rate that has been estimated independently [58]. In this experiment we used two steps of fossil calibration for molecular dating analysis using previously reported 3 fossils and further 6 fossil records from fossilswork.org. The use of mean distribution for priors, without hard upper and lower constraints, reflects the uncertainty in the fossil record and allows posterior estimates to vary in either direction based on their interactions with the other calibration points during analysis. The molecular dating analysis was pursued through MCMCTree and Beast 2.6 in the local institute server PARAM-ISHAN. The finally sorted mitochondrial dataset 13PCG12 was used for the molecular dating

analysis since eliminating third codon positions can reduce the impact of systemic bias [59,60]. In MCMCTree, we used ML approximation by first calculating the ML estimates of the branch lengths, the gradient vector and Hessian matrix, using BaseML and CodeML programs in PAML [61]. Despite the fact that auto correlated models of clock relaxation have been shown to provide a significantly better fit than uncorrelated models on phylogenetic datasets, all analysis were conducted under both models of molecular clock relaxation. In BEAST only uncorrelated lognormal relaxed clock model has been implemented. As parametric prior distribution lognormal distribution has been chosen for summarizing paleontological information because it can assign the highest point probability for the nodal age to be slightly older than the oldest fossil [58].

First three fossil calibrations compatible with our phylogenetic tree were selected from previous literature. (1) Diptera (230 Mya): The most common recent ancestor (MRCA) of all Diptera was calibrated based on the fossil occurrence of *Grauvogelia arzvilleriana* at 230 Mya, 95% prior age interval extended up to 311 Mya in mid of Pennsylvanian epoch of Moscovian age[62,63]. (2) Brachycera (187 Mya): crown group Brachycera was calibrated based on fossil record Paleobolbomyia at 187 Mya with 95% prior interval set to 226 Mya assuming its origin in Triassic at the beginning of Noiran age[3] (3) Schizophora (64 Mya): the crown group Schizophora calibrated as the fossil record of Phytomyzites in 64 Mya. Carolina et al. placed origin of Schizophora clade within the upper Cretaceous (beginning of Cenomonian) to the Paleogene periods (known as K-Pg boundary) yet we reduced the maximum bound up to 90 Mya[64].

These fossils calibration constraints were used with soft bounds under birth-death prior in MCMCTree, because this strategy has been presented to provide the best compromise for dating estimates. The prior on the root age set to the first appearance of the winged insect at 325 Mya in Carboniferous.

Table 6.1: Fossil calibration strategy

Fossil	Min age constrain (Mya)	Max age constrain (Mya)	Clade	Reason of assumption Max age constrain
Grauvogelia	230	311	Diptera	Mid of Pennsylvanian epoch of Moscovian age
Paleobolbomyia	187	226	Brachycera	226 Mya assuming its origin in Triassic at the beginning of Noiran age (Weigman)
Phytomyzites	64	90	Schizophora	Junqueira et al. placed origin of Schizophora clade within the upper Cretaceous (beginning of Cenomonian) to the Paleogene periods (known as K-Pg boundary), we reduced the maximum bound up to 90 Mya.
Fossilworks	50.3	57.5	Tachinidae	Beginning of Eocene thermal maximum 1 (ETM1) 58-55 Mya Temperature rise (5-8)
Fossilworks	50.3	62.1	Oestridae	Danian age right after K-Pg boundary.
Fossilworks	145.5	210	Tabanidae	After separation of continent
Fossilworks	205	245	Chironomidae	After Permian Mass extinction
Fossilworks	99.3	194	Culicidae	After Triassic–Jurassic extinction event
Fossilworks	195	227	Psychodidae	Beginning of Norian age

6.2.3 Molecular Dating Analysis:

Dates of divergence between Diptera were estimated using Bayesian relaxed molecular clock approaches implemented in MCMCTree from PAML package and BEAST 2.6 [65,66]. In MCMCTree we used consensus topology derived from RAXML analysis. We used ML approximation by first calculating the ML estimates of branch lengths, the gradient vector and Hessian matrix, using BaseML and CodeML programs of PAML. BEAST utilizes only uncorrelated relaxed clock models and we used the uncorrelated lognormal model both in Yule and Birth death tree branching prior.

Five fossil calibrations compatible with the tree were selected. These calibrations constrain were used in two stages, at first only three fossil used for calibration and in second stage nine

calibration constraints were used with soft bounds under a birth-death prior in MCMCTree, as this strategy proved to best compromise for dating analysis. The prior on the root age was set at 325 Ma as fossil of first winged insect appeared in late Carboniferous period [5]. In BEAST we used lognormal distribution for confidence intervals constraints as calibration prior with birth-death process on the tree.

In MCMCTree, dating estimations were conducted by running 2 independent MCMC chains were run with following parameters: number of samples = 20000; sampling frequency = 1000; burn in = 20000. 19980000 samples were then summarized to estimate mean divergence date and 95% credibility intervals. In BEAST, the setup using two parallel chain Metroposlis coupled MCMC run for 300 million for analyzing the dataset were established with following parameters: resample every 1000 generations; Delta temperature = 0.1; Optimise delay = 100; Target = 0.234 (default); Number of Initialization Attempts = 10; trace and screen log every 1000 generations. Tracer was used for checking the stationary of MCMC chain and TreeAnnotator was used for deriving maximum clade credibility (MCC) tree from the BEAST analyses. Effective sample size (ESS) value each of the run was measured in Tracer.

6.2.4 Lineage diversification analyses:

Tempo of diversification – the relationship between speciation and extinction rate – was qualitatively analyzed using a lineage-through-time (LTT). The LTT plot created by ape package in R using MCC tree over 1000 randomly sampled trees from BEAST analysis[67]. we calculated Pybus and Harvey's gamma (γ) statistic to test for explosive early diversification[68]. This statistic measures the density of ordered inter-node distances on a phylogeny, to determine if they are evenly distributed (γ = 0; pure-birth), clustered early (negative γ; early burst), or clustered late (positive γ; late burst or high early extinction). As the tree represents an incompletely sampled phylogeny (n = 112), we applied the Monte Carlo randomization in the phytools package (Monte Carlo constant-rates test, function mccr) to generate 1000 null

phylogenies produced under a Yule process using different rho (ρ =1, 0.5 and 0.1)[69]. This will allow us to see if the value of γ from the incomplete tree differs considerably from the null expectation when compared to the randomised values generated by various sampling trees.

6.2.5 Diversification Analyses from molecular dated tree:

To get a more nuanced view of the Diptera diversification dynamic, we used the reversible-jump MCMC (rjMCMC) method implemented in the R package BAMM, which models both rate variation between lineages and rate change with time [70,71]. BAMM differs from previously reported techniques, as it models the position of diversification regimes on the tree as a variable[71]. It incorporates incomplete taxon sampling directly into likelihood calculations; our time trees contained 0.074% of Diptera species. We performed BAMM runs on the multiple MCC phylogeny previously generated, with 10 million generations of Markov Chain Monte Carlo (MCMC) sampling per run and sampling evolutionary parameters every 1000 generations. We assessed convergence of BAMM runs by computing effective sample sizes of log-likelihoods, numbers of processes, and evolutionary rate parameters using the CODA library for R; all parameters had effective sample sizes sizes > 200[72]. BAMMtools were used to compare the frequently observed rate regimes with Bayes factors; identify the branches where rate shifts occurred. We calculated the mean speciation, extinction and diversification rates and plotted the rate-through-time curves for major clades of Diptera from the joint posterior density of parameters simulated with BAMM.

We employed TESS' Bayes factor model selection to explicitly test the relative fit of the following series of alternative branching models to our comparative dataset: (i) time-homogeneous birth–death, (ii) episodically-varying rate and (iii) explicit mass-extinction (survival probability = 10%) that incorporates the diversification parameters. Model comparison analyses were applied for both the maximum clade credibility (MCC) tree and 100 trees sampled from the posterior distribution employing diversified sampling strategy. For a

more quantitative evaluation of Diptera diversification over time, we searched for potential evidence of past mass extinctions using the 'compound Poisson process on mass-extinction analysis' (CoMET) in TESS package of R[72,73]. We performed reversible jump Markov chain Monte Carlo (rjMCMC) analyses under the CoMET (Compound Poisson process on Mass-Extinction Times) model with number of expected mass extinctions and number expected rate changes were 3 and Sampling Fraction = 0.0007466. The expected survival probability was set at 0.05 (5 %) and allowed run to a minimum effective sample size of up to 500. Further we have done series of CoMET analysis by tweaking different parameters such as, sampling fraction, without mass-extinction events, number of expected mass extinctions and number of expected rate changes.

6.2.6 Episodic Birth–Death (Tree-Wide) Diversification:

To inspect the impact of specific abiotic events on diversification, such as mass extinction, sudden changes in plate tectonics, or climate change, we used an episodic birth–death model implemented in the R-package TreePar 3.3[28,74]. This model implies that diversification rates (speciation and extinction) are constant across the tree but might vary within discrete time intervals (i.e., episodically). This allows for the detection of discrete changes in speciation and extinction rates that effect simultaneously all lineages in a tree (i.e., tree-wide rate shifts) caused by global mass extinction or environmental events that impact all lineages at once[74]. We used the function "bd.shifts.optim" in TreePar to compare the likelihood of five episodic birth–death models from zero (constant-rate model) to four diversification rate shifts during the evolutionary history of the Diptera. We performed 3 independent TreePar analyses for each time tree (8 time trees) allowing and disallowing Yule (pure birth) and ME (mass extinction) events (argument ME/Yule = TRUE/FALSE)[15]. For these analyses, we set the "grid" to 5 million years for estimation of rate shifts, the "end" as the age of the Diptera and "start" as the

present (= 0 Mya). In addition to AIC scores, the significance level is employed to determine the best-fit across the multiple models, allowing for gradually more shifts during the evolution.

6.2.7 Time-Dependent Diversification:

We assessed variation in species diversification over time by performing a time-dependent diversification analyses using RPANDA[75]. We evaluated ten models (the first two of which were null models) 1) speciation rate is constant over time with no extinction (Yule null model); 2) speciation and extinction rates are constant; and others models had exponential and linear dependencies of speciation and extinction 3,7) speciation rate varies over time with a single extinction rate that is constant over time; 4,8) speciation rate varies over time with no extinction; 5,9) speciation rate is constant but extinction rate varies over time; and 6,10) both speciation and extinction rates vary over time.

6.2.8 Environment-Dependent Diversification:

We investigated the impact of environmental change on diversification using paleoenvironment-dependent models in RPANDA 1.9[75]. We employed proxies for three environmental variables that properly represent paleoclimatic change, such as paleotemperatures, sea level, and historical atmospheric oxygen, which may have had consequences for Diptera diversification, as inferred from other species[13,74]. By adapting the technique of Condamine et al.[76], we established eight hierarchical models for each of the environmental data, and each of these models had exponential and linear dependencies of speciation and extinction 1) speciation varies with the paleoenvironmental variable and no extinction; 2) speciation varies with the paleoenvironmental variable and constant extinction; 3) constant speciation, and extinction varies with the paleoenvironmental variable; and 4) both speciation and extinction vary with the paleoenvironmental variable. We then computed the AIC of each fit and plotted speciation and extinction rate over time according to the model with the lowest AIC.

6.2.9 Diversity-Dependent Diversification:

To inspect whether biotic interactions within Diptera influenced their diversification we implemented diversity-dependent models found by Etienne et al. 2012 implemented in the R-package DDD 2.7 [23]. We applied 6 different models: 1) speciation depends linearly on diversity without extinction (DDL); 2) speciation depends linearly on diversity with extinction (DDL+E); 3) speciation depends exponentially on diversity with extinction (DDX+E); 4) speciation does not depend on diversity and extinction depends linearly on diversity (DD+EL); 5) speciation does not depend on diversity and extinction depends exponentially on diversity (DD+EX); 6) speciation and extinction rates linearly depend on diversity (DDL+EL). The starting carrying capacity was set to reflect the existing species diversity, and the carrying capacity (K) is evaluated using the models and parameters. Further to test whether clade-wide shift in diversification parameters occurred or not we implemented diversity-dependent diversification model with a shift in the parameters[77] . We applied 4 different models: 1) diversity dependent Key shift in K (carrying capacity) (SR1), 2) diversity dependent Key shift in K and μ (extinction rate) (SR2), 3) diversity dependent Key shift in K and λ (speciation rate) (SR3), 4) diversity dependent Key shift in K, λ and μ (SR4).

We employed the fitdAICrc function in the R package laser as an alternate test of slowing diversification and departure from constant rate, as well as to detect rate shifts in the presence of extinction [78]. PureBirth and birth-death models were rate-constant diversification models (RC) that were fitted to the data. The rate-variable (RV) models tested were the density-dependent models with exponential and logistic variants (DDX and DDL) and the yule2rate models, which allow for two different rates of speciation across the phylogeny. The models were fitted to the branching times of the maximum clade credibility (MCC) trees generated by BEAST, with shifts permitted exclusively at those times[79].

6.2.10 Trait-Dependent Diversification:

We evaluated the effect of trait changes as a potential driver of macroevolution using state-dependent speciation and extinction (SSE) family of models, in which extinction and speciation rates are associated with phenotypic evolution of a trait along a phylogeny[25]. First we used Nematocera and Brachycera as binary trait and implemented in BiSSE; second we further divided Brachycera as Lower Brachycera and Schizophora and three character states were implemented in MuSSE[3,80,81]. To investigate if trait evolution influenced speciation, extinction, or transition rates, we ran seven distinct models in BiSSE and six models in MuSSE. We calculated the posterior density distribution using Bayesian MCMC analyses (10,000 steps) with the best-fit models, as well as the rates of speciation, extinction, and transition for all eight MCC trees. Further we compared net-diversification rate estimated from best-fit models of all MCC trees.

6.2.11 Diversity dynamics using fossil sampling data:

Fossil's occurrence data of Diptera flies were downloaded from Paleobiology Database (PaleoDB, http://www.paleobiodb.org/). DivDyn package of R was used to estimate origination and extinction rate from fossil occurrence data[37]. We also used PyRate on fossils occurrence data assuming gamma distributed variation rate. PyRate was implemented through reversible jump Markov chain Monte Carlo (rjMCMC) on 10 replicates of fossils dataset[36].

6.3 Result and Discussion:

6.3.1 Divarication time estimation using mitochondrial data:

Here in this analysis, we used four different software (MCMCTree, BEAST, treePL and RelTimeOLS) for time estimation of major diversification events. Two different calibration strategies (3 fossils and 9 fossils) were used along with the assumption of different substitution rate for 2 codon positions (1-1L: different nucleotide substitution for 2 codon positions, 2L: same nucleotide substitution for 2 codon position) and 3rd codon position was removed for ease

of analysis. The result of diversification events mainly contradictory for stem lineages whereas the crown lineages of Diptera (Brachycera, 1-9) show almost similar diversification time. Based on our investigation and the outcomes of the various strategies, we inferred that the origination of Diptera happened over a long stretch of time, from the Carboniferous to the Triassic Period [3,85].

The diversification age estimated by different methods and tactics for the major families of stem lineages in the Nematocera suborder is highly variable, but the major families in the crown group Brachycera are rather preserved (Fig. 6.7). We have only adopted Bayesian relaxed clock estimation of BEAST using Yule and Birth-Death branching priors for further downstream investigation; hence we have only emphasized on the estimated output of BEAST analysis for elaborative comparison on divergence time estimations.

According to our BEAST analysis, regardless of branching prior process the partitioned dataset (1-1L) with 9 fossils calibration suggested late Carboniferous origination of Diptera around 346.9 (Yule or Y) Ma and 360 Ma (Birth-Death or BD) (Fig. 6.2 (Left), Fig. 6.4 (Left)) whereas, 3 fossils calibration proposed Early Carboniferous or Late Permian origin 296.6 Ma (Y1-1L3F) and 301.7 Ma (BD1-1L3F) (Fig. 6.1 (Left), Fig. 6.3 (Left)). In case of non-partitioned dataset or same substitution rate for codon position (2L), 9 fossils calibration suggested Late Permian around 280 Ma (BD2L9F) (Fig. 6.4 (Right)) and 3 fossils calibration suggested early Permian origin (~253- 258 Ma) of Diptera flies (Fig. 6.3 (Right), Fig. 6.1 (Right)), which are in agreement with the results of certain previous studies[3,4,6,85]. The variations in node ages were also matched by similar variation in the in the lower and upper bounds of the 95% CIs of each analysis. However, the 95% CIs for node ages across the tree were generally very wide, such that there was typically some overlap between the date estimates from each analysis.

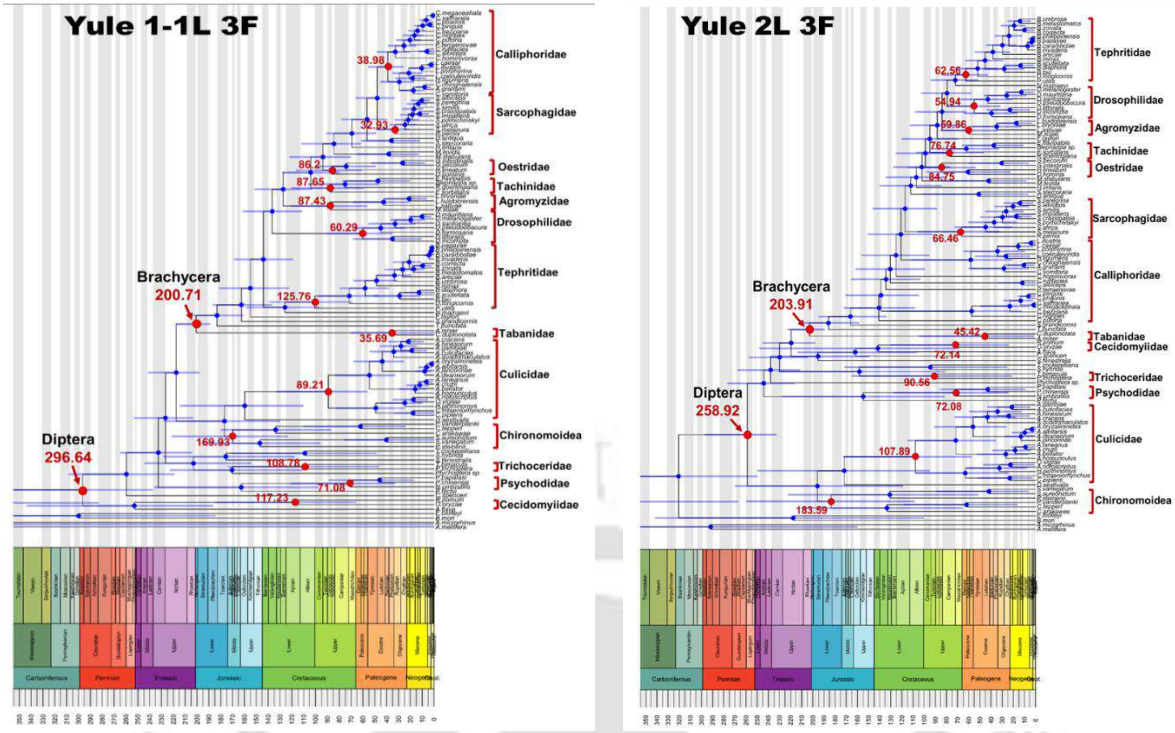


Figure 6.1: Time calibrated phylogeny of Diptera using three fossil calibration using Yule model in BEAST. Different substitution rate for codon position (Left, Y1-1L3F). Same substitution rate for codon position (Right, Y2L3F).

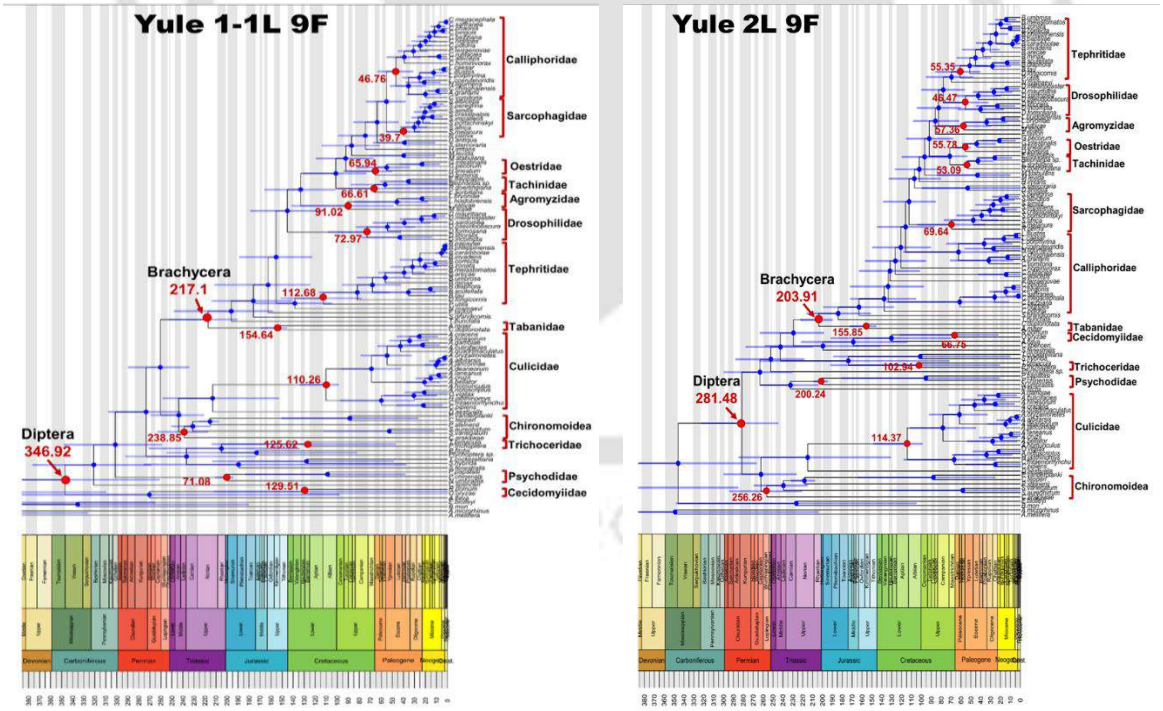


Figure 6.2: Time calibrated phylogeny of Diptera using nine fossil calibration using Yule model in BEAST. Different substitution rate for codon position (Left, Y1-1L9F). Same substitution rate for codon position (Right, Y2L9F).

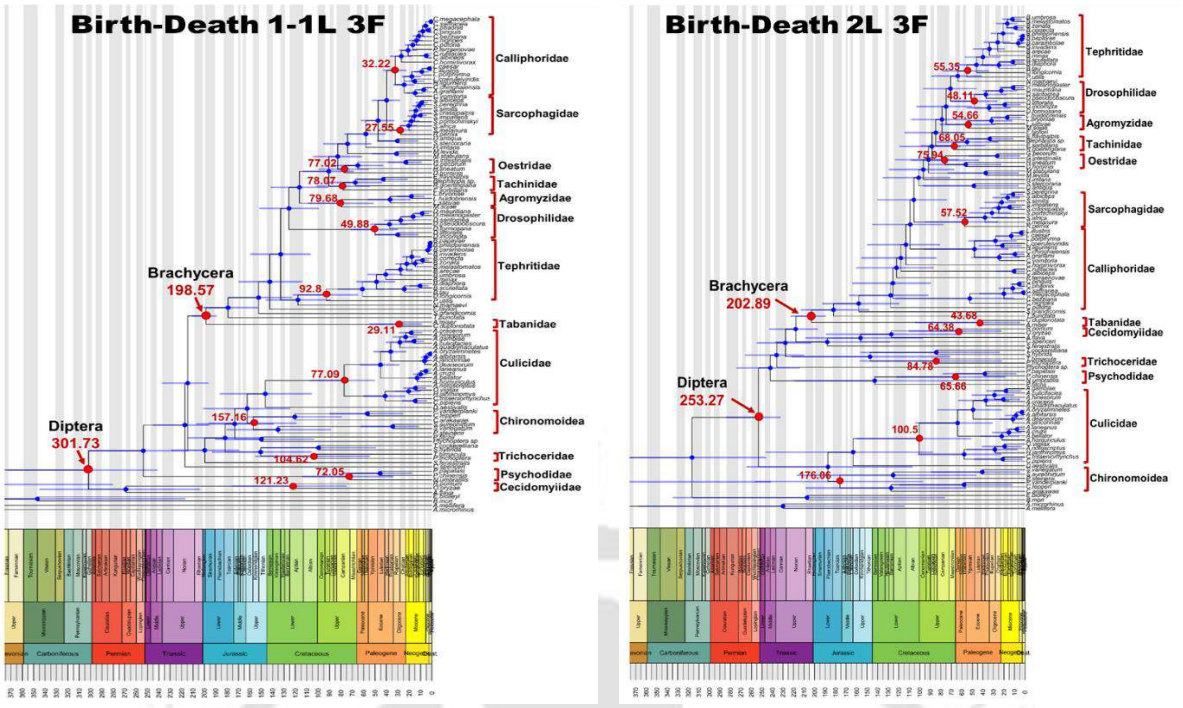


Figure 6.3: Time calibrated phylogeny of Diptera using three fossil calibration using Birth-Death model in BEAST. Different sustitution rate for codon position (Left, BD1-1L3F). Same substitution rate for codon position (Right, BD2L3F).

