UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA MOLECULAR

Carlos Congrains Castillo

Padrões evolutivos inferidos por transcriptomas de moscas-das-frutas do gênero *Anastrepha* (Diptera: Tephritidae)

São Carlos – SP

Dezembro/2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA MOLECULAR

Carlos Congrains Castillo

Padrões evolutivos inferidos por transcriptomas de moscas-das-frutas do gênero *Anastrepha* (Diptera: Tephritidae)

Tese de Doutorado apresentada ao Programa de Pós-graduação em Genética Evolutiva e Biologia molecular do Centro de Ciências Biológicas e da Saúde da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciências com área de concentração em Genética Evolutiva.

Orientador: Prof. Dr. Reinaldo Alves de Brito

São Carlos – SP
Dezembro/2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Biológicas e da Saúde
Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Carlos Congrains Castillo, realizada em 21/12/2017:

Prof. Dr. Reinaldo Otavio Alvarenga Alves de Brito
UFSCar

Prof. Dr. Evandro Marsola de Moraes
UFSCar

Caio Freire

Prof. Dr. Caio Cesar de Melo Freire
UFSCar

Profa. Dra. Vera Nisaka Solferini
UNICAMP

Prof. Dr. José Salvatore Leister Patané
Instituto Butantan

Dedico esse trabalho à minha esposa Karla,
meus pais Ada e Carlos e minha irmã Ada
pela confiança em mim depositada.

AGRADECIMENTOS

Ao Prof. Dr. Reinaldo Brito por sua amizade, conselhos, extensas discussões e porque apesar de ter passado por diferentes situações durante estes anos manteve seu espírito firme que permitiu concluir este trabalho.

À minha família que nunca deixou de acreditar em mim. À minha esposa Karla que foi um excepcional suporte acadêmico e emocional. Aos meus pais por estar sempre perto de mim apesar da distância física. À Ada, minha irmã, quem me ajudou durante este difícil caminho.

À Manu, Felipe, Samira, Emeline, Mário, Isabel, Víctor, Iderval, André, Janaína, Aline, Chris meus amigos do laboratório, com os que passei muitas horas e fizeram mais fácil este doutorado.

Ao Jorge pela amizade de não sei quantos e os bons momentos com a turma do basquete.

À Danila ter me dado sua confiança para analisar os dados dos *Rhodnius*, que finalmente foi uma importante publicação. Ao Elder e o Prof. Marco pela parceria científica e amizade. Ao Prof. Belo por ter gentilmente cedido a lupa para a análise das amostras. À Paola por ter me ensinado os segredos da programação em Python.

Ao programa de pós-graduação, ao CNPq por ter concedido a bolsa (160002/2013-3), à FAPESP (2010/20455-4) e CAPES (PVE 056/13) pelo financiamento do projeto.

"*... these forms may still be only ... varieties; but we have only to suppose the steps of modification to be more numerous or greater in amount, to convert these forms into species ... thus species are multiplied.*"

Charles Darwin 1859

RESUMO

Anastrepha é um gênero que possui uma grande diversidade de espécies, cuja distribuição geográfica inclui a região tropical e subtropical das Américas. Espécies deste gênero têm sido classificadas em 21 grupos de espécies, sendo o grupo *fraterculus* o de maior importância econômica porque inclui espécies com hábitos polífagos consideradas importantes pragas para a agricultura. Este grupo apresenta também espécies crípticas e proximamente relacionadas que provavelmente divergiram com fluxo gênico, tais como o complexo *A. fraterculus, A. obliqua* e *A. sororcula*. Apesar da importância de *Anastrepha*, não estão claros os mecanismos associados à sua rápida diversificação, assim como as relações filogenéticas das espécies do grupo. Para investigar padrões e mecanismos de diferenciação de espécies nesse gênero, focamos em dois aspectos correlatos de sua evolução. Em um primeiro passo, investigamos transcriptomas de machos e fêmeas expressos em tecidos reprodutivos e cefálicos de *A. fraterculus* e *A. obliqua*, e identificamos que genes com expressão aumentada em machos evoluem mais rápido do que os aumentados em fêmeas e os não enviesados, o que pode ser atribuído a uma combinação de seleção positiva e relaxada. Alguns dos genes sob seleção positiva com expressado enviesada em machos estão relacionados a fertilidade e comportamento de corte, o que sugere que poderiam estar envolvidos no processo de diferenciação destas espécies. Este trabalho também contribuiu para resolver as relações filogenéticas entre importantes espécies do gênero, para tal investigamos 20 transcriptomas de tecido reprodutivo de fêmeas pertencentes a 10 linhagens. Esses dados permitiram a identificação de milhares de genes ortólogos a partir dos quais inferimos relações filogenéticas baseadas em métodos coalescentes multiespécies. As filogenias revelaram que o grupo *serpentina* está em posição basal ao resto de grupos amostrados e o grupo *bistrigata* se posicionou como grupo irmão do grupo *fraterculus*. As relações entre espécimes do grupo *fraterculus* mostraram que *A. obliqua* é uma única linhagem, enquanto que o complexo *A.*