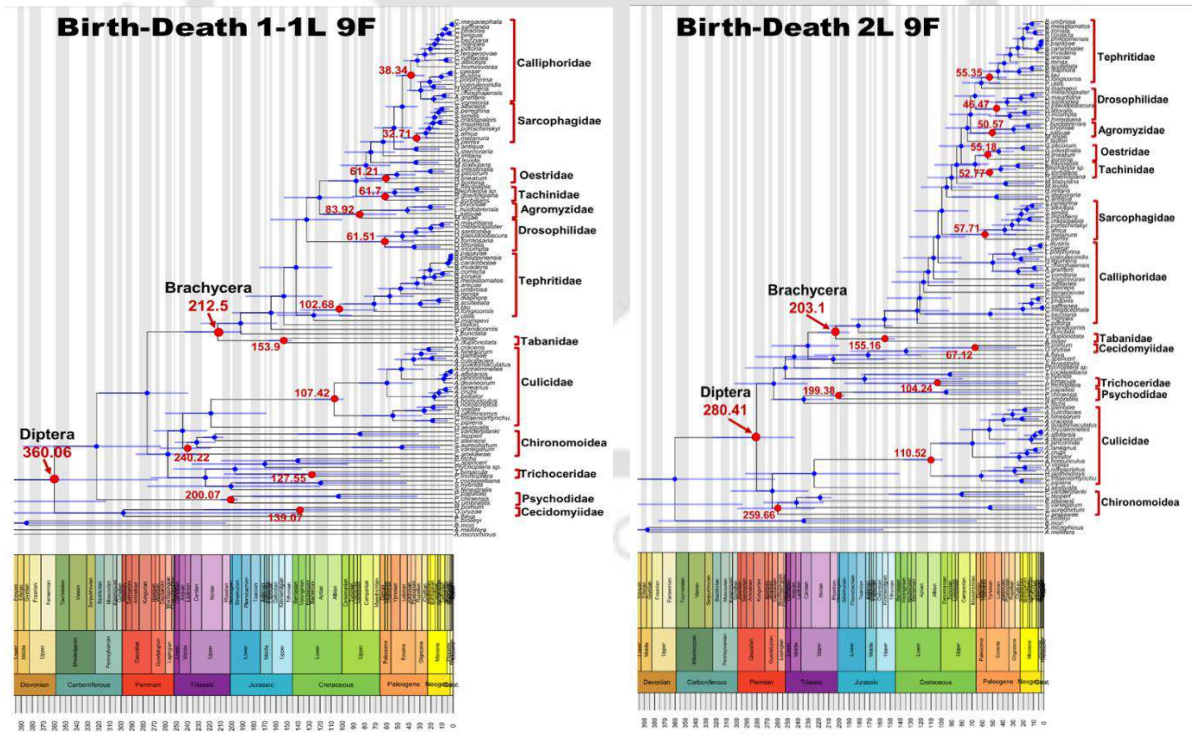


Figure 6.4: Time calibrated phylogeny of Diptera using nine fossil calibration using Birth-Death model in BEAST. Different substitution rate for codon position (Left, BD1-1L9F). Same substitution rate for codon position (Right, BD2L9F).

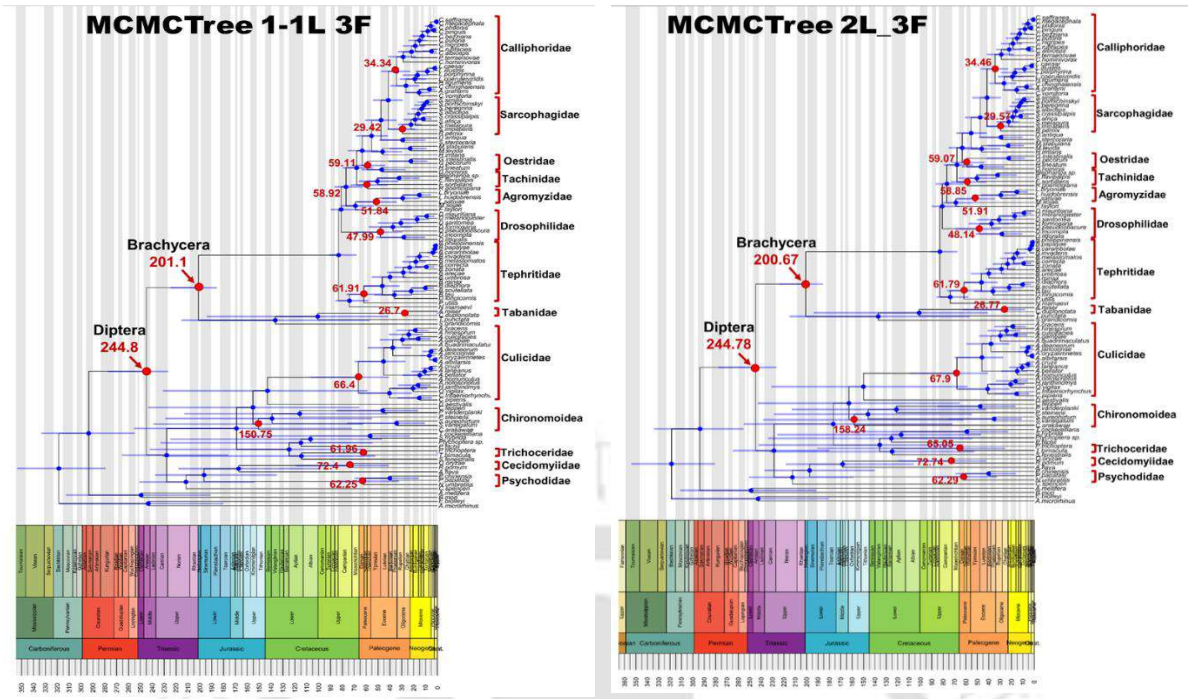


Figure 6.5: Time calibrated phylogeny of Diptera using three fossil calibration using Birth-Death model in MCMCTree. Different substitution rate for codon position (Left, mcmctree1-1L3F). Same substitution rate for codon position (Right, mcmctree2L3F).

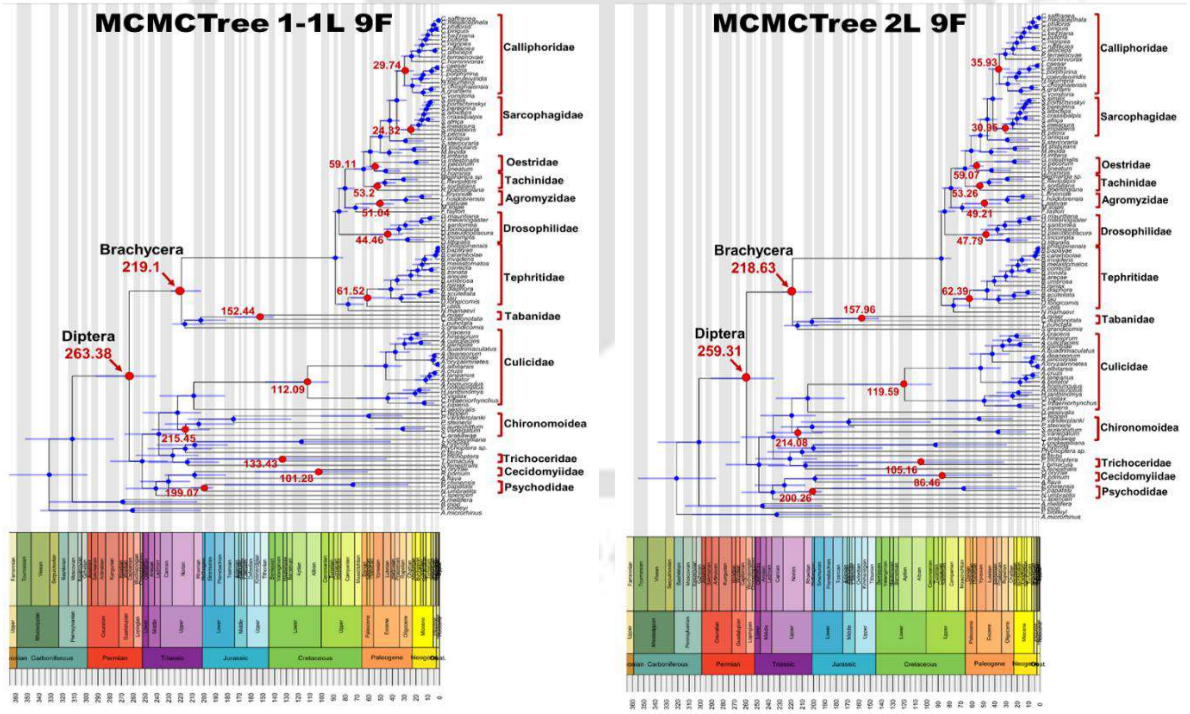


Figure 6.6: Time calibrated phylogeny of Diptera using three fossil calibration using Birth-Death model in MCMCTree. Different sustitution rate for codon position (Left, mcmctree1-1L9F). Same substitution rate for codon position (Right, mcmctree2L9F).

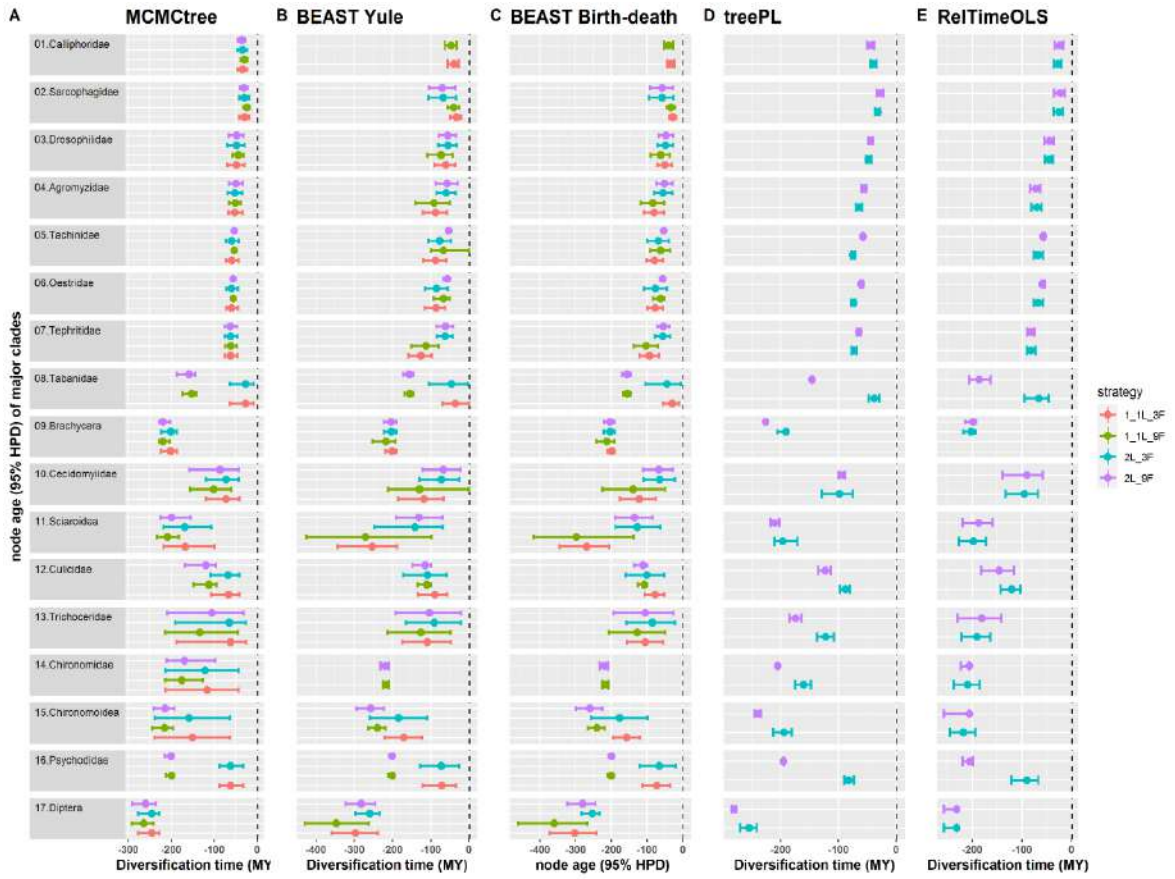


Figure 6.7: Estimated diversification time comparison of major clades of Diptera using different methods and strategies from mitochondrial data. (A) Bayesian estimation of divergence times using approximate likelihood method; (B) Bayesian estimation of divergence times using Yule branching process; (C) Bayesian estimation of divergence times using Birth-Death branching process; (D) Divergence time estimation using Penalized likelihood method; (E) Divergence time estimation using Ordinary least square method.

Our analysis suggests that the last common ancestor of extant Brachycera existed in the late Triassic which is earlier than the previous estimation[6]. The partitioned dataset with 9 fossils calibration shows earlier split of Brachycera lineages around ~217 Ma for Y1-1L9F and ~212.5 Ma for BD1-1L9F (Fig. 6.2 (Left), Fig. 6.4 (Left)). While other five MCC trees resulted ~200.7–203.9 Ma which are on or before Triassic–Jurassic (Tr-J) extinction event. Whereas, BD1-1L3F predicted Brachycera origin 198.6 Ma which is after the Tr-J event and previous estimation[6] (Fig. 6.3 (Left)). Our study places the Schizophora clade's emergence between the Late Jurassic and the Early Cretaceous, with the partitioned dataset (1-1L) putting it in the Late

Jurassic (143-163 Ma) and the non-partitioned dataset (2L) putting it in the Early Cretaceous (123-140 Ma).

Various lower Diptera or Nematocera families exhibit dissimilar estimated divergence times, such as the Culicidae family, which split around 100-114 Mya but partitioned and 3 fossil calibration MCC trees (Y1-1L3F and BD1-1L3F) show much later diversification of the Culicidae family at ~89 Mya and ~77 Mya, respectively (Fig. 6.1 (Left), Fig. 6.3 (Left)). The Chironomoidea superfamily diverged much earlier (Late Permian for non-partitioned dataset and Early Triassic for partitioned dataset) according to 9 fossil calibrations than 3 fossil calibrations (Jurassic) (Fig. 6.1 - Fig. 6.4). The Sciaroidea split occurred in the Early Cretaceous, as per estimates from non-partitioned datasets, while estimates from partitioned datasets reveal a far earlier divergence at the Permian Period (Fig. 6.7 B,C). Cecidomyiidae, the primary family of the Sciaroidea superfamily, diverged in the Early Cretaceous, according to partitioned data, whereas non-partitioned data suggest Late Cretaceous diversification (Fig. 6.7 B,C). Psychodidae divergence occurred during the Triassic–Jurassic (Tr-J) extinction event (~200 Myr), as estimated by 9 fossils, although 3 fossils suggest that Psychodidae diversification occurred substantially later, prior to the Cretaceous–Paleogene (K-Pg) extinction event (~66 Myr) (Fig. 6.7 B,C).

Diversification of Brachycera was estimated to have ~202-203 Mya by without partitioned molecular dataset; partitioned molecular dataset and 3 fossils calibration with Yule and Birth-Death process show diversification time of Brachycera 200.71 Myr and 198.57 Myr respectively, which indicate the time roughly around Tr-J extinction event (Fig. 6.1 - 6.4, Fig. 6.7 B,C). The Tabanidae family is a lower Brachycera family, and the estimated divergence time contradicts several methods used here, including 3 fossil calibrations that estimated Eocene epoch (~29-45 Mya) diversification while 9 fossil calibrations resulted in much earlier diversification of Tabanidae around Late Jurassic (~153-155 Mya) (Fig. 6.1 - 6.4, Fig. 6.7

B,C). The Schizophora origination period estimated in this study was earlier than previously reported, with BD2L9F estimated 123.49 Mya (Early Cretaceous) being the most recent and Y1-1L9F estimated 162.67 Myr (Upper Jurassic) being the earliest[3,6]. The major Schizophora lineages exhibit conflicting divergence times based on different approaches, such as Tephritidae and Agromyzidae appeared after the K-Pg (~66 Mya) event based on non-partitioned dataset, while as per partitioned dataset these families existed before the K-Pg event (Fig. 6.7 B,C). In contrast, Tachinidae and Oestridae family split on or after K-Pg event according to 9 fossils calibration (Tachinidae: ~52-66 Myr and Oestridae: ~55-65 Myr) but 3 fossils (Tachinidae: ~68-87 Myr and Oestridae: ~75-86 Myr) demonstrate it existed before the K-Pg event which is earlier than previously reported[6] (Fig. 6.1-6.4, Fig. 6.7 B,C). The Drosophila split occur after K-Pg (~66 Mya) extinction event estimated from this analysis except for the MCC tree Y1-1L9F (~73 Myr)[6] (Fig. 6.1-6.4, Fig. 6.7 B,C). According to the non-partitioned dataset, the Sarcophagidae diverged roughly Late Cretaceous to Early Paleogene (~57-69 Mya), however the partitioned dataset suggests that the Sarcophagidae diverged considerably more recently, around Late Eocene to Early Oligocene epoch (~39-27 Mya) (Fig. 6.7 B,C). The divergence of Calliphoridae around ~32-46 Mya as per partitioned dataset whereas, the non-partitioned dataset unable to distinguish Calliphoridae as monophyly. Overall, the estimation with nine fossil calibration resulted more earlier diversification time than with three fossil calibrations. The partitioned molecular data (1-1L substitution) with nine fossil calibrations display more earlier diversification time and BEAST Birth-Death branching prior strategy show earlier diversification time.

6.3.2 Lineage diversification analysis:

To assess diversification trends, decipher how the speciation occurred, and to observe impact of major events on the dated tree over time, the Lineages through time (LTT) plot was drawn from the result of BEAST analysis (Fig. 6.8). The LTT plots were generated by plotting log-

lineages through time based on species-level chronograms using the R package ape[67]. The gamma (γ) statistic was obtained using phylogenetic node distances to find whether the tempo of diversification differs considerably from the tempo of diversification in a pure-birth null model, and it may be used to measure whether diversification occurred in an early or late burst[68]. Accordingly, the negative (-ve) γ signifies early burst and positive (+ve) γ signifies late burst and early extinction. This analysis suggests that the number of fossil calibration points (3 and 9), molecular data partitioning (no partition (2L) and partitioned (1-1L), and branching procedure (Yule and Birth-Death) all have an impact on γ statistics estimation. Only Yule_3F_2L shows -ve γ value of -0.5573 with *p-value* of 0.57 and other MCC trees display +ve γ value. The MCC trees like Yule_9F_1-1L (γ = 1.7198, *p-value* = 0.085), Birth-Death_3F-1-1L (γ = 2.3859, *p-value* = 0.017) and Birth-Death_9F-1-1L (γ = 2.588, *p-value* = 0.0097) differed significantly from a model of constant rate diversification and supporting the late burst diversification model. Whereas, among all estimated γ values from the time trees only

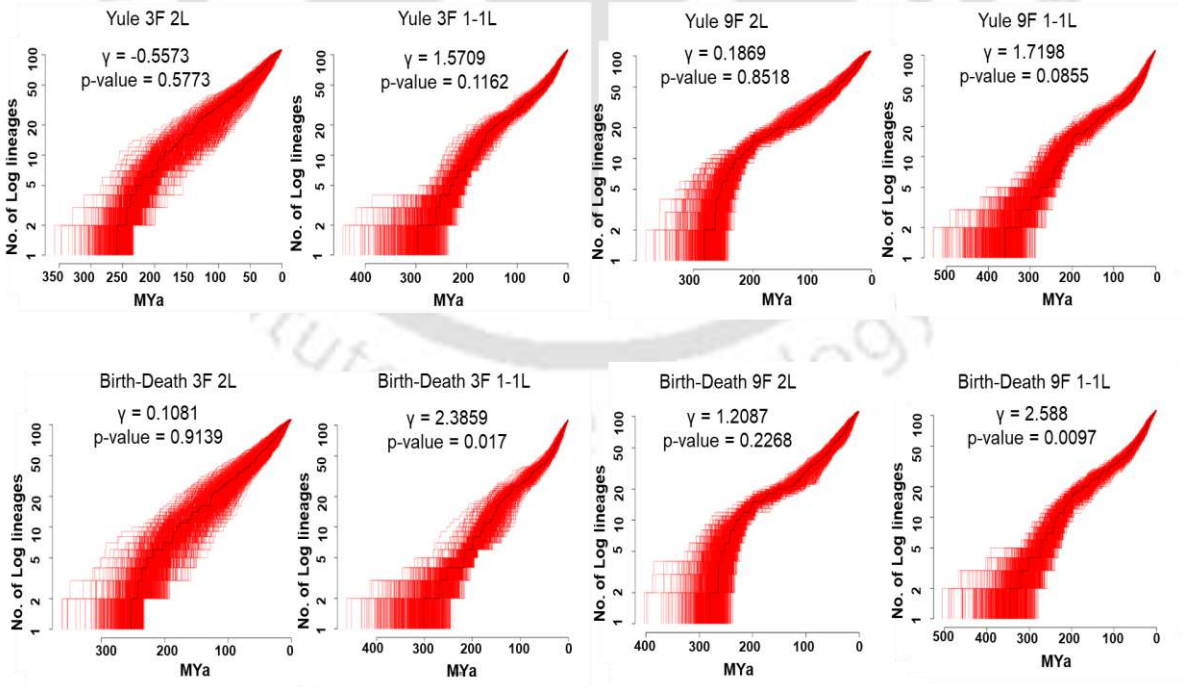


Figure 6.8: Lineages through time (LTT) plot of Diptera using Yule (Pure-birth, top row) and Birth-Death (bottom row) dated tree. Red lines are 1000 posterior trees sampled from BEAST analysis; the black line is MCC (maximum clade credibility) tree.

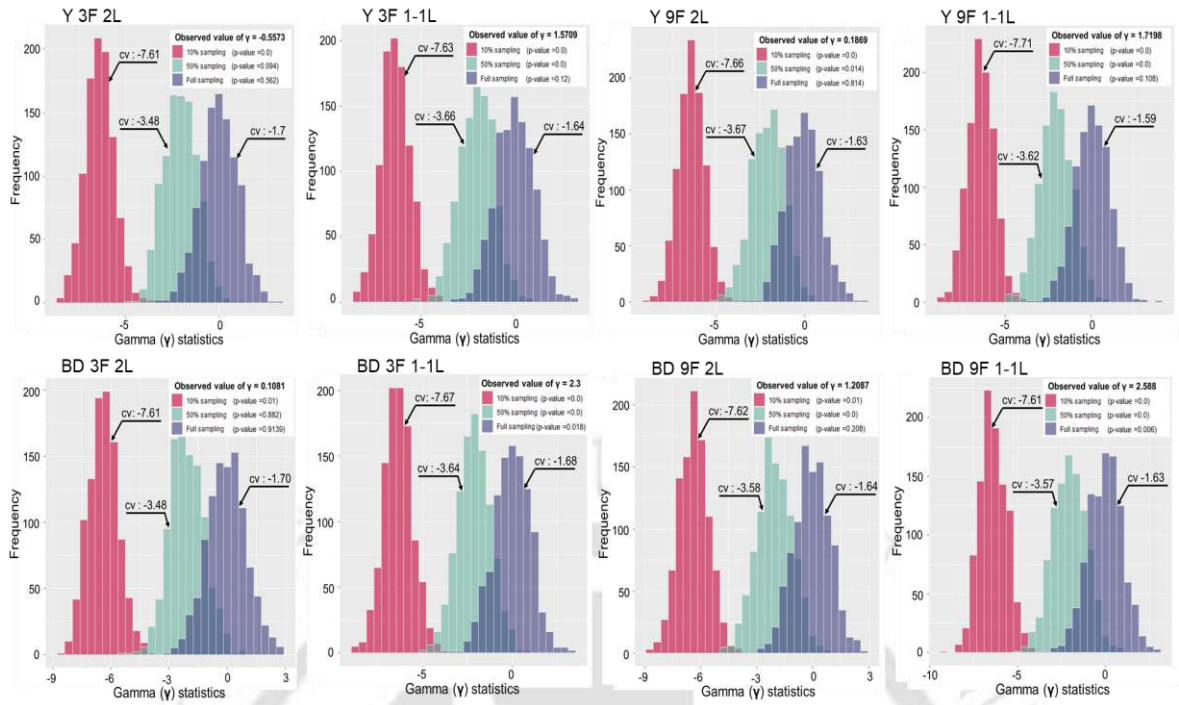


Figure 6.9: Monte Carlo Constant Rate (MCCR) test using Yule (Pure-birth, top row) and Birth-Death (bottom row) dated tree. MCCR analysis was run to generate 1000 null phylogenies produced under a Yule process using different sampling (ρ =1, 0.5 and 0.1).