fraterculus do Brasil foi dividido em dois grupos, sendo um provavelmente *A. fraterculus* Brazil 1 e a outra não foi possível de relacionar com as linhagens do complexo previamente encontradas. Além disso, apesar de *A. distincta* apresentar caraterísticas ecológicas divergentes (especialistas para um hospedeiro) em relação à *A. fraterculus* (*s.l.*) e *A. turpiniae,* a árvore a posicionou como sendo proximamente relacionadas a essas espécies. Também foram encontrados um alto nível de incongruência entre árvores de genes devido ao sorteamento incompleto de linhagens e à introgressão. Interessantemente, detectamos extenso sinal de introgressão ancestral entre as linhagens do grupo *fraterculus*. Este estudo não apenas aumentou consideravelmente a informação genética disponível para estas espécies de importância econômica, mas também determinou que seleção positiva e hibridação estão envolvidas na diferenciação do grupo *fraterculus*.

ABSTRACT

Anastrepha is a species-rich genus, with a geographic distribution that covers tropical and subtropical regions of the Americas. Twenty-one species groups have been recognized in this genus, the most economically relevant being the *fraterculus* group, which has several representatives which are considered important agricultural pests. Moreover, this group also bears cryptic and closely related species that probably diverged with gene flow such as species in the *A. fraterculus* complex, *A. obliqua* and *A. sororcula*. Despite their importance, the evolutionary mechanisms involved in the rapid diversification of this group is unclear as well as their phylogenic relationships. To investigate the mechanisms of differentiation of these species, we analyzed male and female transcriptomes from reproductive and head tissues of *A. fraterculus* and *A. obliqua*. We found that male-biased genes evolve faster than female-biased and unbiased genes due to positive selection and relaxed selective constraints. Some of the positively selected and male-biased genes are involved with courtship behavior and fertility, which suggests that these genes may be involved in the differentiation of these species. We also investigated the phylogenetic relationships of *Anastrepha* lineages evaluating 20 female reproductive transcriptomes belonging to different 10 lineages. We inferred a cluster of high-quality orthologs and reconstructed a robust phylogeny of this group based on thousands of genes using multispecies coalescent methods. Our phylogenetic analysis revealed that the *serpentina* group is basal to other groups here evaluated, and the *bistrigata* group is sister to the *fraterculus* group. The relationships among *fraterculus* groups specimens showed that *A. obliqua* formed only one lineage, whereas *A. fraterculus* complex from Brazil was divided into two groups, one probably including individuals from *A. fraterculus* Brazil 1 and another which its assignment to previously reported lineage remains unknow. Furthermore, although *A. distincta* has divergent ecological traits (host specialist) to *A. fraterculus* (*s.l.*) and *A. turpiniae*, the tree revealed that it is closely related to these species. We also found high levels of gene tree discordance due to incomplete lineage sorting and introgression. Interestingly, our findings indicated extensive ancestral introgression among *fraterculus* group lineages. Our study increased the genetic data available for these economically important species, and established a role of positive selection and hybridization in the diversification of the *fraterculus* group.

LISTA DE FIGURAS

CHAPTER I - INTRODUCTION AND GOALS

CHAPTER II- EVIDENCE OF ADAPTIVE EVOLUTION AND RELAXED CONSTRAINTS IN SEX-BIASED GENES OF SOUTH AMERICAN AND WEST INDIES FRUIT FLIES (DIPTERA: TEPHRITIDAE)

CHAPTER III - PHYLOGENOMIC APPROACH REVEALS SIGNATURES OF INTROGRESSION AMONG NEOTROPICAL TRUE FRUIT FILES (ANASTREPHA: TEPHRITIDAE)

LISTA DE TABELAS

CHAPTER II- EVIDENCE OF ADAPTIVE EVOLUTION AND RELAXED CONSTRAINTS IN SEX-BIASED GENES OF SOUTH AMERICAN AND WEST INDIES FRUIT FLIES (DIPTERA: TEPHRITIDAE)

CHAPTER III - PHYLOGENOMIC APPROACH REVEALS SIGNATURES OF INTROGRESSION AMONG NEOTROPICAL TRUE FRUIT FILES (ANASTREPHA: TEPHRITIDAE)

Sumário

CHAPTER I

INTRODUCTION AND GOALS

CHAPTER I – INTRODUCTION AND GOALS

1.1. Introduction

1.1.1. Tephritidae, *Anastrepha* and *fraterculus* groups

Drosophilidae and Tephritidae are two dipteran families whose members are commonly referred to as fruit flies. However, the great majority of Drosophilidae species are indirectly associated to fruits, because they feed on fungi and bacteria that live on decaying fruits (ASHBURNER; GOLIC; HAWLEY, 2005). On the other hand, Tephritidae feeds on fruits, because of that they are also referred as to true fruit flies. This family includes more than 500 genera, among them four which are the most economically important: *Ceratitis*, *Bactrocera*, *Rhagoletis* and *Anastrepha* (ALUJA; NORRBOM, 1999; NORRBOM, 2004).

Anastrepha Schiner (Diptera: Tephritidae) is a highly diverse genus, harboring more than 270 species (NORRBOM et al., 2012 onwards). Species of this genus are endemic and widely distributed in tropical and subtropical regions of the Americas (ALUJA, 1994). Such a tremendous diversity is in the genus has been divided into 21 groups based on morphology and chromosomal numbers (NORRBOM; ZUCCHI; HERNÁNDEZ-ORTIZ, 1999; NORRBOM et al., 2012 onwards). This thesis is focused on the *fraterculus* group, from which over half of the 34 described species have been reported in Brazil (NORRBOM et al., 1999; ZUCCHI, 2000a), including some which are major pests with generalist habits (ALUJA, 1994; NORRBOM et al., 1999). *A. fraterculus*, *A. obliqua* and *A. sororcula* are considered major pests because of their wide distribution and generalist habits (MALAVASI; ZUCCHI; SUGAYAMA, 2000; ZUCCHI, 2000b). These fruit flies use a great variety of fleshy fruits for oviposition, hence larval stages occurred in the endocarp causing mechanical damage and facilitating fungal and bacterial infections (ALUJA, 1994; DUARTE; MALAVASI, 2000) that has drastic economic consequences.

Although several pest management methods have been proposed to mitigate their agricultural impact, the applicability of these techniques in the field is limited by taxonomic uncertainties among closely related and even cryptic species such as *A. fraterculus* complex (HERNÁNDEZ-ORTIZ et al., 2012).

1.1.2. Morphological traits

Morphological data has been very useful to clarify taxonomical aspects in the genus *Anastrepha*, especially among phylogenetically distant species. However, there are several unresolved questions regarding the phylogeny and identification of species of this genus, particularly in species of the *fraterculus* group. The identification of species in this group is a challenging task because of the different structures that should be analyzed by morphometry techniques and the necessity of a highly trained taxonomist to correctly perform the task. Morphology-based classification system of this group includes analysis of color pattern of the subscutellum (thorax), color pattern and morphology of the wings and morphometry of the aculeus tip (ZUCCHI, 2000b). Although most species of this group can be correctly classified based on these traits, characters, several species are harder to tell apart, so much so that individuals of *A. obliqua*, *A. sororcula* and *A. fraterculus* (*sensu latu*) showed overlapping measurements even in the aculeus tip, a key taxonomic trait (PERRE et al., 2014). Morphology-based identification is even more challenging because *fraterculus* group includes not only closely related species, but also cryptic species, such as the *A. fraterculus* cryptic complex (HERNÁNDEZ-ORTIZ et al., 2012). Even though some authors have proposed morphometry of larvae and pupae as a possible solution (BARBOSA; TOVAR; BRESSAN-NASCIMENTO, 2005; FRÍAS; SELIVON; HERNÁNDEZ-ORTIZ, 2006; CANAL et al., 2015), this is still an extremely complicated task. In this context, genetic molecular tools may significantly contribute to help solving these taxonomic issues.