Yule_9F_2L (γ = 0.1869, *p-value* = 0.85) and Birth-Death_3F-2L (γ = 0.1081, *p-value* = 0.91) show satisfactorily significant low γ value (nearer to 0), thus rejecting an early- or late-burst of diversification.

Positive values of γ as estimated by seven of eight MCC trees can be attributed to either increased diversification rates or species turnover, since recently diverged lineages have not been "pruned" by extinction, resulting in an overabundance of nodes closer to the present ("pull of the present")[86,87]. According to a study published in 2010 by Liow et al., when trees were simulated under a birth-death process with varying rates of speciation, γ tended to be positivelybiased and difficult to identify slowdowns in net diversification because short branches at the tree's tips tend to mask the slowdown signal[88].

To test for significance a critical value was calculated by running a Monte Carlo Constant Rate (MCCR) analysis using R package. The MCCR test produced all negative critical values in

1000 simulations with varied sampling (ρ =1, 0.5, and 0.1) of taxa formed under a pure-birth model employing the mccr (Fig. 6.9). With ρ = 0.1 MCCR test yielded critical value of γ for all MCC trees ~-7.6, *p-value* = 0.0. With ρ = 0.5 seven out of eight MCC trees produced significant critical value of γ ~-3.6; except for BD_3F_2L, which produced insignificant critical value of γ of -3.48, *p-value* = 0.882. With ρ = 1 seven out of eight MCC trees produced insignificant critical value of γ of ~-1.6 except for BD_9F_1-1L, which provided a significant critical value of γ -1.63, *p-value* = 0.006.

As a result of this experiment, it appears that by decreasing the sampling fraction (ρ), the critical value of γ significantly become less than 0 ($p < 0.01$), preferring the early burst and then deceleration of diversification. Since the number of taxa included in this study is limited in comparison to the overall numbers of fly linages found in nature, therefore the MCCR test implies that diversification occurred mostly near the root and a preference towards early diversification. This kind of phenomenon have long recognized by scientists that insufficient taxonomic sampling would result in more branching deeper in a phylogeny (on average)[68,89].

6.3.3 Trait-Independent Diversification rate shifts:

The likelihood of the BAMM MCMC reached convergence, and the post-burn-in (25% and 10%) ESS values were >200 (except Y_3F_2L, although, un-burn ESS >200) for rate shift analyses of all eight MCC trees from BEAST analysis. A model with one shift received a higher BF in comparison to the models with zero, two, three, and four shifts for all trees. The BAMM analysis shows about 2 to 3 diversification rate shifts in the time tree. The trees from non-partitioned molecular data (2L) display 2 rate shifts at node 121 (Culicidae) and 157(BD) /156(Y) (Brachycera). Whereas, partitioned molecular data (1-1L) display 3 rate shifts at node 135 (Culicidae), 157 (Brachycera) and 197 (Calliphoridae and Sarcophagidae) (Fig. 6.10). It was also found that having three rate shifts by 1-1L datasets had a higher probability than having two rate shifts by 2L datasets (Fig. 6.10).

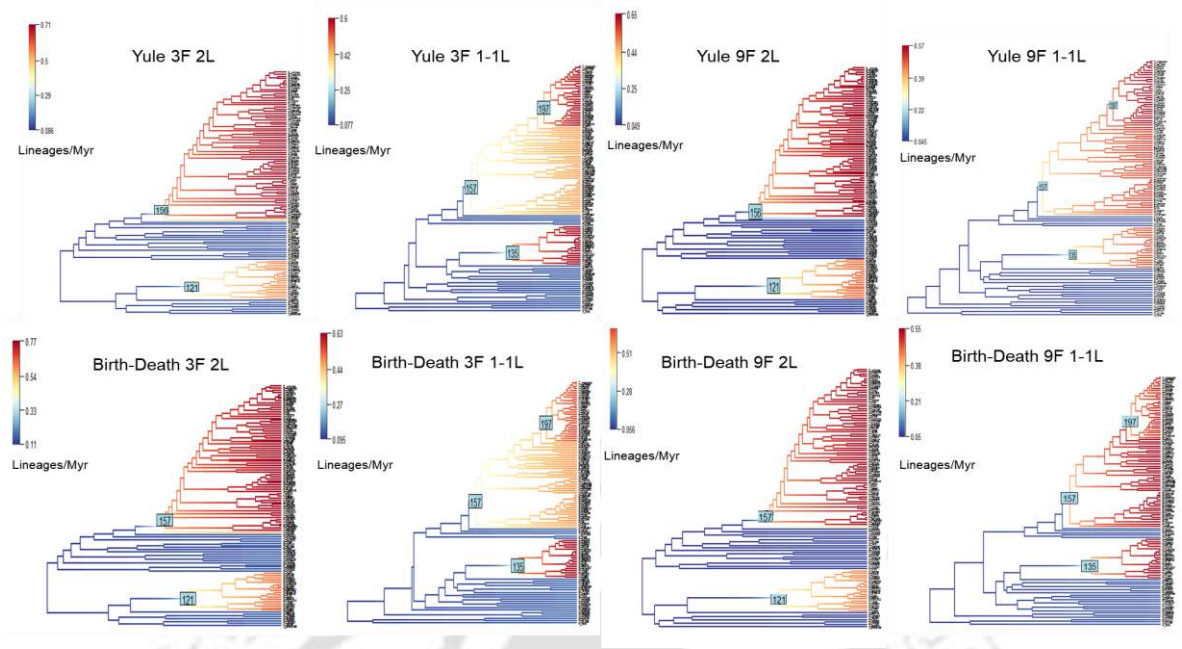


Figure 6.10: Phylorate plot of Diptera with branches coloured according to net diversification rate (Myr−1), resulting from Bayesian Analysis of Macro evolutionary Mixtures (BAMM) of 8-time trees. Light blue rectangular box indicates diversification rate shifts.

The mean evolutionary rate estimation by different softwares (BAMMTools, geiger and ape) in MCC trees show that BD_3F_2L tree exhibit highest speciation and extinction rate by BAMMTools (λ = 0.421875 lineages /My, μ = 0.359482 lineages /My) whereas, estimation by ape displays highest speciation and extinction rate for BD_3F_1-1L tree (λ = 0.022755 lineages /My, μ = 0.01282 lineages /My). The net diversification rate estimated by BAMMTools shows highest net diversification for BD_3F_1-1L tree (net.div = 0.076792 lineages /My) and ape shows highest net diversification for BD_3F_2L tree (net.div = 0.01426 lineages /My) (Table 6.2). Further, the rate through time analysis provides a robust understanding of macroevolutionary dynamics that have shaped the distribution of Diptera species. Mainly the nodes with rate-shifts are only showing in the figures suggests that Schizophora of Brachycera and Culicidae family of Nematocera have experienced vigorous rise in their speciation rate. Background Diptera (except Schizophora and Culicidae) shows slow rate of evolution. Therefore, the net-diversification of Schizophora and Culicidae significantly higher in nature.

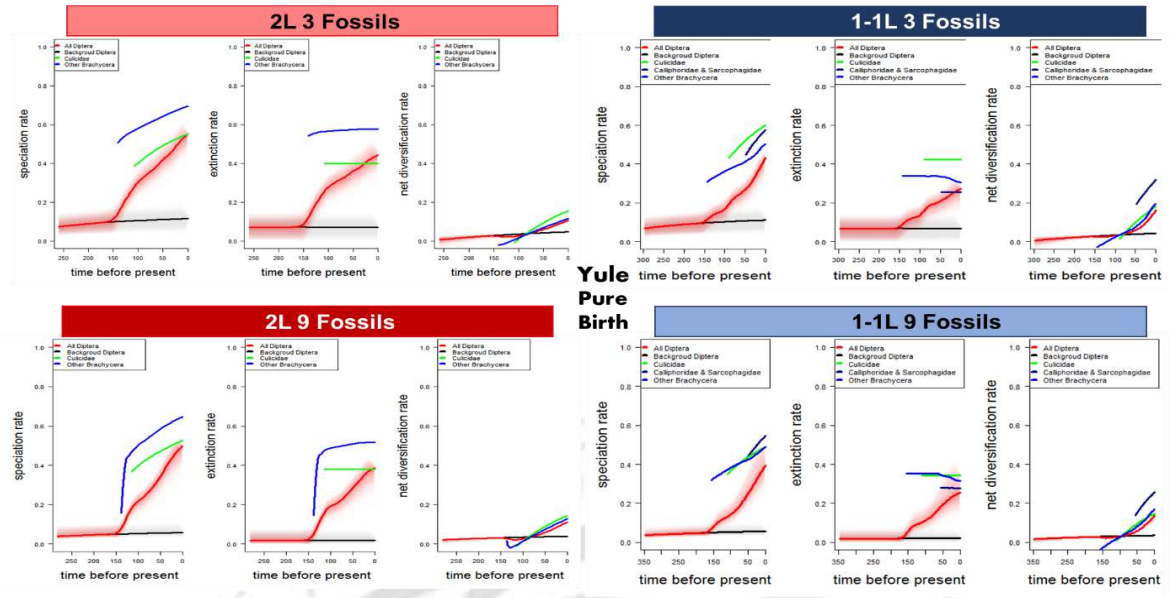


Figure 6.11: Rate through time (RTT) plot from BAMM analysis using Yule (Pure birth) dated tree. Red : All Diptera, Blue : Schizophora (Upper Brachycera), Green : Culicidae, Black : Background Diptera

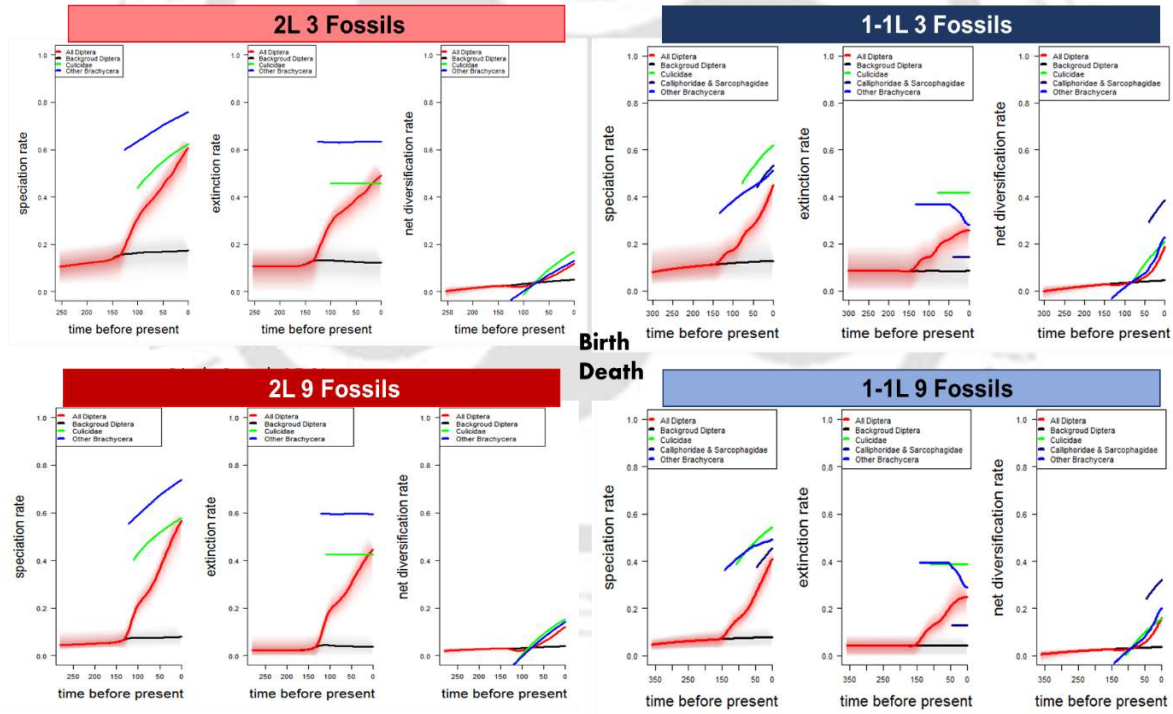


Figure 6.12: Rate through time (RTT) plot from BAMM analysis using Birth-death dated tree. Red: All Diptera, Blue: Schizophora (Upper Brachycera), Green: Culicidae, Black: Background Diptera

Table 6.2: Mean evolutionary rate derived by different tools (BAMMTools, Geiger, and Ape) from eight MCC trees

	BAMMTools			**geiger**			**manual calculation in ape**		
Calibration Approach (MCC tree)	Spec. rate (λ) (lineages /My)	Ext. rate (μ) (lineages /My)	Net.div (λ-μ)	Net.div (e=0)	Net.div (e=0.45)	Net.div (e=0.90)	Spec. rate (b)	Ext. rate (d)	Net div. rate (b-d)
BD_9F_1-1L	0.23603	0.171525	0.064505	*0.01118*	*0.010568*	*0.006781*	0.019297	0.011465	*0.007832*
BD_9F_2L	0.322819	0.262862	0.059957	0.014355	0.01357	0.008707	0.015944	0.005521	0.010423
BD_3F_1-1L	0.274878	0.198086	**0.076792**	0.013341	0.012612	0.008092	**0.022755**	**0.01282**	0.009935
BD_3F_2L	**0.421875**	**0.359482**	0.062393	**0.015893**	**0.015024**	**0.00964**	0.014949	0.000689	**0.01426**
Y_3F_2L	0.389553	0.330313	0.05924	0.015546	0.014696	0.00943	0.013523	*0*	0.013523
Y_3F_1-1L	0.261173	0.190727	0.070446	0.01357	0.012828	0.008231	0.018485	0.007781	0.010704
Y_9F_2L	0.299635	0.241153	*0.058482*	0.0143	0.013519	0.008674	*0.01284*	0.001055	0.011786
Y_9F_1-1L	*0.229897*	*0.169727*	0.06017	0.011603	0.010969	0.007038	0.015813	0.007129	0.008684

Table 6.3: Evolutionary rates of three major clades from Bayesian time trees (MCC trees) generated using different calibration and branching process

Calibration Approach	BAMMTools estimation	Calliphoridae and Sarcophagidae	Culicidae	Other Brachycera
BD_9F_1-1L	Mean speciation rate (λ) (lineages/My)	***0.483038***	0.447557	**0.501527**
	Mean extinction rate (μ) (lineages/My)	0.200618	0.342382	0.378572
	net.div	0.28242	0.105175	0.122955
BD_9F_2L	Mean speciation rate (λ) (lineages/My)		0.481569	0.701175
	Mean extinction rate (μ) (lineages/My)		0.382909	0.604848
	net.div		0.098661	0.096327
BD_3F_1-1L	Mean speciation rate (λ) (lineages/My)	0.525409	0.535804	0.459573
	Mean extinction rate (μ) (lineages/My)	**0.166709**	0.38573	0.327759
	net.div	**0.358701**	**0.150074**	**0.131814**
BD_3F_2L	Mean speciation rate (λ) (lineages/My)		**0.550265**	**0.723367**
	Mean extinction rate (μ) (lineages/My)		**0.436029**	**0.637549**
	net.div		0.114236	0.085819
Y_3F_2L	Mean speciation rate (λ) (lineages/My)		0.489128	0.672465
	Mean extinction rate (μ) (lineages/My)		0.38289	0.595676
	net.div		0.106238	***0.076789***
Y_3F_1-1L	Mean speciation rate (λ) (lineages/My)	**0.554136**	0.515064	***0.439967***
	Mean extinction rate (μ) (lineages/My)	0.270927	0.388427	***0.326604***
	net.div	0.283209	0.126637	0.113363
Y_9F_2L	Mean speciation rate (λ) (lineages/My)		0.44427	0.60863
	Mean extinction rate (μ) (lineages/My)		0.348827	0.521963
	net.div		***0.095443***	0.086667
Y_9F_1-1L	Mean speciation rate (λ) (lineages/My)	0.541533	***0.409545***	0.48787
	Mean extinction rate (μ) (lineages/My)	**0.337776**	***0.307828***	0.384788
	net.div	***0.203758***	0.101717	0.103082

In RTT plot blue line corresponds to the diversification process of upper Brachycera beginning ~150 mya, whereas green line corresponds to Culicidae beginning (~120~95) mya and navy-blue line corresponds to Calliphoridae and Sarcophagidae beginning ~50 mya detected only data-partitioned analysis of BEAST (Fig. 6.11, 6.12). The red line indicated all Diptera which began diversification before 250 mya but rise of speciation started ~150 mya. The diversification rate of different calibrated time trees is not congruent, Y_3F_1-1L display highest mean for speciation rate, λ = 0.554 lineages/My for Calliphoridae and Sarcophagidae whereas, same tree shows lowest λ = 0.439 lineages/My for Brachycera clade. The BD_3F_1-1L tree exhibit highest net diversification rate all three major clades (Calliphoridae and Sarcophagidae: 0.3587 lineages/My; Culicidae: 0.15 lineages/My; Brachycera: 0.1318 lineages/My). The net diversification rate of Calliphoridae and Sarcophagidae clade always

found to be higher than the other two clades in 1-1L dataset. The estimated evolutionary rates three major clades tabulated in Table 6.3. Overall, the BAMM analyses suggest a comprehensive increase of Diptera diversification rate through time since Late Cretaceous or Early Eocene depending on the MCC trees used in the analysis (Fig. 6.19, 6.20). Whereas, speciation and extinction rate both got pace during early Cretaceous after the emergence of lineages with high diversification namely Schizophora and Culicidae.

6.3.4 Diversity-Dependent Diversification:

Since gamma statistics from lineage through time (LTT) plots provide inadequate information to identify diverse kinds of diversification dynamics, numerous alternative modelling techniques have been developed to detect non-homogeneous diversification[11,87]. Diversity-dependence has also been presented as an explanation for patterns of species richness in which the rate of speciation varies as a function of the number of taxa present at any particular time. [90]. This has already been modeled in a range of studies in order to look for signatures of an adaptive radiation[23,91]. Diversity-dependence is a valuable technique since it appears to be consistent with adaptive radiation theory, which states that when niches are filled, the rate of diversification should decline[8].

Herein, diversity-dependent diversification models estimated limited carrying capacity for six out of eight MCC trees for complete Diptera lineages (Table 6.4). Diversity dependence estimated for different MCC trees were not similar such as, Y1-1L3F, Y1-1L9F, BD1-1L3F and BD1-1L9F favored DD+EX model in which extinction increases exponentially with accumulated diversity. The MCC trees like Y2L9F and BD2L9F favored DDX+E model in which speciation declines exponentially with diversity and non-zero extinction. The Y2L3F provides support for the DDL model, which has a greater carrying capacity (528.2) and shows that speciation declines linearly as accumulated diversity increases without extinction. These seven MCC trees better fitted than CR model. Whereas, BD2L3F supported CR model

with an unlimited carrying capacity where diversity does not dependent on speciation and extinction rate.

We also tested whether clade wide shift occurred in diversification parameters like intrinsic speciation rate, extinction rate or clade-level carrying capacity (Table 6.5). The result suggested that six of eight MCC trees supported only shift in carrying capacity (K). Whereas, Y1-1L9F favored a model with shift in carrying capacity and extinction rate and BD1-1L9F favored a model with shift in carrying capacity and speciation rate. The results of each MCC trees reveal that carrying capacity increases after the shift, and the extinction rate for Y1-1L9F increases after the shift, whereas the speciation rate for BD1-1L9F drops after the shift. In this analysis, the timing of the shift was calculated from each MCC trees, and the results indicated that the shift occurred mostly during the Cretaceous period, ranging from Early to Late epochs, depending on the MCC trees employed (Y1-1L9F: 74.5 Ma (latest), BD2L3F: 127.7 Ma (earliest)).

Furthermore, the BDL (birth–death likelihood) analysis revealed that an RV (rate-variable) model fit seven of the eight MCC trees well, with yule3rate supported for six MCC trees and yule2rate favored for Y2L3F. BD2L3F, on the other hand, preferred an RC (rate-constant) model pureBirth to have the best fit model. The six MCC trees supported by yule3rate model show three distinct diversification rates in which diversification rate r2 is lesser than r1 and r3. The time of rate shift detected by yule3rate in incongruent with each other such as, BD1-1L3F, BD1-1L9F and Y1-1L3F detected timing of shift (st2) at Paleogene period. The MCC trees like BD2L9F, Y2L9F and Y1-1L9F detected st2 at Cretaceous period. However, this improvement was not significant when compared to a null sample of 5000 simulated trees (P = 0.9998).

Table 6.4: Best fitted model from Diversity-Dependent Diversification analyses in DDD

MCC trees	Model Name	Model	NP	logL	AICc	Lambda	Mu	K	r
Y2L3F	Linear dependence of speciation rate without extinction	DDL	2	-583.1086	1170.327	0.0152	NA	528.2	NA
Y1-1L3F	Exponential dependence of extinction rate	DD+EX	3	-573.7903	1153.803	0.0241	2.00E-05	115.26	NA
Y2L9F	Exponential dependence of speciation rate with extinction	DDX+E	3	-591.6743	1189.571	0.1334	0.01276	170.07	NA
Y1-1L9F	Exponential dependence of extinction rate	DD+EX	3	-592.8279	1191.878	0.0214	1.00E-05	115.18	NA
BD2L3F	Linear independence of speciation rate with extinction	CR	2	-574.8681	1153.846	0.0149495	0.000690444	Inf	NA
BD1-1L3F	Exponential dependence of extinction rate	DD+EX	3	-563.1759	1132.574	0.027	1.00E-04	113.83	NA
BD2L9F	Exponential dependence of speciation rate with extinction	DDX+E	3	-583.3142	1172.851	0.2009	0.01796	128.38	NA
BD1-1L9F	Exponential dependence of extinction rate	DD+EX	3	-583.2874	1172.797	0.0239	4.00E-05	112.87	NA

Table 6.5: Best fitted model from Diversity-Dependent Diversification with a shift in the parameter analyses in DDD

MCC trees	Model Name	Model	df	loglik	AICc	lambda_1	mu_1	K_1	lambda_2	mu_2	K_2	t_shift
Y2L3F	diversity dependent Key shift in K	SR_1	5	-583.152	1176.87	0.01725204	6.12E-06	68.09265	0.01725204	6.12E-06	312.443	120.9641
Y1-1L3F	diversity dependent Key shift in K	SR_1	5	-574.0697	1158.705	0.04405877	0.01735861	61.82431	0.04405877	0.01735861	117.5654	109.0269
Y2L9F	diversity dependent Key shift in K	SR_1	5	-591.3179	1193.202	0.01874062	0.00103779	45.28735	0.01874062	0.00103779	252.3958	102.9459
Y1-1L9F	diversity dependent Key shift in K and mu	SR_2	6	-590.7531	1194.306	0.02142852	3.18E-03	48.52615	0.02142852	0.002527992	258.5021	74.55976
BD2L3F	diversity dependent Key shift in K	SR_1	5	-573.3056	1157.177	0.02156067	0.002273435	25.50601	0.02156067	0.002273435	252.025	127.7624
BD1-1L3F	diversity dependent Key shift in K	SR_1	5	-562.9905	1136.547	0.04890525	0.02219875	79.90073	0.04890525	0.02219875	119.8953	82.77272
BD2L9F	diversity dependent Key shift in K	SR_1	5	-580.5277	1171.622	0.07694906	0.02071825	68.32667	0.07694906	0.02071825	110.3091	104.2463
BD1-1L9F	diversity dependent Key shift in K and lambda	SR_3	6	-583.9801	1180.76	0.03936637	0.01950363	117.3639	0.0232748	0.01950363	304.6017	125.7577

The slowdown in lineage diversification rates through time is a well-known pattern of diversification[24,91]. This density-dependent tendency can be explained by an early greater opportunity for occupying fresh ecological niches while competitive pressure is minimal, allowing for a rapid diversification rate[91–93]. As the niche gets saturated and the competition for ecological space grows, the rate of speciation decreases. The increase in carrying capacity after the Cretaceous period shift can be attributed to either a decline in speciation rate and an increase in extinction rate of previously originated species; or the emergence of a new environment as a result of spread to a new area or external influences modifying the environment; or a key innovation, such as acquiring organismal phenotypes that promote by allowing escape from competition for niche space[77,94].