1.1.3. Phylogeny of *Anastrepha*

The first study of the molecular phylogeny of this genus produced a weakly supported tree based on the mitochondrial gene 16S rRNA (McPheron et al., 1999). Furthermore, a phylogeny produced by the nuclear gene *period* produced a similar unresolved phylogenetic inference as 16S rRNA (Barr; Cui; McPheron, 2005). Low resolution of these two phylogenetic reconstructions of these species may probably be due to insufficient sampling, small variation of the molecular marker and the use of only one gene. Recently, Mengual et al. (2017) inferred the *Anastrepha* phylogeny using nuclear and mitochondrial genes from 146 species. Despite the notably sampling effort, they reported only seven species groups as monophyletic and in general relationships among groups failed to be strongly established.

Most published phylogenies focusing on the *fraterculus* group species showed weak branches support, such as the reconstruction based on the mitochondrial gene Cytochrome c oxidase subunit I (COI) revealed *A. fraterculus* (*s.l.*), *A. sororcula* and *A. obliqua* as non-monophyletic lineages (Smith-Caldas et al., 2001). A reanalysis of this data along with *A. suspensa* supported the monophyly of that species (Boykin et al., 2006). A phylogeographic study of *A. obliqua* based on two mitochondrial genes found large variation among populations and suggested that this species may be a cryptic species complex such as *A. fraterculus* (Ruiz-Arce et al., 2012). In addition, a phylogeny based on nuclear genes supported most of the fraterculus group recognized species as monophyletic including *A. obliqua*, which seemed to be subdivided into two lineages (Scally et al., 2016). From these studies, we can infer that the low resolution from previous phylogenies may be due to cryptic species such as *A. fraterculus* complex, recent divergence, ancestral polymorphism and hybridization among lineages.

1.1.4. Gene and species tree

With the development of molecular genetic techniques, it became easier to use different molecular markers to reconstruct phylogenies, that would not necessarily agree with one another, making it important to differentiate species trees and gene trees. The former represents the actual relationships among the analyzed species, while the latter represents the evolutionary history of ortholog genes of these species (DEGNAN; ROSENBERG, 2009). Tree topologies by individual genes may show incongruent patterns among them and compared to the species tree, which can be more accentuated in the case of gene trees inferred from closely related species and species with large effective population sizes (HELED; DRUMMOND, 2010). This gene tree discordance may be evident because gene copies from different species can be more related than intraspecies copies because of shared ancestral polymorphism segregating in these lineages (Figure 1A). Another possible source of gene and species tree incongruence is gene flow between species (Figure 1B), which would produce strong effects in phylogenies inferred by genes with uniparental inheritance such as mtDNA (HAILER et al., 2012). Because of the different processes that may be occurring, the probability to infer accurate species trees increases when more genes are sampled (PAMILO; NEI, 1988).

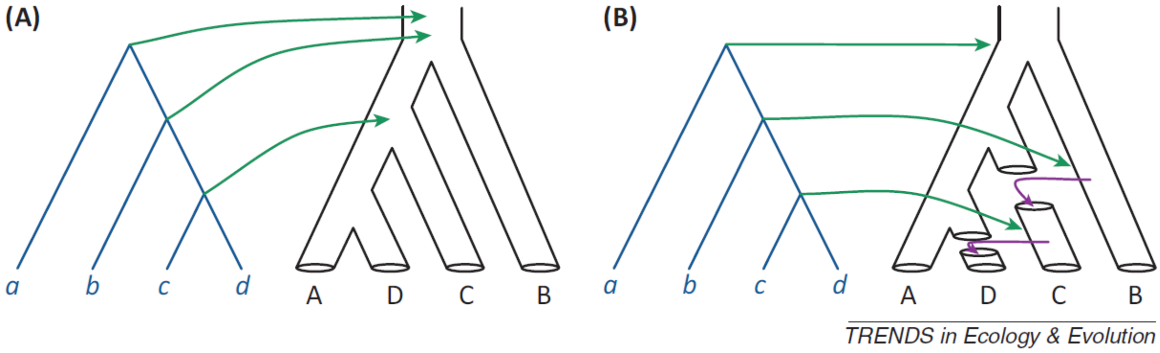


Figure 1. Sources of gene tree discordance. Gene trees are shown in solid lines and species tree in tubes. (A) Phylogenetic incongruence due to incomplete lineages sorting (ILS). (B) Phylogenetic incongruence due to horizontal gene transfer of introgression. Reprinted from Trends in Ecology & Evolution, 28, Nakhleh L., Computational approaches to species phylogeny inference and gene tree reconciliation, 723, Copyright (2013), with permission from Elsevier.

The ability to identify new genetic markers at a reasonable cost and even in species without available genomic information has increased manifold recenty with the the advent of next-generation sequencing technologies (SHOKRALLA et al., 2012; MCCORMACK et al., 2013), which have been widely used to reconstruct phylogenies of a great variety of taxa (SONG et al., 2012; HENRIQUEZ et al., 2014; ILVES; LÓPEZ-FERNÁNDEZ, 2014; KAWAHARA; BREINHOLT, 2014; GARRISON et al., 2016). However, the use of a great number of different markers does not preclude the investigation of gene trees discordance due to incomplete lineage sorting, intraspecific recombination (e.g. gene conversion, meiotic recombination) and hybridization or horizontal gene transfer (RANNALA; YANG, 2008). These types of methodological issues can be particularly important if a gene concatenation approach is applied (SONG et al., 2012). For that reason, a wide variety of approaches referred to as “multispecies coalescent methods” have been developed to correctly deal with multi gene data sets (LIU, 2008; HELED; DRUMMOND, 2010; LARGET et al., 2010; MIRARAB; WARNOW, 2015). Although, some of them may be computationally exhaustive and their application to genome-wide dataset may still be unsuitable, others can be useful for large datasets (CHIFMAN; KUBATKO, 2014; MIRARAB; WARNOW, 2015; VACHASPATI; WARNOW, 2015).

Using enough genomic data, these phylogenetic methods are robust enough to take into account incomplete lineage sorting (ILS), but hybridization may produce incongruent species tree inferences (LEACHÉ et al., 2014). In such cases, multispecies coalescent networks can evaluate both ILS and hybridization (YU et al., 2011). Each hybridization event is modelled as a reticulation in the network, which is associated with the probability that an allele have been transferred from the ancestral lineage to the recipient lineage (inheritance probability) (YU; DEGNAN; NAKHLEH, 2012). Several statistical frameworks have been proposed to infer species networks such as maximum likelihood, Bayesian inference, maximum parsimony and maximum pseudo-likelihood (YU; BARNETT; NAKHLEH, 2013; YU et al., 2014; YU; NAKHLEH, 2015; WEN; YU; NAKHLEH, 2016), though, only the two latter methods are suitable to evaluate genome-scale dataset in a reasonable time. In this thesis, we aimed to infer both a species-tree and a species-network based on multispecies coalescent methods using a large genetic dataset from *Anastrepha* species, which helped us understand phylogenetic relationships and patterns of introgression produced by rapid radiation of this genus.

1.1.5. Speciation model

Mechanisms leading to the formation of new species have been intriguing evolutionary biologists since before the publication of the book "On the Origin of Species" by the naturalist Charles Darwin (1859). In the book, Darwin emphasizes the role of natural selection in the speciation process through adaptation to the environment, though also indicating a role for sexual selection. However, for several decades most studies have focused on understanding allopatric speciation using neutral genetic markers (AVISE, 2000; HICKERSON et al., 2010). Lately, some studies have found evidence of sympatric speciation in animal species (BOLNICK; FITZPATRICK, 2007; MALLET, 2007).