6.3.5 Tempo and mode of lineage diversification:

For the majority of resampled trees, TESS' marginal likelihood model comparison revealed a preference (Bayes factors (BF) > 100) for variable-rates models (episodically-varying and explicit mass extinction rate birth–death models) over a time-homogeneous birth–death mode, confirming our expectation that time-homogeneous processes cannot explain lineage diversification dynamics in Diptera (Table 6.6). Moreover, Model-fit comparisons based on a set of 100 trees evenly sampled from de posterior distribution reveal that most of the trees provide decisive support (BF > 100) for the episodically varying rate model (Figure not shown here).

According to the CoMET analysis, post-Cretaceous speciation rates increased after crossing the Palaeocene–Eocene boundary. This analysis also shows that the Bayes factor provides strong support in the rise of speciation rate (BF >2.0) following the Cretaceous–Paleogene (K–Pg) extinction event (66 Mya) (Fig. 6.13). There was no evidence for mass extinction in these Diptera (BF <2.0) as per CoMET analysis (Fig. 6.14). Rates of extinction, which are typically

Table 6.6: Comparison between different models inferred from TESS based Bayes Factor and Marginal Likelihood

			Bayes Factor							Bayes Factor		
	M0		M1				M0			M1		
		ML	CBD	EBD	MEBD			ML	CBD	EBD	MEBD	
Y_3F_2L	CBD	-2265.3	0	-4391.42	-1839.17		BD_3F_2L	CBD	-2364.32	0	-4467.39	-2460.39
	EBD	-69.5865	4391.423	0	2552.257			EBD	-130.63	4467.389	0	2007.003
	MEBD	-1345.72	1839.166	-2552.26	0			MEBD	-1134.13	2460.386	-2007	0
Y_3F_1-1L	CBD	-2874.63	0	-5188.35	-3755.82		BD_3F_1-1L	CBD	-3086.3	0	-5412.46	-4050.58
	EBD	-280.456	5188.349	0	1432.53			EBD	-380.067	5412.456	0	1361.88
	MEBD	-996.721	3755.819	-1432.53	0			MEBD	-1061.01	4050.575	-1361.88	0
Y_9F_2L	CBD	-2146.01	0	-4154.91	-1802.29		BD_9F_2L	CBD	-2317.87	0	-4269.2	-2618.95
	EBD	-68.5523	4154.906	0	2352.613			EBD	-183.265	4269.204	0	1650.255
	MEBD	-1244.86	1802.292	-2352.61	0			MEBD	-1008.39	2618.949	-1650.26	0
Y_9F_1-1L	CBD	-2597.4	0		-2241.94		BD_9F_1-1L	CBD	-2796.88	0	-4909.6	18646.08
	EBD	-279.098	4636.601	0	2394.659			EBD	-342.078	4909.597	0	23555.67
	MEBD	-1476.43	2241.942		0			MEBD	-12119.9	-18646.1	-23555.7	0

ML: Marginal Likelihood; EBD: EpisodicBD, CBD: ConstBD, MEBD: MassExtinctionBD

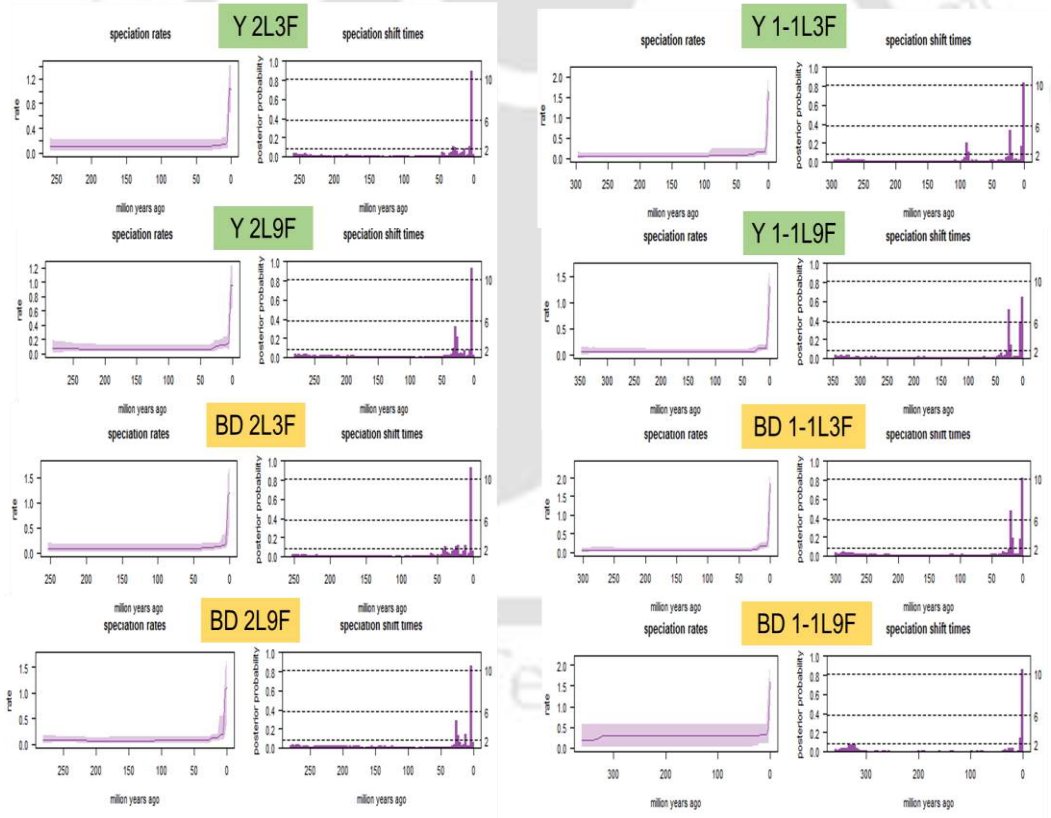


Figure 6.13: Rates of speciation through time estimated from the MCC tree using the CoMET function in TESS. Plots highlight the variation of the CoMET results obtained from different trees. All analyzes used a minimum threshold value of effective samples (ESS) of 500.

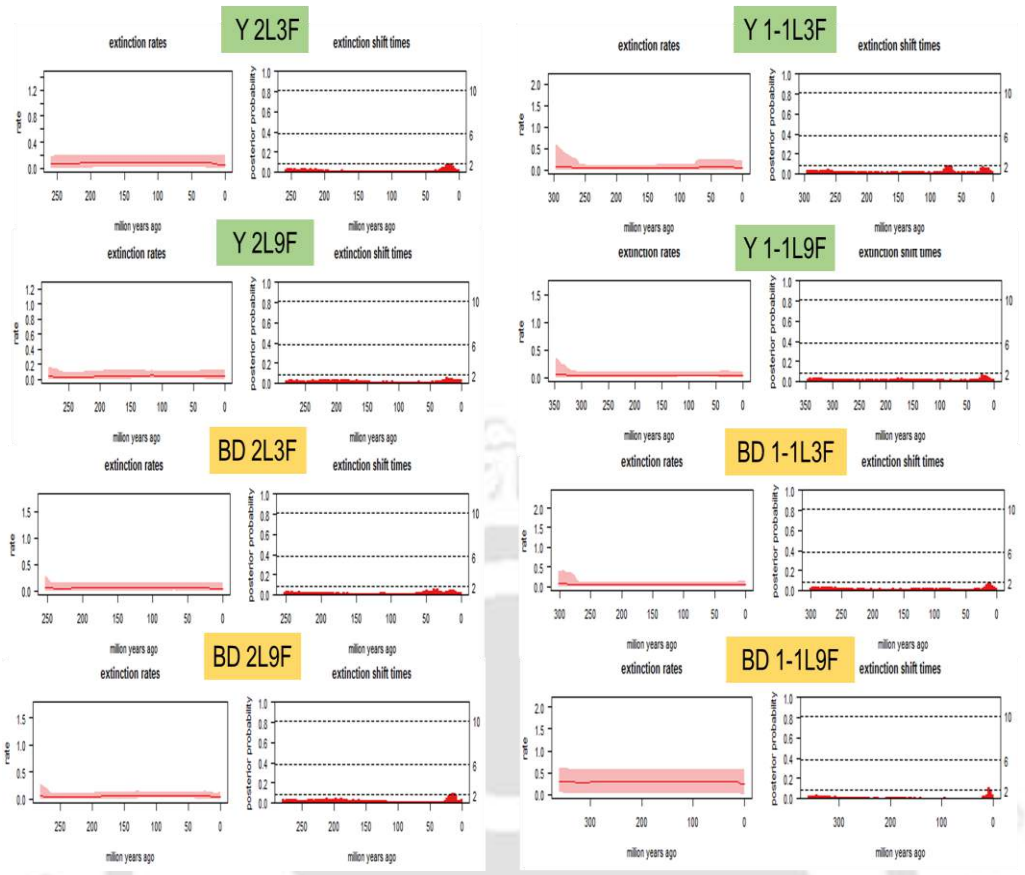


Figure 6.14: Rates of extinction through time estimated from the MCC tree using the CoMET function in TESS. Plots highlight the variation of the CoMET results obtained from different trees. All analyzes used a minimum threshold value of effective samples (ESS) of 500.

difficult to estimate using phylogeny alone, were estimated to be low and constant through time[95,96]. Although, little rise in extinction rate observed after 50 MY (BF ~2.0) in all MCC trees. Only Y1-1L3F display an additional rise in extinction rate around ~75-65 mya (BF ~2.0) and it can be corresponded to K–Pg extinction event.

6.3.6 Tree-Wide diversification rate shift:

An evolutionary model was proposed by T. Stadler namely birth–death-shift process, in which all rates are estimated simultaneously throughout the evolution of a phylogeny, instead of looking for local slope changes in the LTT plot[28]. It has several advantages over the LTT plot, including the following: this likelihood framework requires no input besides the dated tree; it allows slope changes without predicting a false rate change; it can infer rates in the recent and

distant past by assuming random sampling for both complete and incomplete phylogenies; this likelihood technique is resilient to flaws in speciation time estimates and unresolved nodes (polytomies); this method using the complete tree can easily explain the stochastic factors that account for the speciation pattern of a limited number of lineages in a short time span[28].

We used this maximum likelihood-based birth-death-shift (BDS) model with mass extinctions disabled to explicitly test for global changes in the diversification of Diptera lineages across time, with 0 to 4 rate shifts allowed during evolutionary peroid[28]. The BDS modelling using the 8-time trees (MCC) displayed that one rate shift is most likely, as indicated by the lowest AIC value (except Y2L3F and BD2L3F (no shift)) (Table 6.7). Although the time of diversification rate shift from the rest of the MCC trees are incongruent, as determined by our analyses. The Y2L9F and BD2L9F show significant rate shift about 140 mya (P = 0.97) and 125 mya (P = 0.99). High initial rate of diversification (r2 = 0.0345 (Y2L9F), r2 = 0.0321 (BD2L9F)) and downshift in diversification rate (r1 = 0.0121 (Y2L9F), r1 = 0.0115 (BD2L9F)) has been observed. Whereas, Y1-1L9F and BD1-1L9F show significant rate shift about 40 mya (P = 0.94) and 65 mya (P = 0.96) (Figure 6.21, 6.22). Low initial diversification rate (r2 = 0.0105 (Y1-1L9F), r2 = 0.0144 (BD1-1L9F)) and followed by upshift in diversification rate (r1 = 0.0361 (Y1-1L9F), r1 = 0.0152 (BD1-1L9F). With the exception of Y1-1L9F, the turnover of other three MCC trees appear to be increase after the shift. Only Y1-1L3F and BD1-1L3F exhibit similar and significant rate shift ~50 million years ago (mya) (P = 0.93, 0.92) and AIC (1155.644, 1133.895) values were also lower than other estimation. The low initial rate of diversification (r2 = 0.0162 (Y1-1L3F), r2 = 0.0167 (BD1-1L3F)) and high turnover (ε2= 0.9978 (Y1-1L3F), 0.9983 (BD1-1L3F)), followed by increase in diversification rate (r1= 0.0231 (Y1-1L3F), 0.019 (BD1-1L3F)) and increase in turnover (ε1 = 0.9989 (Y1-1L3F), 0.9993 (BD1-1L3F)). High turnover indicates that both dipteran speciation and extinction rates are faster which increases after rate shift meaning extinction rate is in rise.

Although the turnover is always appearing to be less than 1, which suggest that extinction rate never exceeded speciation rate.

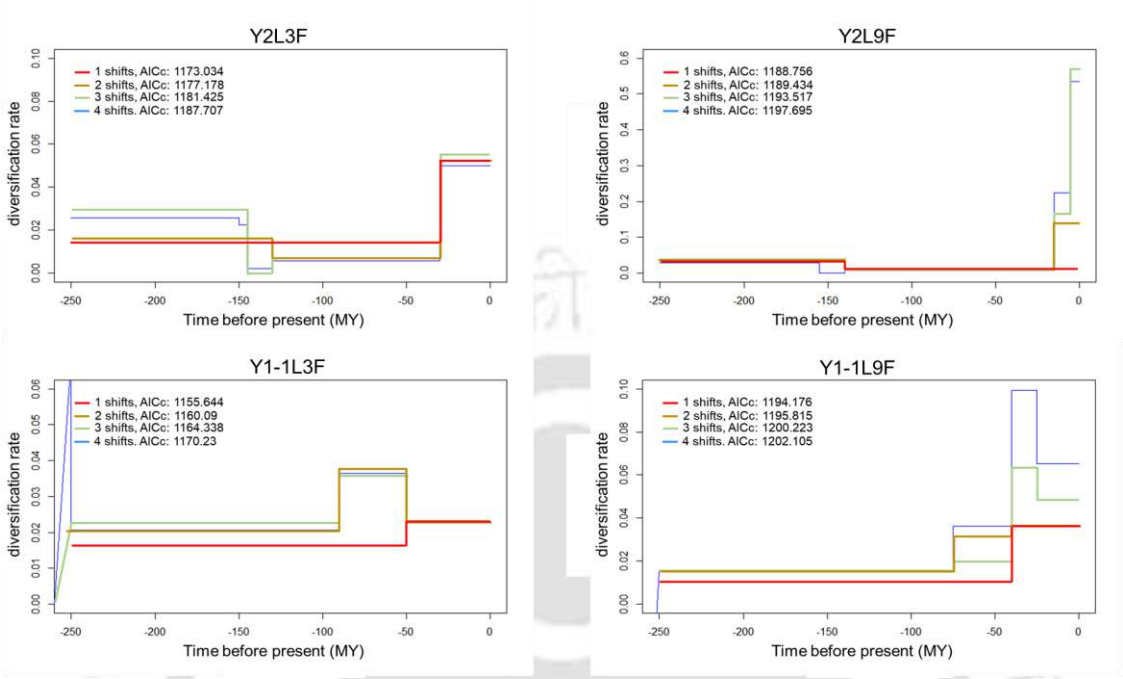


Figure 6.15: Maximum-likelihood diversification rate estimates (per million years) for Dipteran time tree generated from Yule branching process. 1-4 rate shifts allow for the diversification rate estimation.

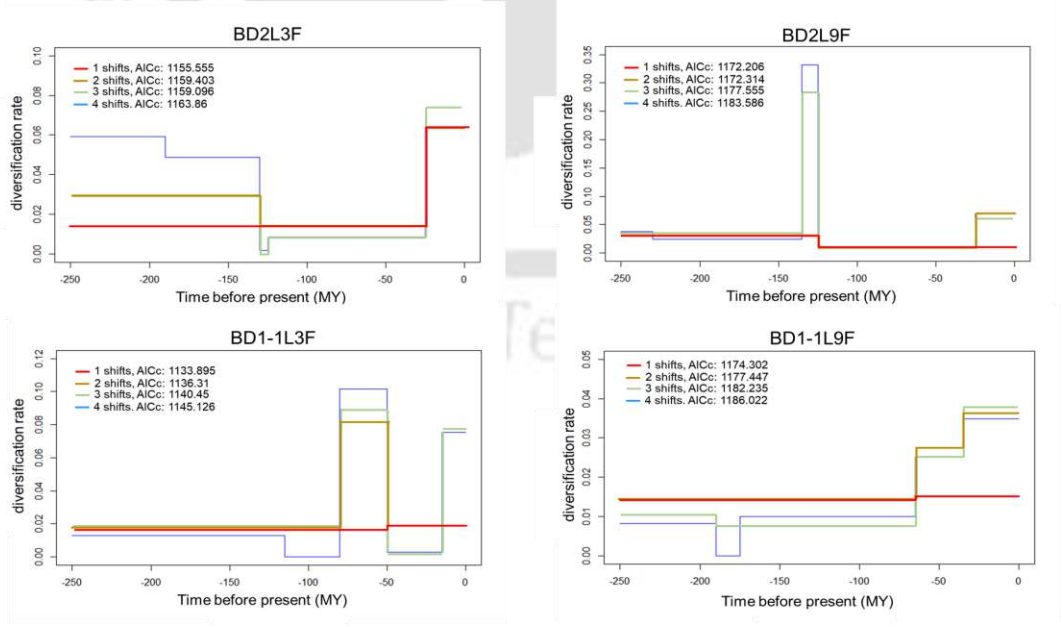


Figure 6.16: Maximum-likelihood diversification rate estimates (per million years) for Dipteran time tree generated from Birth-Death branching process. 1-4 rate shifts allow for the diversification rate estimation.

Table 6.7: Best fitted model from episodic diversification analyses in TreePar without mass extinction

Birth Death rate shift	Y2L3F	BD2L3F	Y1-1L3F	BD1-1L3F	Y2L9F	BD2L9F	Y1-1L9F	BD1-1L9F
Model	BD with No Shift time	BD with No Shift time	BD with 1 Shift time	BD with 1 Shift time	BD with 1 Shift time	BD with 1 Shift time	BD with 1 Shift time	BD with 1 Shift time
NP	2	2	5	5	5	5	5	5
logL	-583.2358	-574.8681	-572.5391	-561.6646	-589.0952	-580.8199	-591.8049	-581.8678
AICc	1170.582	1153.846	1155.644	1133.895	1188.756	1172.206	1194.176	1174.302
significance			0.9300224	0.9246575	0.9717743	0.9911795	0.9431579	0.9652097
DivRate1	0.0147	0.0143	0.0231	0.019	0.0121	0.0115	0.0361	0.0152
Turnover1	0.9991	0.9993	0.9989	0.9993	0.9993	0.9995	0.9975	0.9994
ShiftTime1	NA	NA	50	50	140	125	40	65
DivRate2	NA	NA	0.0162	0.0167	0.0345	0.0321	0.0105	0.0144
Turnover2	NA	NA	0.9978	0.9983	0.9891	0.9888	0.998	0.9979
ShiftTime2	NA	NA	NA	NA	NA	NA	NA	NA

Table 6.8: Best fitted model from episodic diversification analyses in TreePar with mass extinction

Mass extinction	Y2L3F	BD2L3F	Y1-1L3F	BD1-1L3F	Y2L9F	BD2L9F	Y1-1L9F	BD1-1L9F
Model	BD with no ME	BD with no ME	BD with 1 ME	BD with 1 ME	BD with no ME	BD with no ME	BD with 1 ME	BD with 1 ME
NP	2	2	4	4	2	2	4	4
logL	-583.2358	-574.8681	-572.7699	-561.6896	-593.636	-586.628	-591.8856	-581.8735
AICc	1170.582	1153.846	1156.106	1133.945	1191.382	1177.366	1194.337	1174.313
significance			0.9141754	0.9229694			0.9389132	0.9650314
DivRate	0.0147	0.0143	0.3729	0.3554	0.0118	0.0104	0.3148	0.2926
Turnover	0.9991	0.9993	0.9993	0.9994	0.9993	0.9995	0.9992	0.9994
ShiftTime1	NA	NA	50	50	NA	NA	75	65
survival probability	NA	NA	0.0173	0.0171	NA	NA	0.0149	0.0146

When mass extinctions are allowed, only four of the eight MCC trees constructed using partitioned datasets supported one mass extinction event, as shown in Table 6.8. However, the estimated time of mass extinction event of MCC trees were not similar, Y1-1L9F and BD1-1L9F favored mass extinction event around 75 and 65 mya (*P* = 0.93, 0.96) the survival probability of 1.49% and 1.46% respectively. Whereas, Y1-1L3F and BD1-1L3F both MCC trees possess lowest AIC values (1156.106, 1133.945) among other time-trees and supported mass extinction at ~50 mya (*P* = 0.93, 0.92) with the survival probability of 1.73% and 1.71% respectively. Specifically, it appears that MCC tree, BD1-1L9F supported upshift and mass extinction around 65 mya is closely situated around the K-Pg mass extinction[97]. Another interesting thing to note is that two MCC trees, Y1-1L3F and BD1-1L3F, supported upshift and mass extinction approximately 50 mya, which corresponds to the end of the Paleocene–Eocene Thermal Maximum (PETM)[98].

6.3.7 Trait-Dependent Diversification:

In the BiSSE analysis of eight distinct MCC trees, two models were shown to be the best fitted (Table 6.9). Three MCC trees (BD2L3F, Y2L3F, Y2L9F) fitted best in which speciation (λ) and transition (q) rates were variable and extinction (μ) remained equal between the species distributed. The other five MCC trees (BD2L9F, BD1-1L3F, BD1-1L9F, Y1-1L3F, Y1-1L9F) that best fit the data were those with variable λ and μ and without any transition from Brachycera to Nematocera (q10~0).

On the other hand, the second-best model fitted for first set of three MCC trees in which λ and μ were variable and there was no transition from Brachycera to Nematocera (q10~0) and for second set of five trees λ and q were variable and μ0 ~ μ1. According to second best model of first set of three MCC trees and best model of second set of five MCC trees the transition rate from Nematocera to Brachycera was detected but could not be detected in the opposite direction. Under the best BiSSE models, Bayesian analysis by MCMC credibility intervals

Table 6.9: Results of the trait-dependent diversification (BiSSE) models on the binary traits. The model with all speciation parameters free is supported by the lowest AICc and and ΔAICc. Adding more free parameters of diversification did not improve significantly the likelihood.

MCC Trees	Rank of Best model	Best Model	Df	lnLik	AICc	ΔAICc	w	λ0	λ1	μ0	μ1	q01	q10
BD2L3F	1	Equal_extinction	5	-576.45	1162.905	0	3.32E-01	0.010285	0.018932	1.65E-05	1.65E-05	2.78E-04	1.12E-07
	2	No_transition1_0	5	-576.45	1162.907	0.00158	3.31E-01	0.010289	0.018898	6.44E-05	2.75E-05	2.87E-04	0
BD2L9F	2	Equal_extinction	5	-583.71	1177.424	0.003494	3.38E-01	0.008094	0.019267	2.13E-06	2.13E-06	2.04E-04	2.48E-07
	1	No_transition1_0	5	-583.71	1177.421	0	3.38E-01	0.008077	0.019312	4.84E-06	2.54E-06	2.17E-04	0
BD1-1L3F	2	Equal_extinction	5	-565.4	1140.8	0.260077	0.306885044	0.012962	0.025081	5.62E-03	5.62E-03	2.43E-04	7.78E-07
	1	No_transition1_0	5	-565.27	1140.54	0	0.349502845	0.011853	0.027049	3.67E-03	9.17E-03	3.13E-04	0
BD1-1L9F	2	Equal_extinction	5	-582.54	1175.071	0.312061	0.305712753	0.008869	0.021163	2.44E-03	2.44E-03	1.67E-04	8.98E-07
	1	No_transition1_0	5	-582.38	1174.759	0	0.357335901	0.008219	0.023147	1.24E-03	6.19E-03	2.32E-04	0
Y2L3F	1	Equal_extinction	5	-585.59	1181.182	0	3.34E-01	0.009688	0.017107	1.12E-06	1.12E-06	2.70E-04	5.02E-09
	2	No_transition1_0	5	-585.59	1181.187	0.004604	3.33E-01	0.009692	0.017116	6.35E-07	1.43E-05	2.72E-04	0
Y2L9F	1	Equal_extinction	5	-592.07	1194.143	0	3.41E-01	0.008099	0.017238	4.25E-08	4.25E-08	2.11E-04	6.54E-08
	2	No_transition1_0	5	-592.08	1194.15	0.006992	3.40E-01	0.008024	0.017153	5.02E-05	2.35E-06	2.11E-04	0
Y1-1L3F	2	Equal_extinction	5	-576.28	1162.57	0.055738	3.25E-01	0.010267	0.020798	1.83E-03	1.83E-03	2.26E-04	1.47E-07
	1	No_transition1_0	5	-576.26	1162.514	0	3.35E-01	0.009919	0.021442	1.16E-03	3.12E-03	2.72E-04	0
Y1-1L9F	2	Equal_extinction	5	-592.21	1194.412	0.015918	3.36E-01	0.007295	0.017987	1.34E-05	1.34E-05	1.94E-04	1.01E-06
	1	No_transition1_0	5	-592.2	1194.396	0	3.38E-01	0.007325	0.018335	3.43E-05	7.48E-04	2.01E-04	0

Abbreviations are denoted as follows: NP, number of parameters; logL, log-likelihood; AICc, corrected Akaike Information Criterion; ΔAICc, the difference in AICc between the model with the lowest AICc and the others. Parameter estimates are denoted as follows: λ, speciation rate; μ, extinction rate; q, transition rate.

Table 6.10: Results of the trait-dependent diversification (MuSSE) models on the multiple traits. The model with all speciation parameters free is supported by the lowest AICc and and ΔAICc. Adding more free parameters of diversification did not improve significantly the likelihood.

MCC Trees	Best Models	Df	lnLik	AICc	ΔAICc	w	λ1	λ2	λ3	μ1	μ2	μ3	q12	q13	q21	q23	q31	q32
BD2L3F	free.lambda	5	-575.45	1160.9	2.10916	0.224611	0.032282	2.56E-02	2.08E-02	7.48E-03	7.48E-03	7.48E-03	2.85E-04	0	2.85E-04	2.85E-04	0	0.000285
	free.mu	5	-574.39	1158.8	0	0.644806	0.027654	0.02765	0.027654	2.99E-07	9.31E-03	1.70E-02	2.55E-04	0	2.55E-04	2.55E-04	0	0.000255
BD2L9F	free.lambda	5	-578.68	1167.4	0.99768	3.20E-01	0.033164	2.39E-02	1.64E-02	6.03E-03	6.03E-03	6.03E-03	2.21E-04	0	2.21E-04	2.21E-04	0	2.21E-04
	free.mu	5	-578.18	1166.4	0	5.27E-01	0.028269	0.02827	0.028269	7.73E-08	1.03E-02	2.09E-02	1.84E-04	0	1.84E-04	1.84E-04	0	1.84E-04
BD1-1L3F	free.lambda	5	-564.94	1139.9	0	0.457161	0.047014	0.03836	3.08E-02	2.20E-02	2.20E-02	2.20E-02	0.000341	0	0.000341	0.000341	0	0.000341
	free.mu	5	-564.95	1139.9	0.03508	0.449214	0.038188	0.03819	0.038188	0.009676	2.14E-02	3.05E-02	0.000302	0	0.000302	0.000302	0	0.000302
BD1-1L9F	free.lambda	5	-577.97	1165.9	0	4.97E-01	0.039288	2.86E-02	2.05E-02	1.28E-02	1.28E-02	1.28E-02	2.42E-04	0	2.42E-04	2.42E-04	0	2.42E-04
	free.mu	5	-578.26	1166.5	0.59048	3.70E-01	0.030574	0.03057	0.030574	2.26E-06	1.45E-02	2.46E-02	1.98E-04	0	1.98E-04	1.98E-04	0	1.98E-04
Y2L3F	free.lambda	5	-584.99	1180	1.79812	2.49E-01	0.027277	2.27E-02	1.79E-02	4.19E-03	4.19E-03	4.19E-03	2.78E-04	0	2.78E-04	2.78E-04	0	2.78E-04
	free.mu	5	-584.09	1178.2	0	6.13E-01	0.024671	0.02467	0.024671	5.03E-07	6.02E-03	1.37E-02	2.53E-04	0	2.53E-04	2.53E-04	0	2.53E-04
Y2L9F	free.lambda	5	-587.84	1185.7	0.82833	3.46E-01	0.027766	2.13E-02	1.43E-02	2.94E-03	2.94E-03	2.94E-03	2.25E-04	0	2.25E-04	2.25E-04	0	2.25E-04
	free.mu	5	-587.43	1184.9	0	5.23E-01	0.024858	0.02486	0.024858	7.69E-06	6.27E-03	1.66E-02	1.98E-04	0	1.98E-04	1.98E-04	0	1.98E-04
Y1-1L3F	free.lambda	5	-576.39	1162.8	0.44813	4.01E-01	0.03878	0.03214	2.49E-02	1.51E-02	1.51E-02	1.51E-02	0.000337	0	0.000337	0.000337	0	0.000337
	free.mu	5	-576.17	1162.3	0	5.02E-01	0.032233	0.03223	0.032233	0.005467	1.50E-02	2.42E-02	0.000299	0	0.000299	0.000299	0	0.000299
Y1-1L9F	free.lambda	5	-588.66	1187.3	0	4.49E-01	0.032503	2.45E-02	1.68E-02	8.17E-03	8.17E-03	8.17E-03	2.28E-04	0	2.28E-04	2.28E-04	0	2.28E-04
	free.mu	5	-588.71	1187.4	0.09361	4.29E-01	0.026184	0.02618	0.026184	3.44E-06	9.35E-03	1.96E-02	1.88E-04	0	1.88E-04	1.88E-04	0	1.88E-04

Abbreviations are denoted as follows: NP, number of parameters; logL, log-likelihood; AICc, corrected Akaike Information Criterion; ΔAICc, the difference in AICc between the model with the lowest AICc and the others. Parameter estimates are denoted as follows: λ, speciation rate; μ, extinction rate; q, transition rate.

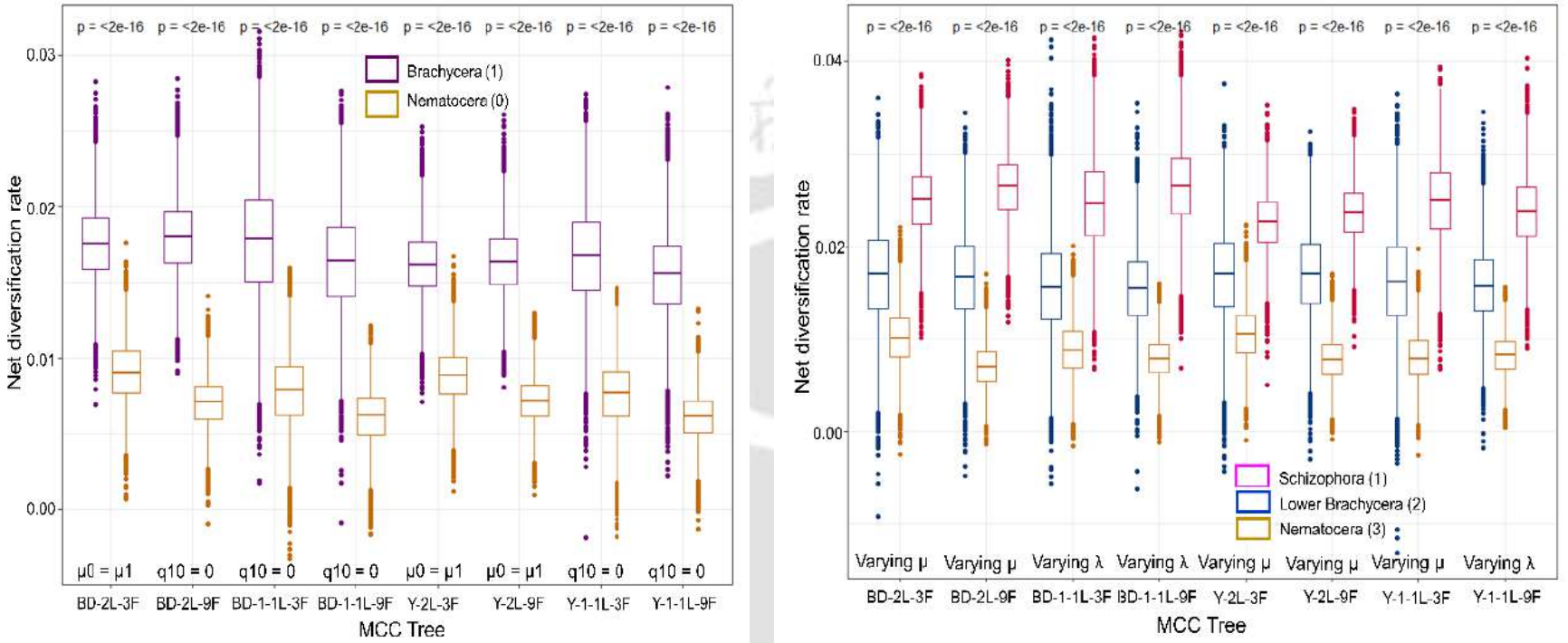


Figure 6.17: Box-plot of posterior data among eight MCC trees derived from Bayesian analysis of best models selected by two state (Brachycera and Nematocera) BiSSE (left) and three state (Schizophora, Lower Brachycera and Nematocera) MuSSE (right).

showed net diversification rates of Brachycera were significantly higher than Nematocera regardless of the different MCC trees (Fig. 6.17 (left)).

The MuSSE analysis on different lineage trait of Diptera tabulated in Table 6.10. It suggests that five MCC trees (BD2L3F, BD2L9F, Y2L3F, Y2L9F, Y1-1L3F) were fitted best where speciation (λ) and transition (q) rates estimated as equal among different traits whereas, the extinction (μ) rates were variable. The rest of three MCC trees (BD1-1L3F, BD1-1L9F, Y1-1L9F) were fitted best in which speciation (λ) rates were varied among lineage trait and extinction (μ) and transition (q) rates estimated as equal. Under the best MuSSE models, Bayesian analysis by MCMC credibility intervals exhibited net diversification rates of Schizophora were significantly higher than Lower Brachycera and net diversification rates of Nematocera significantly lower than other two lineage traits regardless of the different MCC trees (Fig. 6.17(right)).

6.3.8 Environment-Dependent Diversification:

We investigated alternative theories that could explain the mechanisms causing the observed macroevolutionary dynamics using a paleoenvironment-dependent diversification model (PDDM) applied to the Diptera phylogeny. The best fitted ML-based outcome of PDDM implemented in RPANDA is shown in Table 6.11. Five models are compared for eight MCC trees, and each was previously selected in a reciprocal series of time-dependent models, of temperature-dependent models, of atmospheric oxygen-dependent models, of sea level-dependent models, and of angiosperm-dependent models. The analysis of time-dependent models revealed that a constant birth-death model [BcstDcst] was the best fit all MCC trees, suggesting that both the rate of speciation and the rate of extinction remain constant (Table 6.11). For temperature-dependent models, the best fit for all MCC trees was a model with varying extinction rates, where extinction rate varies linearly and speciation rate is unaffected by temperature fluctuation. The negative (-ve) values of β (rate of variation of the extinction)

for Y2L3F and BD2L3F suggest a negative correlation with temperature variation, implying that higher temperatures result in lower extinction and *vice versa*. Whereas positive (+ve) values of β for other six MCC trees suggested a positive correlation with temperature fluctuation, implying that higher temperatures result in higher extinction and *vice versa* (Fig. 6.18). The majority of MCC trees had high extinction rates during warmer climates, according to our results, which is consistent with previous research that suggested that past global warming was a significant source of high extinction rates [99,100].

With past atmospheric oxygen levels, only two MCC trees (Y2L3F and Y2L9F) showed constant extinction rate and speciation rate linearly dependent on the atmospheric oxygen level. The -ve value of α (rate of variation of the speciation) for Y2L3F implied a negative correlation with oxygen level variation, inferring that higher oxygen levels reduced speciation rate and *vice versa*. The Y2L9F showed +ve α, indicating a positive correlation with oxygen level fluctuations as such higher oxygen level led to an increased rate of speciation and *vice versa*. Whereas, other six MCC trees exhibit constant speciation rate and extinction rate have linear dependence on atmospheric oxygen (Fig. 6.19). Only BD2L3F exhibited a -ve β, indicating that extinction has a negative relationship with oxygen level fluctuation, implying that higher oxygen levels decreased extinction rate and *vice versa*. The remaining five MCC trees have +ve β, which means a positive correlation of extinction rate with oxygen level variation such that elevated oxygen level increased the extinction rate, and *vice versa*. This analysis suggests that majority of the MCC trees supported the decrease of extinction rate during decrease of oxygen level such as during Jurassic period and early Eocene (Fig. 6.19). This finding backs with previous studies that found environmental constraints generated by historical fluctuations in atmospheric oxygen concentrations facilitated in species extinction by selectively eliminating species with high rates of energy utilisation[101].

The models built to assess the impact of past sea-level fluctuation for Y2L9F and BD1-1L3F showed constant speciation rate and linearly variable extinction rate with -ve β for Y2L9F

Table 6.11: Time and paleo-environment dependent best models for Diptera diversification with the phylogeny-based diversification models

MCC Tree	Dependent on	Models	NP	logL	AICc	Lambda	Alpha	Mu	Beta
Y2L3F	Time	BcstDcst	2	-583.24	**1170.58**	16.6379	NA	16.6232	NA
	Temperature	BcstDTemp.var_LIN	3	-583.19	1172.61	15.7715	NA	15.7533	-0.000194704
	Oxygen	BOxy.varDcst_LIN	3	-583.15	1172.53	15.6137	-0.000749347	15.5801	NA
	Sea level	BSea.varDcst_LIN	3	-582.49	1171.2	13.0599	-0.000124025	13.0315	NA
	Angiosperm	BcstDAngio.var_EXPO	3	-583.19	1172.61	17.2698	NA	17.2539	0.000158399
Y1-1L3F	Time	BcstDcst	2	-576.07	**1156.25**	24.7588	NA	24.7481	NA
	Temperature	BcstDTemp.var_LIN	3	-576.03	1158.29	25.7105	NA	25.7027	0.000154533
	Oxygen	BcstDOxy.var_LIN	3	-576.06	1158.35	25.1442	NA	25.1386	0.000203708
	Sea level	BSea.varDcst_LIN	3	-575.85	1157.93	22.3093	-6.31E-05	22.2918	NA
	Angiosperm	BAngio.varDAngio.var_LIN	4	-573.99	1156.35	10.6371	16.14608014	10.6172	-16.15562218
Y2L9F	Time	BcstDcst	2	-593.64	1191.38	17.2056	NA	17.1939	NA
	Temperature	BcstDTemp.var_LIN	3	-593.58	1193.39	18.1614	NA	18.1531	0.000181676
	Oxygen	BOxy.varDcst_LIN	3	-593.5	1193.21	18.5491	0.000818809	18.5582	NA
	Sea level	BcstDSea.var_LIN	3	-589.58	**1185.38**	10.3725	NA	10.3319	-0.000267874
	Angiosperm	BAngio.varDAngio.var_LIN	4	-592.81	1193.99	1.74345	16.99402953	1.72192	-16.99842027
Y1-1L9F	Time	BcstDcst	2	-595.57	1195.25	21.1773	NA	21.1686	NA
	Temperature	BcstDTemp.var_LIN	3	-595.42	1197.06	22.8169	NA	22.8131	0.000241402
	Oxygen	BcstDOxy.var_LIN	3	-595.43	1197.08	22.6382	NA	22.6476	0.000705329
	Sea level	BSea.varDcst_LIN	3	-593.18	**1192.58**	15.1843	-0.000195875	15.1548	NA
	Angiosperm	BAngio.varDAngio.var_LIN	4	-592.81	1193.99	1.74345	16.99402953	1.72192	-16.99842027
BD2L3F	Time	BcstDcst	2	-574.87	**1153.85**	20.0285	NA	20.0143	NA
	Temperature	BcstDTemp.var_LIN	3	-574.86	1155.95	19.6433	NA	19.6277	-7.75E-05
	Oxygen	BcstDOxy.var_LIN	3	-574.86	1155.94	19.6325	NA	19.6117	-0.000259331
	Sea level	BSea.varDcst_LIN	3	-574.18	1154.58	16.1086	-0.000120422	16.0814	NA
	Angiosperm	BcstDAngio.var_EXPO	3	-574.85	1155.92	20.4946	NA	20.4794	9.83E-05
BD1-1L3F	Time	BcstDcst	2	-565.11	**1134.33**	30.4648	NA	30.4548	NA
	Temperature	BcstDTemp.var_LIN	3	-565.02	1136.27	32.1632	NA	32.1578	0.000243749
	Oxygen	BcstDOxy.var_LIN	3	-565.06	1136.33	31.6681	NA	31.6723	0.000557926
	Sea level	BcstDSea.var_LIN	3	-565.11	1136.45	30.6098	NA	30.6001	2.74E-06
	Angiosperm	BcstDAngio.var_EXPO	3	-564.15	1134.52	34.6273	NA	34.6119	0.000372722
BD2L9F	Time	BcstDcst	2	-586.63	1177.36	21.3661	NA	21.3557	NA
	Temperature	BcstDTemp.var_LIN	3	-586.34	1178.91	23.8849	NA	23.8823	0.000408812
	Oxygen	BcstDOxy.var_LIN	3	-586.17	1178.57	24.1572	NA	24.1832	0.001429541
	Sea level	BSea.varDcst_LIN	3	-582.22	**1170.67**	13.1256	-0.000279126	13.0855	NA
	Angiosperm	BAngio.varDAngio.var_LIN	4	-581.87	1172.11	0.13939	18.87357669	0.09758	18.89565076
BD1-1L9F	Time	BcstDcst	2	-586.18	1176.47	25.8453	NA	25.8375	NA
	Temperature	BcstDTemp.var_LIN	3	-586	1178.22	27.7711	NA	27.7684	0.000255955
	Oxygen	BcstDOxy.var_LIN	3	-585.97	1178.17	27.8575	NA	27.8725	0.000891802
	Sea level	BSea.varDcst_LIN	3	-584.75	1175.73	20.301	-0.000153607	20.2768	NA
	Angiosperm	BAngio.varDAngio.var_LIN	4	-583.53	**1175.43**	4.75715	21.67550507	4.73881	-21.6814154

suggested that extinction was negatively correlated with past sea-level fluctuations while +ve β for BD1-1L3F positively correlated to past sea-level fluctuations. The remaining six MCC trees had a constant extinction rate and a linearly variable speciation rate, with -ve α indicating that speciation was negatively associated with past sea-level variations, with higher sea levels

reduced the speciation rate and *vice versa*. The majority of MCC trees indicate that rising sea levels slowed the rate of speciation, but one MCC tree fostered extinction which is consistent with earlier research (Fig. 6.20). Topographic abnormalities on the shelf induce the formation of separate embayments or basins when sea level rises and falls, which can isolate populations of a species, enabling allopatric speciation during times of rising sea level[99,102]. According to 2015 research by Condamine, Fabien L., et al., the rise and fall of sea levels through time is an important factor determining diversification opportunities because it creates (low sea level) or erases (high sea level) regions where radiations took place[99].

We have also tested the influence of the relative level of angiosperm on the evolution of Diptera (Table 6.11). It exhibited that three MCC trees (Y2L3F, BD2L3F, BD1-1L3F) favoured a model that had a constant speciation rate and exponentially variable extinction rate with +ve β. Where, extinction was positively correlated to past angiosperm levels such that a higher level

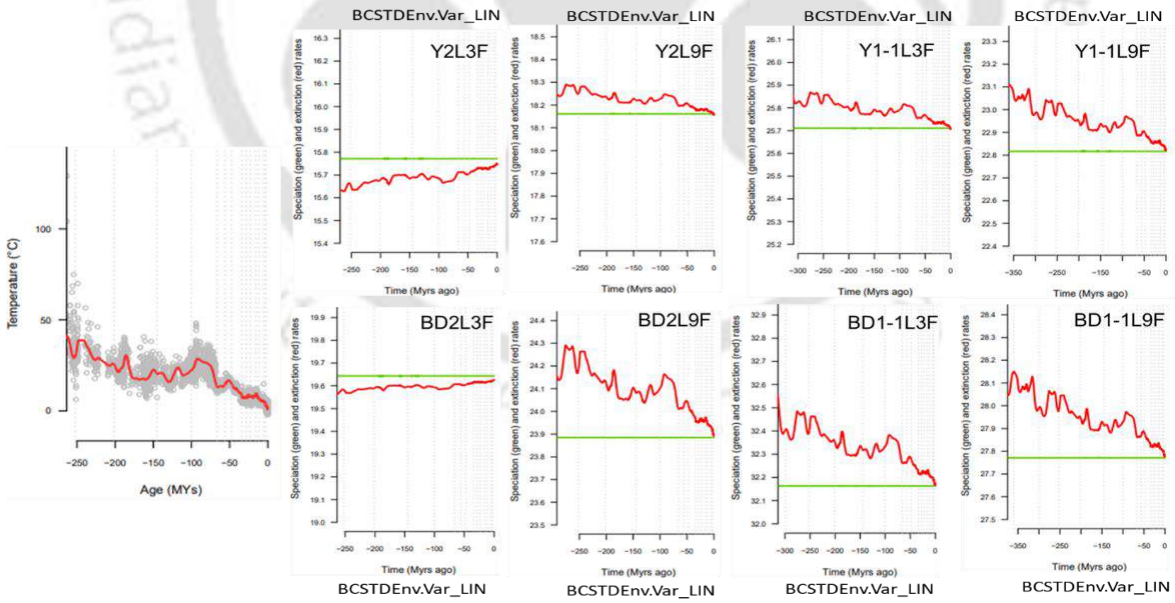


Figure 6.18: Paleo-environment-dependent diversification processes in Diptera flies. Dependence between paleo temperatures and speciation/extinction

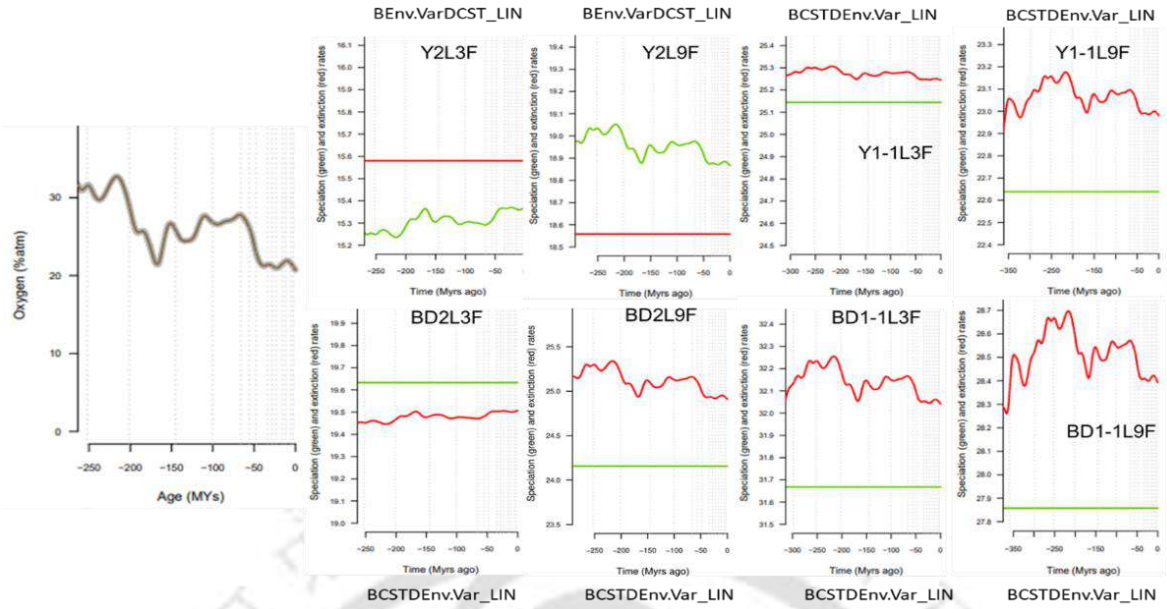


Figure 6.19: Paleo-environment-dependent diversification processes in Diptera flies. Dependence between paleo atmospheric oxygen and speciation/extinction

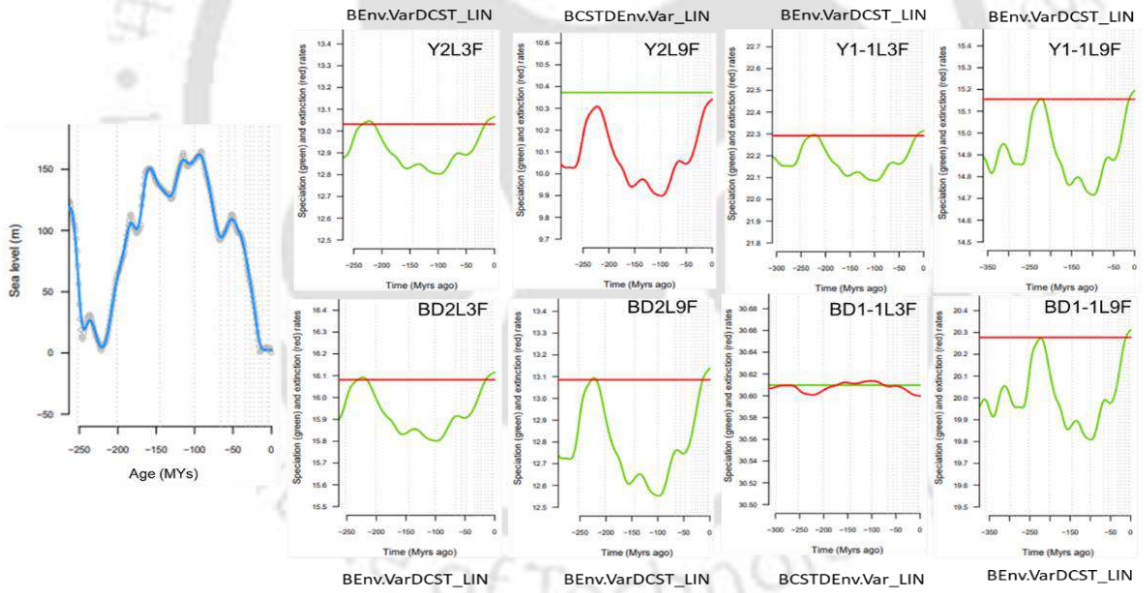


Figure 6.20: Paleo-environment-dependent diversification processes in Diptera flies. Dependence between paleo Sea level and speciation/extinction

of angiosperm caused higher extinction, and *vice versa*. On the other hand, the Y1-1L9F and BD1-1L9F exhibited positively correlated speciation (+ve α) and negatively correlated extinction (-ve β) linearly. The remaining three MCC trees are linear and positively correlated

to both speciation (+ve α) and extinction (+ve β) rate, showing that higher angiosperm levels caused both higher speciation and extinction, and *vice versa*.

Overall, according to the lowest AIC value, the time-dependent model better fits MCC trees produced with 3 fossil calibrations. While the sea-level dependent model better fits MCC trees constructed with 9 fossil calibration, except for BD1-1L9F in which the angiosperm level-dependent model better fits.

6.3.9 Diversity dynamics from fossil sampling data:

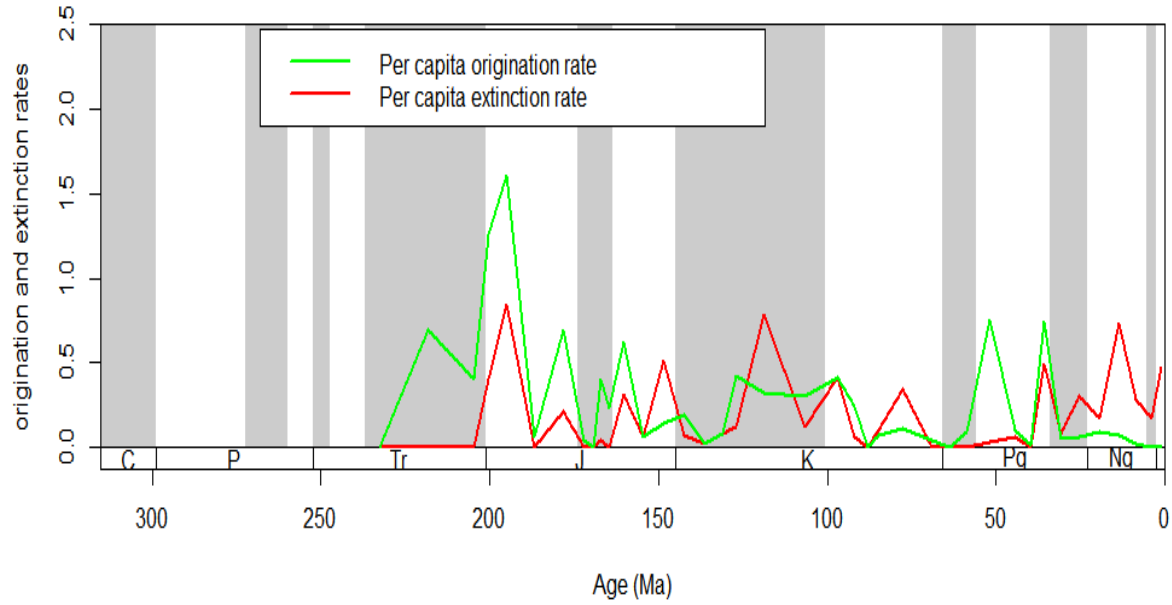


Figure 6.21: Per capita origination and extinction rate derived from fossil sampling data of Diptera. X-axis signifies million years ago.

In this analysis a R package namely, divDyn was used to estimate per capita origination and extinction rate found by Foote 1999[103].This analysis shows that origination Diptera began during early Triassic period. High origination rate in their early history and low origination rate at recent times observed from raw fossil records. A spike in origination rate and low extinction rate after Cretaceous-Paleogene (K-Pg) boundary observed in this study. Low rate of origination and relatively high extinction rate at the end of Jurassic (J), middle of Cretaceous (K) period and in the present evidenced in this analysis (Fig. 6.21).

We also estimated speciation, extinction using the Bayesian PyRate method to analyses the fossil record[36]. This analysis display that from Triassic period, a fairly constant, near-zero diversification rate is estimated no peak of origination or extinction is recovered until Paleogene (Fig. 6.22). Analyses of the fossil orders show a rise in rate of origination in the early Paleogene and Neogene. The increase of extinction rate detected in Jurassic-Cretaceous boundary, K-Pg boundary and in the present.

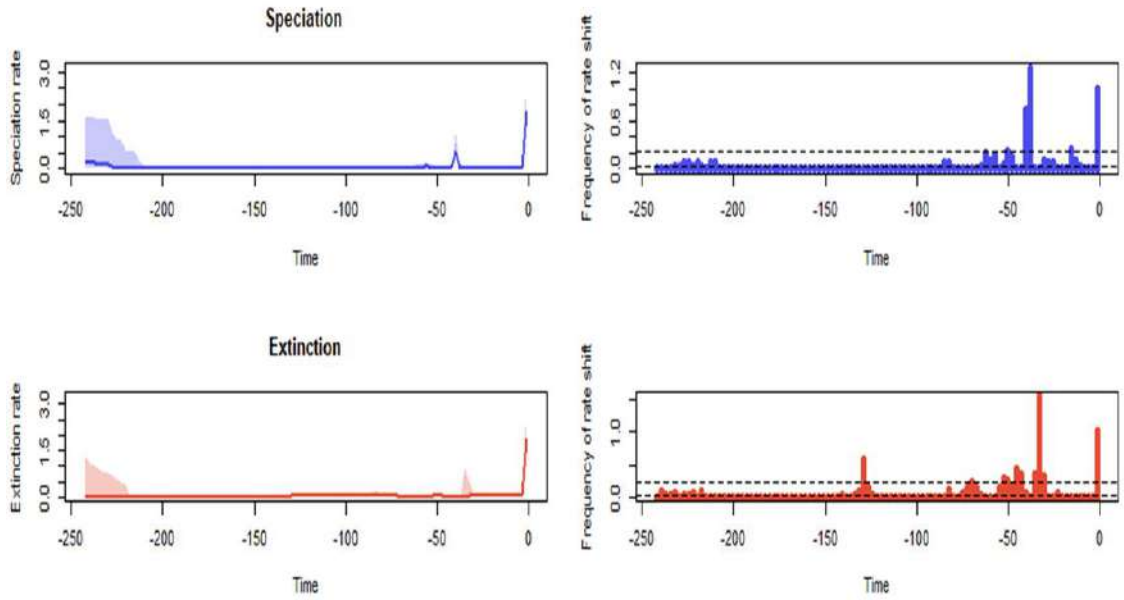


Figure 6.22: Rate-through-time plots for the marginal rates of speciation (blue) and extinction (red) through time for the Diptera obtained through BDMCMC. Solid lines show the mean rate estimates, shaded areas display the associated 95% credibility intervals.

6.4 Conclusion:

Determining the molecular divergence times and the macroevolutionary factors that influence speciation, extinction and richness is an important aspect of evolutionary biology[15,17,104,105]. Most macroevolutionary dynamics relied on molecular dating based on single branching prior processes, with little attention paid to the influence of diverse fossil calibrations or variable molecular data partitioning[15,74,105]. Furthermore, previous studies have emphasized on a single component to explain a specific diversity pattern, such as historical events that promote diversification and extinction[16,100,106], or species ecology and clade competitions as the primary

drivers of diversification[21,107]. The combined impact of multiple factors has been studied less often, and on a smaller geographic and temporal scale, or with a narrower range of approaches[108–110]. Herein, our investigation of molecular dating using alternative branching prior procedures with partitioned and non-partitioned molecular data, as well as multiple fossil calibration strategies, yields new insights about Diptera fly evolution. Furthermore, analysing and comparing the effects of putative abiotic and biotic factors driving Diptera evolution using current macroevolutionary dynamics techniques opens up a comprehensive vista of understanding.

The Cretaceous Period began 145.0 million years ago and ended 66 million years ago, making it the last and longest of the Mesozoic Era's three periods. The Cretaceous Period witnessed major changes in the placement of Earth's landmasses, which was to be anticipated given the period's length. South America separated from Africa in the Late Jurassic to Early Cretaceous, with India, Australia, and Antarctica on the other side. Also, It should be noted that although angiosperm appeared in the Jurassic, vociferous radiation happened throughout the Cretaceous [54,111]. In this context, our macroevolutionary result from BAMM suggest that acceleration of both speciation and extinction of Diptera commenced during early Cretaceous after the emergence of lineages with high diversification rate namely Schizophora in Brachycera and Culicidae in Nematocera. The BDL (birth–death likelihood) analysis revealed that 9 fossils calibration mainly supported their first-rate shift in the Cretaceous period. The density-dependent rate shift analysis suggest that the transition took place largely throughout the Cretaceous period, from early to late epochs, and it was found that carrying capacity improved as a result of it. The sea level was higher than it had ever been for much of the Cretaceous period, and this was a major influence on the paleogeography of the time. When sea levels risen during the Late Cretaceous, marine waters submerged the continents, resulting in relatively shallow epicontinental seas. In this scenario, our findings show that the rate of

speciation in Diptera lineages declined as sea levels increased. The temperature during the Cretaceous Period was significantly higher than today's, maybe the highest on a global scale at any period throughout the Phanerozoic Eon. According to our findings, the Diptera experienced high extinction rates in warmer temperatures, which is consistent with earlier studies suggesting that historical global warming was a primary driver of high extinction extinction rates[64]. After the Cretaceous period, the Paleogene period began, which is divided into three epochs: Paleocene, Eocene, and Oligocene. Between 60 and 50 mya, during the end of the Paleocene and the beginning of the Eocene, a series of hyperthermal events occurred, causing a rise in global average temperature and changes in the Earth's carbon cycle[112,113]. In this context, BAMM's macroevolutionary diversification analysis suggests that crown groups (e.g. Calliphoridae and Sarcophagidae) began diversification around 50 mya and net-diversification rate of Diptera got pace around the K-Pg boundary. The CoMET analysis show increase in speciation rate after the Eocene hyperthermal events. As per the paleo-environmental investigations, sea-level, temperature, and atmospheric oxygen levels have all declined since 50 million years ago, resulting in more diversification and lower extinction rates.

Since the emergence of Diptera three mass extinction event happened and our analysis suggest that Diptera might have escaped earlier two mass extinction events. Our estimation of witnessing mass extinction by Diptera mainly concentrated on late Cretaceous to Eocene period. Earlier studies suggested extinction rate estimates from molecular phylogenies may be dubious[95,96]. Although, analysis with TreePar with 9 fossils calibration supported the significant rate shift and mass extinction in the Cretaceous. Early Cretaceous downshift is seen in Y2L9F and BD2L9F, while late Cretaceous upshift observed in BD1-1L9F at 65 mya. The Cretaceous mass extinction is supported by Y1-1L9F and BD1-1L9F at 75 and 65 mya, respectively. However, according to CoMET analysis, only Y1-1L3F exhibit an extra surge in extinction rate during 75-65 mya. The detect mass extinction in late Cretaceous can be

attributed to the K–Pg extinction event. In addition, our analysis found mass extinctions in the Eocene period, such as Y1-1L9F showing a significant rate upshift in the Eocene at 40 mya in TreePar analysis. Around 50 mya, Y1-1L3F and BD1-1L3F supported significant rate upshift as well as mass extinction. Furthermore, the CoMET study suggests a little rise in the extinction rate approximately 50 MY. This early Eocene mass extinction event might correspond to the Eocene hyperthermal events[112]. In contrast, fossil-only analysis with PyRate identified a rise in extinction rate during the Jurassic-Cretaceous boundary, the K-Pg boundary, and in the present. While DivDyn analysis indicates high extinction during the end of the Triassic, end of the Jurassic (J), middle of the Cretaceous (K), and present. Therefore, the fossil-based approach DivDyn was able to detect the Triassic-Jurassic (T-J) mass extinction event as well as three other extinction events. Both the PyRate and DivDyn methods revealed a high level of extinction towards the present, implying that population levels may be reduced due to competition and dispersion restrictions, increasing the risk of extinction[114].

Our findings also imply that Diptera diversity is partly diversity driven, since diversity accumulated through time but began to shrink when the extinction rate increased and speciation rate decrease. The slowdown in lineage diversification rates through time is a well-known phenomenon in diversification[24,91]. As the niche gets saturated and the competition for ecological space grows, the rate of diversification decreases. The rate through time plot in BAMM analysis was able to capture this tendency, showing that as the rate of speciation increased, so did the rate of extinction. TreePar analysis supports this, revealing a similar situation in which the estimated turnover is quite high (> 0.9) while being less than 1, indicating that extinction never outpaced speciation rate.

Biological interactions between distantly related lineages are likely to have played an important role in clade diversification during geologic history[107,115]. Significant disparities in speciation rates were found in lineages with diverse traits, according to our findings. Brachycera has much

greater rates of speciation than Nematocera, according to many models (BAMM, BiSSE, MuSSE). When Brachycera is divided into two categories, Schizophora and Lower Brachycera, the Schizophora has a far faster rate of diversification than the other two. Interestingly, lowest diversification rate associated with Nematocera without Culicidae family (Background Diptera in RTT plot). The extant Nematocera are mostly aquatic lineages with long antennae, and they first appeared as fly species during the Carboniferous to Permian periods, before diversifying throughout the Triassic period[3]. Lower Brachycera are primarily terrestrial lineages, many of which are flower visitors with long proboscides for nectar feeding[3]. They evolved in the late Triassic and radiated in the early Jurassic, shortly after angiosperms appeared[54]. The Schizophora initially appeared in the late Jurassic to early Cretaceous, but the most explosive radiation occurred in the late Cretaceous to Eocene, according to our findings[3,116]. With the development of new key innovations such as the ptilinal sac (improved escape mechanism for the fly from its puparium), highly modified parasitism, and so on, they are able to invade previously underutilized niches[2,3]. This might be the potential reason of Schizophora for higher diversification rate than others as it is related with the survivability and feeding habit[53]. According to BiSSE analysis, transition from Nematocera to Brachycera was conceivable but not allowed in the reverse direction.

In adaptive radiations, a rapid burst of diversification is expected, followed by a reduction in speciation rates when niches are filled; this is due to a diversity-dependence effect[23]. A transition to a new ecological resource (for example, a new host) may represent a turning point in terms of adaptation, allowing for more speciation at the beginning of the radiation than declines later as the number of species accumulate[117]. This research found a pattern of decreasing speciation or rising extinction rates with standing diversity of Diptera, confirming the concept as an example of adaptive radiation driven by the emergence of a new ecological resource[92].

For biological radiations to succeed, both extrinsic conditions and intrinsic traits must function in harmony or sequentially across time, and both abiotic and biotic factors promote species diversity[118]. We argue that RQ and CJ-type factors are intricately related to each other, and that they both enhance and inhibit species diversification. Several factors, ranging from past climate change to sea-level changes to the appearance of angiosperms, can be used to describe the diversification of Diptera. Our findings also suggest that clade-specific diversity-dependence, as well as ecological and biotic interactions for survivorship and feeding habit associations, can produce differential diversification capacity in closely related species living in the same rapidly changing environment. In particular, the combined influence of temperature and geological changes, together with novel biotic interactions, appears to have aided Diptera radiation. Sea level variation promoted allopatric speciation, but it also enhanced habitat heterogeneity, resulting in the emergence of new niches (ecological divergence). The invasion of such niches resulted in new host species relationships (such as angiosperms) and the evolution of unique morphological adaptive traits, such as extended proboscides for effective nectar feeding. Or Schizophora's modified parasitism allows them to invade various niches in a very short period of time, making them one of the most vigorous diversification bursts in geological history[3].

Incomplete taxon sampling, inaccurate divergence time estimates, and small phylogenies are all serious difficulties that are difficult to resolve. This can result in erroneous parameter estimations and model selections. The accuracy of our analyses and findings may be hampered by such methodological restrictions[119]. This study, however, demonstrates that several techniques must be integrated to adequately address the diversity of any lineage in relation to abiotic and biotic factors. The current diversity patterns are the result of the combined influence of numerous variables (species traits, environmental drivers, and mass extinction), rather than a single factor, and we propose a comparably multidimensional approach for studying these

patterns. As long as the basic hypothesis testing component of these inquiries is kept in mind, these studies will remain our fundamental understanding of Earth's rich and deep history of remarkable biological variety. We believe that our strategy may open up new avenues for future research into other model groups with broader phylogenies.

6.5 References:

1. Stork, N. E., McBroom, J., Gely, C. & Hamilton, A. J. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7519–7523 (2015).
2. Grimaldi, D. A. & Engel, M. S. *Evolution of the insects*. (Cambridge University Press, 2005).
3. Wiegmann, B. M. *et al.* Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* **108**, 5690–5695 (2011).
4. Montagna, M. *et al.* Recalibration of the insect evolutionary time scale using Monte San Giorgio fossils suggests survival of key lineages through the End-Permian Extinction. *Proc. R. Soc. B* **286**, (2019).
5. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science (80-.).* **346**, 763–767 (2014).
6. Zhao, Z. *et al.* The Mitochondrial Genome of Elodia flavipalpis Aldrich (Diptera: Tachinidae) and the Evolutionary Timescale of Tachinid Flies. *PLoS One* **8**, 61814 (2013).
7. Simpson, G. G. *The Major Features of Evolution*. *The Major Features of Evolution* (Columbia University Press, 1953). doi:10.7312/SIMP93764/HTML.
8. Schluter, D. *The ecology of adaptive radiation*. (Oxford University Press, 2000).
9. Erwin, D. H. Lessons from the past: Biotic recoveries from mass extinctions. *Proc. Natl. Acad. Sci.* **98**, 5399–5403 (2001).
10. Benton, M. J. The Red Queen and the Court Jester: Species Diversity and the Role of Biotic and Abiotic Factors Through Time. *Science (80-.).* **323**, 728–732 (2009).
11. Moen, Daniel, and H. M. Why does diversification slow down ? *Trends Ecol. Evol.* **29**, 190–197 (2014).
12. Condamine, F. L., Rolland, J. & Morlon, H. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecol. Lett.* **22**, 1900–1912 (2019).
13. Kong, H. *et al.* Phylogenomic and Macroevolutionary Evidence for an Explosive Radiation of a Plant Genus in the Miocene. *Syst. Biol.* (2021) doi:10.1093/SYSBIO/SYAB068.
14. Van Valen, L. A new evolutionary law. *Evol. Theor* **1**, 1–30 (1973).
15. Condamine, F. L., Rolland, J., Höhna, S., Sperling, F. A. H. & Sanmartín, I. Testing the role of the Red Queen and Court Jester as drivers of the macroevolution of Apollo butterflies. *Syst. Biol.* **67**, 940–964 (2018).
16. Barnosky, A. D. Distinguishing the effects of the Red queen and Court Jester on Miocene mammal evolution in the northern Rocky Mountains. *J. Vertebr. Paleontol.* **21**, 172–185 (2001).
17. Morlon, H. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**, 508–525 (2014).
18. Stadler, T. Recovering speciation and extinction dynamics based on phylogenies. *J. Evol. Biol.* **26**, 1203–1219 (2013).
19. Höhna, S. The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. *J. Theor. Biol.* **380**, 321–331 (2015).
20. Morlon, H., Parsons, T. L. & Plotkin, J. B. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16327–16332 (2011).

21. Rabosky, D. L. *et al.* Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**, 1–8 (2013).

22. Alfaro, M. E. *et al.* Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl. Acad. Sci.* **106**, 13410–13414 (2009).

23. Etienne, R. S. *et al.* Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc. R. Soc. B Biol. Sci.* **279**, 1300–1309 (2012).

24. Rabosky, D. L. & Lovette, I. J. Explosive Evolutionary Radiations: Decreasing Speciation Or Increasing Extinction Through Time? *Evolution (N. Y).* **62**, 1866–1875 (2008).

25. Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a Binary Character's Effect on Speciation and Extinction. *Syst. Biol.* **56**, 701–710 (2007).

26. Ng, J. & Smith, S. D. How traits shape trees: new approaches for detecting character state-dependent lineage diversification. *J. Evol. Biol.* **27**, 2035–2045 (2014).

27. Condamine, F. L., Rolland, J. & Morlon, H. Macroevolutionary perspectives to environmental change. *Ecol. Lett.* **16**, 72–85 (2013).

28. Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6187–6192 (2011).

29. May, M. R., Höhna, S. & Moore, B. R. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods Ecol. Evol.* **7**, 947–959 (2016).

30. Jetz, W. & Pyron, R. A. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nat. Ecol. Evol.* **2**, 850–858 (2018).

31. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nat. 2012 4917424* **491**, 444–448 (2012).

32. Alfaro, M. E. *et al.* Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. *Nat. Ecol. Evol. 2018 24* **2**, 688–696 (2018).

33. Varga, T. *et al.* Megaphylogeny resolves global patterns of mushroom evolution. *Nat. Ecol. Evol.* **3**, 668–678 (2019).

34. Louca, S. & Pennell, M. W. Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505 (2020).

35. Slater, G. J., Price, S. A., Santini, F. & Alfaro, M. E. Diversity versus disparity and the radiation of modern cetaceans. *Proc. R. Soc. B Biol. Sci.* **277**, 3097–3104 (2010).

36. Silvestro, D., Salamin, N. & Schnitzler, J. PyRate: a new program to estimate speciation and extinction rates from incomplete fossil data. *Methods Ecol. Evol.* **5**, 1126–1131 (2014).

37. Kocsis, Á. T., Reddin, C. J., Alroy, J. & Kiessling, W. The <scp>r</scp> package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods Ecol. Evol.* **10**, 735–743 (2019).

38. Magall´on, S. & Magall´on, M. Using Fossils to Break Long Branches in Molecular Dating: A Comparison of Relaxed Clocks Applied to the Origin of Angiosperms. *Syst. Biol.* **59**, 384–399 (2010).

39. Foster, C. S. P. *et al.* Evaluating the Impact of Genomic Data and Priors on Bayesian Estimates of the Angiosperm Evolutionary Timescale. *Syst. Biol.* **66**, 338–351 (2017).

40. Leary, M. a O. *et al.* The Placental Mammal Ancestor and the Post–K-Pg Radiation of Placentals. *Science (80-.).* **339**, 662–667 (2013).

41. Meredith, R. W. *et al.* Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science (80-.).* **334**, 521–524 (2011).

42. Norman, J. E. & Ashley, M. V. Phylogenetics of Perissodactyla and Tests of the Molecular Clock. *J. Mol. Evol. 2000 501* **50**, 11–21 (2000).

43. Zakharov, E. V, Caterino, M. S. & Sperling, F. A. H. Molecular Phylogeny, Historical Biogeography, and Divergence Time Estimates for Swallowtail Butterflies of the Genus Papilio (Lepidoptera: Papilionidae). *Syst. Biol* **53**, 193–215 (2004).

44. Pulquério, M. J. F. & Nichols, R. A. Dates from the molecular clock: how wrong can we be? *Trends*

Ecol. Evol. **22**, 180–184 (2007).

45. Forest, F. Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. *Ann. Bot.* **104**, 789–794 (2009).

46. Inoue, J., Donoghue, P. C. J. & Yang, Z. The impact of the representation of fossil calibrations on bayesian estimation of species divergence times. *Syst. Biol.* **59**, 74–89 (2010).

47. Dornburg, A., Beaulieu, J. M., Oliver, J. C. & Near, T. J. Integrating Fossil Preservation Biases in the Selection of Calibrations for Molecular Divergence Time Estimation. *Syst. Biol.* **60**, 519–527 (2011).

48. Ho, S. Y. W. & Phillips, M. J. Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. *Syst. Biol.* **58**, 367–380 (2009).

49. Dornburg, A., Brandley, M. C., McGowen, M. R. & Near, T. J. Relaxed Clocks and Inferences of Heterogeneous Patterns of Nucleotide Substitution and Divergence Time Estimates across Whales and Dolphins (Mammalia: Cetacea). *Mol. Biol. Evol.* **29**, 721–736 (2012).

50. Soltis, P. S., Soltis, D. E., Savolainen, V., Crane, P. R. & Barraclough, T. G. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. *Proc. Natl. Acad. Sci.* **99**, 4430–4435 (2002).

51. Donoghue, P. C. J. & Benton, M. J. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. *Trends Ecol. Evol.* **22**, 424–431 (2007).

52. Ho, S. Y. W. & Larson, G. Molecular clocks: when timesare a-changin'. *Trends Genet.* **22**, 79–83 (2006).

53. Bertone, Matthew A., and B. M. W. *True fies (Diptera). The timetree of life* (2009).

54. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl. Acad. Sci.* **107**, 5897–5902 (2010).

55. Blagoderov, V., Grimaldi, D. A. & Fraser, N. C. How time flies for flies: Diverse diptera from the triassic of Virginia and early radiation of the order. *Am. Museum Novit.* **24112**, 1–39 (2007).

56. Cho, S. *et al.* Can Deliberately Incomplete Gene Sample Augmentation Improve a Phylogeny Estimate for the Advanced Moths and Butterflies (Hexapoda: Lepidoptera)? *Syst. Biol.* **60**, 782–796 (2011).

57. C. C. Labandeira. The Evolutionary Biology of Flies. in *The Evolutionary Biology of Flies* (eds. D. K. Yeates & B. M. Wiegmann) 217 (Columbia University Press, 2005).

58. Ho, S. Y. An examination of phylogenetic models of substitution rate variation among lineages. *Biol. Lett.* **5**, 421–424 (2009).

59. Breinholt, J. W. & Kawahara, A. Y. Phylotranscriptomics: Saturated Third Codon Positions Radically Influence the Estimation of Trees Based on Next-Gen Data. *Genome Biol. Evol.* **5**, 2082–2092 (2013).

60. Phillips, M. J., Delsuc, F. & Penny, D. Genome-Scale Phylogeny and the Detection of Systematic Biases. *Mol. Biol. Evol.* **21**, 1455–1458 (2004).

61. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

62. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–7 (2014).

63. Krzemiński, W., Krzemińska, E. & Papier, F. Grauvogelia arzvilleriana sp.n. - the oldest Diptera species [Lower-Middle Triassic of France]. *Acta Zool. Cracoviensia* **37**, (1994).

64. Carolina, A. *et al.* Large-scale mitogenomics enables insights into Schizophora (Diptera) radiation and population diversity. *Sci. Rep.* **6**, 1–13 (2016).

65. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).

66. dos Reis, M. & Yang, Z. Bayesian molecular clock dating using genome-scale datasets. in *Methods in Molecular Biology* vol. 1910 309–330 (Humana Press Inc., 2019).

67. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

68. Pybus, O. G. & Harvey, P. H. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc. R. Soc. London. Ser. B Biol. Sci.* **267**, 2267–2272 (2000).

69. Revell, L. J. phytools : an R package for phylogenetic comparative biology (and other things). 217–223 (2012) doi:10.1111/j.2041-210X.2011.00169.x.

70. Rabosky, D. L. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. *PLoS One* **9**, e89543 (2014).

71. Rabosky, D. L. *et al.* BAMMtools : an R package for the analysis of evolutionary dynamics on phylogenetic trees. 701–707 (2014) doi:10.1111/2041-210X.12199.

72. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R news* **6**, 7–11 (2006).

73. Culshaw, V., Stadler, T. & Sanmartín, I. Exploring the power of Bayesian birth-death skyline models to detect mass extinction events from phylogenies with only extant taxa. *Evolution (N. Y).* **73**, 1133–1150 (2019).

74. Höhna, S., May, M. R. & Moore, B. R. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics* **32**, 789–791 (2016).

75. Aduse-Poku, K. *et al.* Miocene Climate and Habitat Change Drove Diversification in Bicyclus, Africa's Largest Radiation of Satyrine Butterflies. *Syst. Biol.* **0**, 1–19 (2021).

76. Morlon, H. *et al.* RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. *Methods Ecol. Evol.* **7**, 589–597 (2016).

77. Condamine, F. L., Alexandre Antonelli, Laura P. Lagomarsino, Carina Hoorn & Lee Hsiang Liow. Teasing apart mountain uplift, climate change and biotic drivers of species diversification. in *Mountains, Climate and Biodiversity* (eds. Hoorn, C., Allison Perrigo, and & Alexandre Antonelli) 257–272 (John Wiley & Sons, 2018).

78. Etienne, R. S. & Haegeman, B. A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. *Am. Nat.* **180**, (2012).

79. Rabosky, D. L. LASER: A Maximum Likelihood Toolkit for Detecting Temporal Shifts in Diversification Rates From Molecular Phylogenies. *Evol. Bioinform. Online* **2**, 247 (2006).

80. Rabosky, D. L. Likelihood methods for detecting temporal shifts in diversification rates. *Evolution (N. Y).* **60**, 1152–1164 (2006).

81. Fitzjohn, R. G., Maddison, W. P. & Otto, S. P. Estimating Trait-Dependent Speciation and Extinction Rates from Incompletely Resolved Phylogenies. *Syst. Biol.* **58**, 595–611 (2009).

82. FitzJohn, R. G. <tt>Diversitree</tt> : comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* **3**, 1084–1092 (2012).

83. Tong, K. J., Duchêne, S., Ho, S. Y. W. & Lo, N. Comment on 'Phylogenomics resolves the timing and pattern of insect evolution'. 763 (2014) doi:10.1126/science.aaa5460.

84. Nee, S., Mooers, A. 0 & Harvey, P. H. Tempo and mode of evolution revealed from molecular phylogenies. **89**, 8322–8326 (1992).

85. Pennell, M. W., Sarver, B. A. J. & Harmon, L. J. Trees of Unusual Size: Biased Inference of Early Bursts from Large Molecular Phylogenies. *PLoS One* **7**, e43348 (2012).

86. Liow, L. H., Quental, T. B. & Marshall, C. R. When Can Decreasing Diversification Rates Be Detected with Molecular Phylogenies and the Fossil Record? *Syst. Biol.* **59**, 646–659 (2010).

87. Nee, S., May, R. M. & Harvey, P. H. The reconstructed evolutionary process. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **344**, 305–311 (1994).

88. Rabosky, D. L. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. *Ecol. Lett.* **12**, 735–743 (2009).

89. Rabosky, D. L. & Lovette, I. J. Density-dependent diversification in North American wood warblers. *Proc. R. Soc. B Biol. Sci.* **275**, 2363–2371 (2008).

90. Gavrilets, S. & Losos, J. B. Adaptive radiation: Contrasting theory with data. *Science (80-.).* **323**, 732–

737 (2009).

91. Yoder, J. B. *et al.* Ecological opportunity and the origin of adaptive radiations. *J. Evol. Biol.* **23**, 1581–1596 (2010).

92. Hunter, J. P. Key innovations and the ecology of macroevolution. *Trends Ecol. Evol.* **13**, 31–36 (1998).

93. Quental, T. B. & Marshall, C. R. Extinction during evolutionary radiations: reconciling the fossil record with molecular phylogenies. *Evolution (N. Y).* **63**, 3158–3167 (2009).

94. Rabosky, D. L. Extinction rates should not be estimated from molecular phylogenies. *Evol. Int. J. Org. Evol.* **64**, 1816–1824 (2010).

95. Renne, P. R. *et al.* Time scales of critical events around the cretaceous-paleogene boundary. *Science (80-.).* **339**, 684–687 (2013).

96. Zachos, J. C., Dickens, G. R. & Zeebe, R. E. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nat. 2008 4517176* **451**, 279–283 (2008).

97. Condamine, F. L. *et al.* Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. *Sci. Rep.* **5**, 1–11 (2015).

98. Mayhew, P. J., Jenkins, G. B. & Benton, T. G. A long-term association between global temperature and biodiversity, origination and extinction in the fossil record. *Proc. R. Soc. B Biol. Sci.* **275**, 47–53 (2008).

99. Mcalester, A. L. Animal Extinctions, Oxygen Consumption, and Atmospheric History. *J. Paleontol.* **44**, 405–409 (1970).

100. Allmon, W. D. & Smith, U. E. What, if anything, can we learn from the fossil record about speciation in marine gastropods? Biological and geological considerations. *Am. Malacol. Bull.* **29**, 247–276 (2011).

101. Foote, M. Morphological Diversity In The Evolutionary Radiation Of Paleozoic and Post-Paleozoic Crinoids. *Paleobiology* **25**, 1–115 (1999).

102. Benton, M. J. Exploring macroevolution using modern and fossil data. *Proc. R. Soc. B Biol. Sci.* **282**, (2015).

103. Dornburg, A., Townsend, J. P., Friedman, M. & Near, T. J. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. *BMC Evol. Biol.* **14**, 169 (2014).

104. Erwin, D. H. Climate as a Driver of Evolutionary Change. *Curr. Biol.* **19**, R575–R583 (2009).

105. Silvestro, D., Antonelli, A., Salamin, N. & Quental, T. B. The role of clade competition in the diversification of North American canids. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8684–8689 (2015).

106. Drummond, C. S., Eastwood, R. J., Miotto, S. T. S. & Hughes, C. E. Multiple Continental Radiations and Correlates of Diversification in Lupinus (Leguminosae): Testing for Key Innovation with Incomplete Taxon Sampling. *R. Bot. Gard. Kew, Wakehurst Place* **61**, 443–460 (2012).

107. Bouchenak-Khelladi, Y., Onstein, R. E., Xing, Y., Schwery, O. & Linder, H. P. On the complexity of triggering evolutionary radiations. *New Phytol.* **207**, 313–326 (2015).

108. Lagomarsino, L. P., Condamine, F. L., Antonelli, A., Mulch, A. & Davis, C. C. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytol.* **210**, 1430–1442 (2016).

109. Condamine, F. L., Silvestro, D., Koppelhus, E. B. & Antonelli, A. The rise of angiosperms pushed conifers to decline during global cooling. *Proc. Natl. Acad. Sci.* **117**, 28867–28875 (2020).

110. Slotnick, B. S. *et al.* Large-amplitude variations in carbon cycling and terrestrial weathering during the latest Paleocene and earliest Eocene: The record at Mead Stream, New Zealand. *J. Geol.* **120**, 487–505 (2012).

111. Abels, H. A. *et al.* Terrestrial carbon isotope excursions and biotic change during Palaeogene hyperthermals. *Nat. Geosci.* **5**, 326–329 (2012).

112. Liao, J. *et al.* Modelling plant population size and extinction thresholds from habitat loss and habitat fragmentation: Effects of neighbouring competition and dispersal strategy. *Ecol. Modell.* **268**, 9–17 (2013).

113. Voje, K. L., Holen, Ø. H., Liow, L. H. & Stenseth, N. C. The role of biotic forces in driving macroevolution: beyond the Red Queen. *Proc. R. Soc. B Biol. Sci.* **282**, (2015).

114. von Tschirnhaus, M. & Christel Hoffeins. Fossil flies in Baltic amber–insights in the diversity of Tertiary Acalyptratae (Diptera, Schizophora), with new morphological characters and a key based on 1,000 collected inclusions. *Denisia* **26**, 171–212 (2009).

115. Ehrlich, P. R. & Raven, P. H. Butterflies and plants: a study in coevolution. *Evolution (N. Y).* **18**, 586–608 (1964).

116. Donoghue, M. J. & Sanderson, M. J. Confluence, synnovation, and depauperons in plant diversification. *New Phytol.* **207**, 260–274 (2015).

117. Condamine, F. L., Clapham, M. E. & Kergoat, G. J. Global patterns of insect diversification: towards a reconciliation of fossil and molecular evidence? *Sci. Rep.* **6**, 19208 (2016).

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी • Indian Institute of Technology Guwahati •

CHAPTER 7

Summary and Future Prospects

> "Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."
>
> — **Marie Curie**

TH-2821_146106003

CHAPTER 7

Summary and Future Prospects

The entire study was deliberately designed for the exploration of the evolution scenery of Diptera flies. Throughout the thesis, the key theme is to reconstruct Dipteran phylogeny with two main molecular datasets (mitochondrial and genomic) and correlate different molecular phenomena with physical and environmental parameters. Chapter 2 reports the mitochondrial genome of *Blepharipa* sp., an endoparasite of Muga silkworm, and presents an overview of the phylogenetic relationship of the Oestroidea superfamily (n = 36) using mitochondrial genes. Chapter 3 broadens the taxon sampling (Diptera = 112) and investigates the influence of different factors that impose various challenges during the reconstruction of Dipteran phylogeny. Chapter 4 expands the dataset and uses orthologous genes to reconstruct Diptera phylogeny and to estimate divergence time. In Chapter 5, a correlation between dietary pattern and the nucleotide substitution rate of mitochondrial DNA is inferred. Finally, Chapter 6 examines the divergence trends of major Diptera clades, as well as the impact of various biotic and abiotic factors on fly evolution.. A chapter-wise summary and perspective is provided below,

7.1 Mitochondrial genome sequencing of *Blepharipa sp.*, an endoparasite of Muga silkworm, and comparative codon usage analysis with other Oestroidea flies

In this chapter, we reported and annotated first complete mitochondrial genome of *Blepharipa sp.* using Next-Generation Sequencing technique. The mitogenome of *Blepharipa sp.*

comprises typical dipteran gene organization and number (37 genes: 13 protein coding genes, 22 tRNAs, 2 rRNAs). This mitochondrial genome was compared to other species of Oestroidea superfamily as well as certain species of order Diptera in terms of numerous parameters, particularly codon usage pattern. The comparative analysis of protein coding genes leads to unravel the use of AT rich PCGs as well as AT biased 3rd codon position of *Blepharipa sp.* along with its family Tachinidae as compared to other Oestroidea flies. As a consequence, lower number of effective codons present in this family. We postulate that it might be the reason for their endoparasitic life strategy. In addition, the phylogenies of Oestroidea exhibited well-supported monophyly of Sarcophagidae and Calliphoridae family.

Future Prospect:

The mitogenome of *Blepharipa sp.* is the fourth complete mitogenome from this family Tachinidae till now sequenced, out of almost 10,000 species. There are a plenty of scope for generating a greater mitogenomic resources and studying on their molecular basis. It will be very appreciating if in future more analysis on this family and on this postulate confirm or criticize our observation.

7.2 Reconstruction of Diptera phylogeny with larger taxa

Following the previous chapter that lent a drawback of proper phylogenetic relationship of different Diptera flies. Here, reconstructed phylogeny with multiple homogeneous models and method by larger taxon set (116 taxa), provided substandard phylogenetic resolution. The profile of phylogenetic informativeness (PI) which capture the signal inferred by the tree building methods decipher that inconsistency of PI persists across the depth of the tree. This is due to presence of several kind of noise from the dataset namely, synonymous codon substitution, compositional heterogeneity among taxa, among site composition heterogeneity, heterotachy within-site rate variations, homoplasy (character shared by a set of species but not

present in their common ancestor), substitution saturation or reticulate evolution. The coalescent-based species tree also exposed gene-tree discordance with Diptera phylogenetic backbone. Further, inspection of datasets with network represents obvious ambiguity in the signal observed due to rapid and reticulate diversifications may explain conflicts among genes and the challenges for resolving evolutionary relationships. The evidences of other historical processes, such as recombination, or reticulation and their outcome invoked the idea of inferring unique features from the data without limiting our focus to a single tree.

Future Prospect:

In phylogenetic analysis the systematic error from molecular dataset is long-standing problem. Many scientists have committed decades of study to addressing various aspects of this issue. Moreover, it is evident that many phylogenetic histories will entail significant scientific effort to interpret, not necessarily as our methods are still deprived but because the histories are truly very challenging to recover. It opens a new vista for designing more vigorous tool for addressing all the aforementioned systematic issue at a time as I found during my study that a lot individual benchmark tools had been developed in past for different individual problems. In addition, it will be essential in these circumstances to maintain an open mind towards alternative way to infer the data.

7.3 Reconstruction of Dipteran phylogeny and molecular dating using larger dataset (Orthologous genes) and comparative analysis.

The learning from previous chapter was increasing the taxon set led to systematic error in phylogenetic relation. Here in this chapter, constraining the taxon set (52 Diptera, 2 Outgroups) we increased the amount of data by identifying the orthologous protein coding genes from the available genome in the public databases. A general comparison and correlation study done on the identified genes such as correlation of size of the genomes between total gene and CDS

size. Correlation among different codon usage indices of 52 species of identified orthologous genes. Then we identified 335 common orthologous genes that presence in all 54 taxa and used for phylogenetic analysis. Different concatenation and coalescent based method were applied for building phylogenetic tree. The analysis suggest that some node have incongruence in the backbone of species tree. The comparison of phylogenetic trees using multidimensional scaling of different distance-based method in treespace provides multiple clusters of similar species trees and gene trees. The molecular dating analysis also performed implementing various strategies confirmed that Diptera evolved between Late Permian to Late Triassic period.

Future Prospect:

Majority of the evolutionary research on true flies till date based on morphologic characteristics or using molecular marker such as, DNA barcode and few nuclear genes. Although, some contemporary researches have implemented mitochondrial DNA for phylogenetic investigation of Diptera, still the utilization of a large genomic size data for evolutionary analysis is extremely rare. Thus, the use of orthologous genes in evolutionary analyses will undoubtedly offer a cutting-edge insight on the Fly tree of Life. We were extremely enthusiastic to add Tachinidae genome during the early stages of our work (2017-2018), however due to the lack of Tachinidae genome, we were unable to do so. As the number of genomic resources grows at a geometric rate, more efficient and effective methods are required for phylogenetic analysis using enormous genomic datasets.

7.4 Classification of Diptera based on different dietary adaptation and the tempo and mode of their diversification.

In this chapter, we hypothesized that different nucleotide substitution in mitochondrial genes have influenced establishing successful niche in different climates, lifestyles, and different

source of diet. As mitochondria provides more than 95% of the energy of the cell and mitochondrial ATP generation bioenergetic efficiency depends on variability of nutrient molecule and availability of oxygen. Also, mitochondria generate heat through uncoupled respiration and aids in thermoregulation during unusual climatic conditions. Our investigation established that food habit of Diptera is significantly correlated with nucleotide substitution either nonsynonymous (dN) or synonymous (dS) or their ratio (dN/dS). Variations in selection patterns in mitochondrial genes due to change in the source of nutrients mean that dietary ability has a differential effect on mitochondrial energy metabolism. Our findings also reveals that the rate of diversification varies based on dietary habits. In addition, role of dS, dN and ω as continuous trait and we employ phylogenetic modelling to investigate the tempo and mode of evolutionary diversification of these organisms. Furthermore, this chapter also finds that lineages with different food habit evolve towards their optimal continuous molecular trait value.

Future Prospect:

This Chapter was started based on my intuition during literature survey as we observed that Diptera flies are capable with different diet, they omnipresent in different climate and in nature the live as endo-parasite or as ecto-parasite. Since, mitochondria are the energy and heat producing organelle in the cell and it needs nutrition and oxygen to produce energy therefore, changes in the availability of nutrition and oxygen might influence mitochondrial function. In addition to energy production mitochondria also produce heat and similarly which might corelated with climate. We found that diet is significantly corelated with different nucleotide substitution in mitochondria. However, further studies based on nuclear-encoded OXPHOS genes should be of great help in validating these conclusions. To the best of my knowledge, this is the first study of its kind that evaluates nucleotide substitution rate as a continuous trait that correlates with dietary habits of any species group. Thus, the study presented in this chapter

undeniably provides a novel perspective on phenotypic trait and molecular evolution in unison. The framework we present may be expanded with a larger taxon sample, thorough character coding, and data from the fossil record to provide further insights and a more appropriate conclusion.

7.5 Deciphering major diversification time through molecular dating and dynamics of speciation-extinction events of Diptera.

Here in the final chapter, we estimated diversification time of dipteran major clades from mitochondrial datasets using different calibration methods. We compared variation of time assuming different nucleotide substitution rate (similar, different), different branching prior process (yule and birth-death) using different molecular dating methods. This chapter indicates that by increasing the number calibration point, the time of diversification events became older in stem lineages. Different dating methods have a significant influence on the chronology of key Diptera lineages and subsequent macroevolutionary studies. Considerable variation in diversification rates linked with distinct Diptera traits, as well as diversification that is dependent on diversity. According to palaeo-environment diversification models, high extinction rates are caused by global warming, whereas low speciation rates are driven by rising sea levels. This study also showed the role of mass extinction in shaping the current Diptera diversity. This chapter confirms the essence of the confluence of several factors rather than single explanations in modeling diversification within lineages.

Future Prospect:

The estimation of diversification time with mitochondrial dataset with multiple strategy provides incongruence in estimation but gives a broad idea about the range of diversification the Diptera. Lack of appropriate taxon sampling, inability to detect discrete evolutionary units, and challenges describing diversity patterns in temporal and geographical settings are some of

the drawbacks. Nonetheless, this work presents a number of macroevolutionary perspectives for Diptera diversity that have not previously been considered. This chapter presents evidence of Diptera's diversity dynamics to historical climatic change, which have been rarely documented.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी • Indian Institute of Technology Guwahati •

CURRICULUM VITAE

Debajyoti Kabiraj

Bioengineering Research Laboratory (BERL),
Dept. of Biosciences and Bioengineering,
Indian Institute of Technology Guwahati,
India-781039

Personal Profile

Date of birth	22nd May, 1990
Gender	Male
Nationality	Indian
Languages	English, Bengali, Hindi, Assamese
Marital status	Unmarried
Present Address	Bioengineering Research Laboratory (BERL) Department of Biosciences and Bioengineering O-Block, Academic Complex, Indian Institute of Technology Guwahati (IITG) Guwahati-781039, Assam, India Email d.kabiraj@iitg.ac.in, kabir.deb0355@gmail.com Phone no: +91 8486500157, +91 8167840357

Education

Degree	University/Institute	Specialization	Duration	Marks
Doctor of Philosophy	Indian Institute of Technology Guwahati, Assam, India	Biosciences and Bioengineering	2014-	
Master of Technology	West Bengal University of Technology, Kolkata, West Bengal	Bioinformatics	2012-2014	7.96/10
Batchelor of Technology	Haldia Institute of Technology, Haldia, West Bengal	Biotechnology	2008-2012	7.29/10
Higher-Secondary	Siliguri Boys' High School, West Bengal Council for Higher-Secondary Education	PCMB	2008	73.4%
Secondary	Siliguri Boys' High School, West Bengal Board of Secondary Education		2006	83.4%

Research Experience	
Doctoral research on Mito-genomics, Genomics and Evolutionary Biology at the Bioengineering Research Laboratory (BERL), Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati	Doctoral Supervisor: Prof. Utpal Bora **Thesis title: Evolutionary landscape of dipteran insects**
Master's Dissertation at the Department of Bioinformatics, West Bengal University of Technology, Kolkata, West Bengal	Supervisor: Dr. Raja Banerjee (WBUT) Prof. Bidyut Roy (ISI Cal) Dissertation title: Structural changes in mitochondrial protein due to mutation in genes.

Publications	
2022	**Kabiraj, D.,** Chetia, H., Nath, A., Sharma, P., Mosahari, P. V., Singh, D., ... & Bora, U. (2022). Mitogenome-wise codon usage pattern from comparative analysis of the first mitogenome of Blepharipa sp. (Muga uzifly) with other Oestroid flies. Scientific Reports, 12(1), 1-33.
2019	Chetia H, **Kabiraj D**, Bharali B, Ojha S, Barkataki MP, Saikia D, Singh T, Mosahari PV, Sharma P, Bora U. 2019. Exploring the benefits of endophytic fungi via omics, p. 51–81. In. Springer, Cham.
2018	Ojha S, Singh D, Sett A, Chetia H, **Kabiraj D**, Bora U. 2018. Nanotechnology in Crop Protection. Nanomater Plants, Algae, Microorg 345–391.
2018	Mosahari PV, Singh D, Kalita JJ, Sharma P, Chetia H, **Kabiraj D**, Mahanta C, Bora U. Nanotoxicity: Impact on Health and Environment. Environmental Toxicity of Nanomaterials. 2018 Apr 17:21.
2017	Chetia H, **Kabiraj D**, Singh D, Mosahari PV, Das S, Sharma P, Neog K, Sharma S, Jayaprakash P, Bora U. 2017. De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). Gene 611.
2017	Singh D, **Kabiraj D**, Sharma P, Chetia H, Mosahari PV, Neog K, Bora U. 2017. The mitochondrial genome of muga silkworm (Antheraea assamensis) and its comparative analysis with other lepidopteran insects. PLoS One 12.
2017	Chetia H, **Kabiraj D**, Sharma S, Bora U. 2017. Comparative insights to the transportome of Nosema: a genus of parasitic microsporidians. bioRxiv 110809.
2016	Chattopadhyay E, De Sarkar N, Singh R, Ray A, Roy R, Paul RR, Pal M, Ghose S, Ghosh S, **Kabiraj D**, Banerjee R. Genome-wide mitochondrial DNA sequence variations and lower expression of OXPHOS genes predict mitochondrial dysfunction in oral cancer tissue. Tumor Biology. 2016 Sep;37(9):11861-71.
2016	Singh D, Chetia H, **Kabiraj D**, Sharma S, Kumar A, Sharma P, Deka M, Bora U. 2016. A comprehensive view of the web-resources related to sericulture. Database 2016:baw086.
2015	**Kabiraj D**, Kalita J, Chetia H, Singh D, Bora U. 2015. Expanding the frontiers of rice research through omics. Assam Sc. Soc. Vo. p. 1-28.
2015	Kumar A, Chetia H, Sharma S, **Kabiraj D**, Talukdar NC, Bora U. 2015. Curcumin Resource Database. Database 2015:bav070.

Workshops
BIOCONVERSE 2018 (One day workshop of Wildlife ecology and Seri bioresources) organized by Directorate of Sericulture (BTC) & College of Veterinary Sciences (AAU), Khanapara at Manas National Park & BTAD, Assam from 30th Jan- 01st Feb 2018.
"North-East Winter School on Human Genetics 2016-Genetic Analyses of Complex Traits" organized jointly by Dibrugarh University and Indian Statistical Institute Kolkata from 21st to 22nd December, 2016.
AICTE Sponsored short-term course **"ADVANCES IN ENVIRONMENTAL HEATH"** organized by Centre for the Environment, IIT Guwahati on 9th -13th February 2015
TEQIP sponsored **"Winter School 2013"** organized by West Bengal University of Technology on 28th February – 1st March.
Conferences, Seminar and Symposia
"International symposium on biodiversity and biobanking (BIODIVERSE) 2018" with Association for Promotion of DNA Fingerprinting and Other DNA Technologies (ADNAT) at IIT Guwahati from 27th to 29th January, 2018.
One-day Capacity Building Workshop-cum-Brainstorming Meeting on **"River ecosystems and fresh water biodiversity research (REFRESH) 2018"** at IIT Guwahati on 2nd February, 2018.
"Nextgen Genomics, Biology, Bioinformatics and Technologies Conference (NGBT 2017)" at Bhubaneswar, Odisha from 02nd to 4th October, 2017. (Partial grant awarded)
Oral Presentation at **"SeriTech 2017"** organized by Unit of Excellence on Seribiotechnology, Centre for the Environment, IIT Guwahati on 24th January, 2017.
National Symposium on **"IPR in Innovation and Entrepreneurship"** conducted under the Technical Education Quality Improvement Programme sponsored by MHRD, GOI held on 16th March 2015.
Oral Presentation at **"International Conference on Disease Biology and Therapeutics, ICDBT 2014"** organized by Institute of Advance Study in Science & Technology, Guwahati-35, Assam on 3rd -5th December 2014.
Conference on **"Exploitation of Seribiodiverity for Novel Product Development"** at IIT Guwahati from 29th to 30th November, 2014.
Plenary Session on **"FOOD SECURITY & GENETICALLY MODIFIED CROPS"** organized by West Bengal Council of Science & Technology and Calcutta University on 25th July 2013.
"4th Indian Peptide Symposium" organized by Indian Peptide Society on 22nd February 2013.
National Conference on **"Biodiversity: Threats and Conservation through traditional and Biotechnological Approaches"** organized by Dum Dum Motijhil College on 4th-6th February 2012.
"ONCON-2010" organized by School of Biotechnology and Life Sciences, Haldia Institute of Technology on 13th January 2010.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

scientific reports

Check for updates

OPEN

Mitogenome-wise codon usage pattern from comparative analysis of the first mitogenome of *Blepharipa* sp. (Muga uzifly) with other Oestroid flies

Debajyoti Kabiraj[1], Hasnahana Chetia[1], Adhiraj Nath[1], Pragya Sharma[3], Ponnala Vimal Mosahari[2], Deepika Singh[1], Palash Dutta[4], Kartik Neog[4] & Utpal Bora[1,2]✉

Uziflies (Family: Tachinidae) are dipteran endoparasites of sericigenous insects which cause major economic loss in the silk industry globally. Here, we are presenting the first full mitogenome of *Blepharipa* sp. (Acc: KY644698, 15,080 bp, A + T = 78.41%), a dipteran parasitoid of Muga silkworm (*Antheraea assamensis*) found in the Indian states of Assam and Meghalaya. This study has confirmed that *Blepharipa* sp. mitogenome gene content and arrangement is similar to other Tachinidae and Sarcophagidae flies of Oestroidea superfamily, typical of ancestral Diptera. Although, Calliphoridae and Oestridae flies have undergone tRNA translocation and insertion, forming unique intergenic spacers (IGS) and overlapping regions (OL) and a few of them (IGS, OL) have been conserved across Oestroidea flies. The Tachinidae mitogenomes exhibit more AT content and AT biased codons in their protein-coding genes (PCGs) than the Oestroidea counterpart. About 92.07% of all (3722) codons in PCGs of this new species have A/T in their 3rd codon position. The high proportion of AT and repeats in the control region (CR) affects sequence coverage, resulting in a short CR (*Blepharipa* sp.: 168 bp) and a smaller tachinid mitogenome. Our research unveils those genes with a high AT content had a reduced effective number of codons, leading to high codon usage bias. The neutrality test shows that natural selection has a stronger influence on codon usage bias than directed mutational pressure. This study also reveals that longer PCGs (e.g., *nad5*, *cox1*) have a higher codon usage bias than shorter PCGs (e.g., *atp8*, *nad4l*). The divergence rates increase nonlinearly as AT content at the 3rd codon position increases and higher rate of synonymous divergence than nonsynonymous divergence causes strong purifying selection. The phylogenetic analysis explains that *Blepharipa* sp. is well suited in the family of insectivorous tachinid maggots. It's possible that biased codon usage in the Tachinidae family reduces the effective number of codons, and purifying selection retains the core functions in their mitogenome, which could help with efficient metabolism in their endo-parasitic life style and survival strategy.

Insect mitochondria which arose from alpha-proteobacteria have its own circular mitogenome of about 14–20 kb[1–3]. The inner membrane of this organelle harbors five distinct protein complexes for efficient production of energy via oxidative phosphorylation (OXPHOS) process[4,5]. In general, the insect mitogenome has 13 protein-coding genes (PCGs), 2 ribosomal RNAs (rRNAs), 21 to 23 transfer RNAs (tRNAs)[6]. It also contains several non-coding regions with the lengthiest being AT-rich control region (Table 1)[7]. A typical metazoan mitogenome is small in size, maternally inherited, mutation prone, has minimal or no homologous recombination, with conserved gene content, and high genetic polymorphism, making it a potential sequence for barcoding, phylogeography, phylogenetic and molecular dating research[8–10]. However, little attention has been paid to the study of

[1]Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India. [2]Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati, Assam, India. [3]Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati, Assam, India. [4]Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, Jorhat, Assam, India. ✉email: ubora@iitg.ac.in

TH-2821_146106003

PLOS ONE

RESEARCH ARTICLE

The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects

Deepika Singh[1,2], Debajyoti Kabiraj[1], Pragya Sharma[3], Hasnahana Chetia[1], Ponnala Vimal Mosahari[2], Kartik Neog[4], Utpal Bora[1,2]*

1 Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India, **2** Centre for the Environment, Indian Institute of Technology Guwahati, Assam, India, **3** Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati, Assam, India, **4** Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, Jorhat, Assam, India

* ubora@iitg.ernet.in, ubora@rediffmail.com

Check for updates

OPEN ACCESS

Citation: Singh D, Kabiraj D, Sharma P, Chetia H, Mosahari PV, Neog K, et al. (2017) The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. PLoS ONE 12(11): e0188077. https://doi.org/10.1371/journal.pone.0188077

Editor: Daniel Doucet, Natural Resources Canada, CANADA

Received: June 28, 2017

Accepted: October 31, 2017

Published: November 15, 2017

Copyright: © 2017 Singh et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The full annotated mitogenome sequence and SRA data of A. assamensis submitted to NCBI GenBank are available under the accession numbers KU379695 and SRR3948351, respectively.

Funding: The authors thank the Department of Biotechnology, Govt. of India, New Delhi for supporting the research through the UXCEL project

Abstract

Muga (*Antheraea assamensis*) is an economically important silkmoth endemic to the states of Assam and Meghalaya in India and is the producer of the strongest known commercial silk. However, there is a scarcity of genomic and proteomic data for understanding the organism at a molecular level. Our present study is on decoding the complete mitochondrial genome (mitogenome) of *A. assamensis* using next generation sequencing technology and comparing it with other available lepidopteran mitogenomes. Mitogenome of *A. assamensis* is an AT rich circular molecule of 15,272 bp (A+T content ~80.2%). It contains 37 genes comprising of 13 protein coding genes (PCGs), 22 tRNA and 2 rRNA genes along with a 328 bp long control region. Its typical $tRNA^{Met}$-$tRNA^{Ile}$-$tRNA^{Gln}$ arrangement differed from ancestral insects ($tRNA^{Ile}$-$tRNA^{Gln}$-$tRNA^{Met}$). Two PCGs *cox1* and *cox2* were found to have CGA and GTG as start codons, respectively as reported in some lepidopterans. Interestingly, *nad4l* gene showed higher transversion mutations at intra-species than inter-species level. All PCGs evolved under strong purifying selection with highest evolutionary rates observed for *atp8* gene while lowest for *cox1* gene. We observed the typical clover-leaf shaped secondary structures of tRNAs with a few exceptions in case of $tRNA^{Ser1}$ and $tRNA^{Tyr}$ where stable DHU and TΨC loop were absent. A significant number of mismatches (35) were found to spread over 19 tRNA structures. The control region of mitogenome contained a six bp (CTTAGA/G) deletion atypical of other *Antheraea* species and lacked tandem repeats. Phylogenetic position of *A. assamensis* was consistent with the traditional taxonomic classification of Saturniidae. The complete annotated mitogenome is available in GenBank (Accession No. KU379695). To the best of our knowledge, this is the first report on complete mitogenome of *A. assamensis*.

TH-2821_146106003

Gene 611 (2017) 54–65

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer)

Hasnahana Chetia [a], Debajyoti Kabiraj [a], Deepika Singh [a], Ponnala Vimal Mosahari [b], Suradip Das [a], Pragya Sharma [c], Kartik Neog [d], Swagata Sharma [a], P. Jayaprakash [e], Utpal Bora [a,b,*]

[a] Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam 781039, India
[b] Centre for the Environment, Indian Institute of Technology Guwahati, Assam 781039, India
[c] Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati 781014, Assam, India
[d] Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, 785700 Jorhat, Assam, India
[e] Central Silk Board (CSB), Bangalore 506068, Karnataka, India

ARTICLE INFO

Article history:
Received 20 October 2016
Received in revised form 29 January 2017
Accepted 15 February 2017
Available online 17 February 2017

Keywords:
Alimentary canal
Antimicrobial peptide
Lepidoptera
Machilus bombycina
Next generation sequencing
Residual body
Saturniidae
Sericin
Silk gland
Silk gland factor

ABSTRACT

Antheraea assamensis (Lepidoptera: Saturniidae), is a semi-domesticated silkworm known to be endemic to Assam and the adjoining hilly areas of Northeast India. It is the only producer of a unique, commercially important variety of golden silk called "muga silk". Herein, we report the *de novo* transcriptome of *A. assamensis* reared on *Machilus bombycina* leaves for the first time. Short reads generated by high throughput sequencing of cDNA libraries from multiple tissues, viz. alimentary canal, silk gland and residual body of the 5th instar of muga silkworm were assembled into transcripts via a *de novo* assembly pipeline followed by functional annotation and classification. A total of 1,21,433 transcripts were generated from ~231 million raw reads of which ~74% (89,583) were either allocated a functional annotation or categorized under Pfam/COG/KEGG categories. Identification of differentially expressed transcripts and their comparative sequence analysis revealed candidate genes related to silk synthesis, viz. silk gland factor-1 and 3, sericin-like transcript, etc. with conserved forkhead, homeo- and POU domains. Several candidate anti-microbial peptides which may have potential anti-bacterial, anti-fungal or anti-parasitic activity in *A. assamensis* were also identified. T/A and AT/TA were predicted to be the most abundant mono- and di-nucleotide simple sequence repeat markers in the transcriptome. Transcriptome validation was carried out by quantitative real-time PCR (qPCR) amplification of eight transcripts. The resources generated by this study will expand the periphery of existing genomic data on *A. assamensis* facilitating future in-depth studies on its unknown aspects.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Silk is an important cultural and commercial fibre largely obtained from silkworms. Some silkworms have been domesticated by humans over a period of time to exploit their potential in textiles. Still, majority of them remain semi-domesticated or wild. The silk proteins- fibroin and sericin of domesticated mulberry silkworm, *Bombyx mori* (Family: Bombycidae) has been extensively used for tissue engineering applications (Das et al., 2014). Research on these has been further accelerated by discovery of its complete genome and transcriptome (Li et al., 2012; The International Silkworm Genome Consortium, 2008). Similar as well as novel research applications can also be expected from biomaterials of other less-studied semi- or undomesticated silkworms. The dearth of information and hindrances in domestication of these silkworms currently restricts their usage in such applications. One such semi-domesticated silkworm is *Antheraea assamensis* (Helfer), also known as the "muga silkworm". *A. assamensis* ($n = 15$) is a multivoltine, polyphagus silkworm classified under the order - Lepidoptera and family - Saturniidae. It is mostly endemic to the Brahmaputra valley of Assam and adjoining hilly areas of Northeast India (Tikader et al., 2013). Being the sole producer of globally acclaimed "muga silk", a lustrous golden yellow fabric, makes *A. assamensis* one of the most important components of the Assamese silk industry and it hugely contributes towards employment generation in North-Eastern India. The unique quality of this silk-based textile earned it a geographical indication tag

Abbreviations: AaCbp, *Antheraea assamensis* Carotenoid binding protein; AaFhc, *Antheraea assamensis* Fibroin heavy chain; AC, Alimentary Canal; AMP, Anti-microbial peptides; CEG, Core eukaryotic genes; COG, Cluster of Orthologous genes; EST, Expressed Sequence Tag; GO, Gene Ontology; KEGG, Kyoto Encyclopaedia of Genes and Genomes; MISA, MIcroSAtellite Identification Tool; PNTAa, Putative novel transcripts of *Antheraea assamensis*; POU, Pit-Oct-Unc; qPCR, Quantitative real time PCR; RB, Residual body; SG, Silk gland; SRA, Short Read Archive; SSR, Simple sequence repeats.

* Corresponding author at: Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India.
E-mail addresses: ubora@iitg.ernet.in, drutpalbora@gmail.com (U. Bora).

http://dx.doi.org/10.1016/j.gene.2017.02.021
0378-1119/© 2017 Elsevier B.V. All rights reserved.
TH-2821_146106003

DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–31
doi: 10.1093/database/baw086
Review

Review

A comprehensive view of the web-resources related to sericulture

Deepika Singh[1], Hasnahana Chetia[1], Debajyoti Kabiraj[1], Swagata Sharma[1], Anil Kumar[2], Pragya Sharma[3], Manab Deka[3] and Utpal Bora[1,4,5,*]

[1]Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India, [2]Centre for Biological Sciences (Bioinformatics), Central University of South Bihar (CUSB), Patna 800014, India, [3]Department of Bioengineering & Technology, Gauhati University Institute of Science & Technology, Gauhati University, Guwahati, Assam 781014, India, [4]Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India and [5]Mugagen Laboratories Pvt. Ltd, Technology Incubation Centre, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

*Corresponding Author: Tel: +913612582215; Fax: +913612582249; Email: ubora@iitg.ernet.in; ubora@rediffmail.com

Citation details: Singh,D., Chetia,H., Kabiraj,D. *et al.* A comprehensive view of the current web-resources in sericulture and related fields. *Database* (2016) Vol. 2016: article ID baw086; doi:10.1093/database/baw086

Received 21 January 2016; Revised 25 April 2016; Accepted 2 May 2016

Abstract

Recent progress in the field of sequencing and analysis has led to a tremendous spike in data and the development of data science tools. One of the outcomes of this scientific progress is development of numerous databases which are gaining popularity in all disciplines of biology including sericulture. As economically important organism, silkworms are studied extensively for their numerous applications in the field of textiles, biomaterials, biomimetics, etc. Similarly, host plants, pests, pathogens, etc. are also being probed to understand the seri-resources more efficiently. These studies have led to the generation of numerous seri-related databases which are extremely helpful for the scientific community. In this article, we have reviewed all the available online resources on silkworm and its related organisms, including databases as well as informative websites. We have studied their basic features and impact on research through citation count analysis, finally discussing the role of emerging sequencing and analysis technologies in the field of seri-data science. As an outcome of this review, a web portal named SeriPort, has been created which will act as an index for the various sericulture-related databases and web resources available in cyberspace.

Database URL: http://www.seriport.in/

© The Author(s) 2016. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

TH-2821_146106003

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/baw086/2630457 by Indian Institute of Technology Guwahati user on 07 May 2022

Database, 2015, 1–6
doi: 10.1093/database/bav070
Original article

Downloaded from https://academic.oup.com/database/article/doi/10.1093/database/bav070/2433203 by Indian Institute of Technology Guwahati user on 07 May 2022

Original article

Curcumin Resource Database

Anil Kumar[1,2], Hasnahana Chetia[1], Swagata Sharma[1], Debajyoti Kabiraj[1], Narayan Chandra Talukdar[3,*] and Utpal Bora[1,4,*]

[1]Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati (IITG), Assam 781039, India, [2]Centre for Biological Sciences (Bioinformatics), Central University of South Bihar (CUSB), Patna 800014, India, [3]Institute of Advanced Studies on Science and Technology (IASST) Boragaon, Guwahati, Assam 781035, India and [4]Institutional Biotech Hub, Centre for the Environment, Indian Institute of Technology Guwahati (IITG), Assam 781039, India

*Corresponding author: Tel: +91 361 2582215; Fax: +91 361 2582249; Email: ubora@iitg.ernet.in, ubora@rediffmail.com

Correspondence may also be addressed to Narayan Chandra Talukdar. Tel: +91 361 2273058; Fax: +91 361 2273062; Email: nctalukdar@yahoo.com

Citation details: Kumar,A., Chetia,H., Sharma,S., *et al.* Curcumin Resource Database. *Database* (2015) Vol. 2015: article ID bav070; doi:10.1093/database/bav070

Received 29 April 2015; Revised 10 June 2015; Accepted 26 June 2015

Abstract

Curcumin is one of the most intensively studied diarylheptanoid, *Curcuma longa* being its principal producer. This apart, a class of promising curcumin analogs has been generated in laboratories, aptly named as Curcuminoids which are showing huge potential in the fields of medicine, food technology, etc. The lack of a universal source of data on curcumin as well as curcuminoids has been felt by the curcumin research community for long. Hence, in an attempt to address this stumbling block, we have developed Curcumin Resource Database (CRDB) that aims to perform as a gateway-cum-repository to access all relevant data and related information on curcumin and its analogs. Currently, this database encompasses 1186 curcumin analogs, 195 molecular targets, 9075 peer reviewed publications, 489 patents and 176 varieties of *C. longa* obtained by extensive data mining and careful curation from numerous sources. Each data entry is identified by a unique CRDB ID (identifier). Furnished with a user-friendly web interface and in-built search engine, CRDB provides well-curated and cross-referenced information that are hyperlinked with external sources. CRDB is expected to be highly useful to the researchers working on structure as well as ligand-based molecular design of curcumin analogs.

Database URL: http://www.crdb.in

Introduction

Curcumin (diferuloylmethane) is a hydrophobic polyphenol derived from rhizome of the perennial herb turmeric (*Curcuma longa*) which belongs to the ginger family (Zingiberaceae) native to tropical South Asia (1). Numerous traditional usage of turmeric is described in

© The Author(s) 2015. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

TH-2821_146106003

J. Assam Sc. Soc. Vo. 56, No.2
December 2015; PP. 1-28

ISSN 0587-1921

EXPANDING THE FRONTIERS OF RICE RESEARCH THROUGH OMICS

Debajyoti Kabiraj[1], Jonjyoti Kalita[1], Hasnahana Chetia[1], Deepika Singh[1], *Utpal Bora[1,2]

[1]Bioengineering Research Laboratory, Dept. of Biosciences and Bioengineering, Indian Institute of Technology, Guwahati (IITG), Assam 781039, India

[2]Institutional Biotech Hub, Centre for the Environment, Indian Institute of Technology Guwahati, (IITG), Assam 781039, India

**Corresponding author :* Tel: +91 361 2582215; Fax: +91 361 2582249;
Email: ubora@iitg.ernet.in

ABSTRACT

Rice is an important food grain and a staple food for nearly 60% of the world's population. In post-genomic era, comprehensive data from "omics" technologies have been extensively applied to rice to elucidate its cellular intricacies and regulatory mechanisms. The present review deals with current scenario of "omics" applications in understanding this important food crop and highlights the essentiality of genomics and transcriptomics in rice research. It includes a brief coverage of improvement in genetic mapping, rice trait analysis, developmental understanding of the rice embryogenesis etc. aided by high-throughput sequencing technologies. The review also discusses application of basic and applied proteomics and metabolomics research in providing new directions to improvement of qualitative and quantitative traits. These include another proteome and seed development, tackling of biotic and abiotic stress, pathogen infections etc. Towards the end, the review addresses the application of data science into this field, shortlisting relevant database (for the omics) discussed above.

Key words : Rice, *Oryza sativa,* Genomics, Transcriptomics, Proteomics, Metabolomics, Database

TH-2821_146106003