An important question to be addressed is how populations that exchange genetic material accumulate enough differentiation to generate isolation and, with time, diverge into new species. The answer could be in the role of selection in cases of recent speciation with evidence of gene flow (WU; TING, 2004), via two not mutually exclusive evolutionary mechanisms: divergent selection due to ecological niche differences and sexual selection (MARIE CURIE SPECIATION NETWORK, 2012; ARNEGARD et al., 2014). Ecological differences can produce distinct selective pressures in several genomic regions, producing divergent genetic variants between new lineages (RUNDLE; NOSIL, 2005). These genetic regions evolving under divergent selection are referred to as "islands of divergence" (WU, 2001; MICHEL et al., 2010). While these islands show an elevated level of genetic differentiation, the remainder of the genome remains homogenized by gene flow. In early phase of species divergence, genetic drift would not play a crucial role in the differentiation of genomes. As time progresses, regions physically linked to the "islands of divergence" may change their gene frequencies due to genetic hitchhiking effect. Finally, this differentiation can decrease the effect of gene flow and increase the effect of genetic drift, producing divergence at the level of the whole genome (reviewed in NOSIL; FUNK; ORTIZ-BARRIENTOS, 2009). However, genomic regions with low genetic diversity due to strong selective pressures, rather than regions resistant to gene flow, can also produce picks of differentiation throughout the genome, being an alternative explanation for this outcome (CRUICKSHANK; HAHN, 2014). Some authors though question the idea that islands of divergence in the genome would be always caused by a balance between selection and gene flow, indicating that regions with low recombination can appear as differentiated between closely related taxa because of ancestral shared polymorphism and drift (NOOR; BENNETT, 2009).

This model of speciation is compatible with the patterns of differentiation found in the apple maggot *Rhagoletis pomonella* (Tephitidae) (MICHEL et al., 2010), in which two races diverged under gene flow following different host preferences (FEDER et al., 2003). This species has used hawthorn as host during reproduction, but a subpopulation of this fly shifted its host to apple, which had been introduced by North American farmers (BUSH, 1966). This ecological shift led to rapid differences in allele frequencies and partial reproductive isolation of these subpopulations, that became different races (FEDER; CHILCOTE; BUSH, 1990; FEDER; HUNT; BUSH, 1993). Although there is evidence of host preferences of *fraterculus* group species such as *A. distincta* and *A. fraterculus*, most of them such as *A. sororcula*, *A. fraterculus*, *A. zenildae* are polyphagous and have overlapping host plants (ZUCCHI, 2000a), despite having some preferences. Thus, probably this model may be limited to explain the divergence of species in this group.

On the other hand, sexual selection may play an important role in the differentiation of species in *Anastrepha*. Males from most *Anastrepha* fruit flies display lek mating behavior (ALUJA et al., 1999), in which males aggregate to attract females by displaying visual, acoustic and/or chemical signals. Furthermore, during lekking, *Anastrepha* males extrude lateral and anal pouches that enhance the dispersion of pheromones, whose dispersion is aided by rapid wing fanning (NATION, 1989). Differences in courtship behavior and pheromone composition may be key factors that led to reproductive isolation among *fraterculus* group species. Chemical studies have revealed that morphotypes of *A. fraterculus* complex display different pheromone composition (CÁCERES et al., 2009; VANÍČKOVÁ et al., 2015). These attributes should increase sexual selection in species of this group, fostering evolution in genes associated with these processes. In fact, genes related to reproduction such as courtship behavior have been found to evolve under natural selection, which may indicate sexual selection acting on these genes (SOBRINHO; DE BRITO, 2010; 2012). Large-scale genetic data analysis provides a tool for the search of sex-biased expressed genes evolving under positive selection, which may be involved in the differentiation of *Anastrepha* species.

1.2. Objectives

The main goal of this study is to contribute to the understanding of the evolutionary patterns involved in the rapid radiation of the diverse genus *Anastrepha* focusing on the *fraterculus* specie group. For that, the following specific objectives have been proposed:

1. Analyze male and female transcriptomes expressed in head and reproductive tissue from closely related species *A. fraterculus* and *A. obliqua* to investigate tissue and sex expression patterns.
2. Determine evolutionary forces (adaptive or non-adaptive) related to sex-biased expressed genes in the transcriptomes of *A. fraterculus* and *A. obliqua*.

3. Select a set of candidate genes potentially involved in the differentiation and divergence of both species.
4. Analyze transcriptomes from some *Anastrepha* lineages to establish a robust phylogeny of the *fraterculus* species group as well as other related species groups.
5. Evaluate sources of gene tree discordance such as incomplete lineage sorting and hybridization that could be confounding phylogenetic inferences in *Anastrepha*.

1.3. References

ALUJA, M. Bionomics and management of *Anastrepha*. **Annual review of entomology,** v. 39, n. 1, p. 155-178, 1994.

ALUJA, M.; NORRBOM, A. **Fruit flies (Tephritidae): phylogeny and evolution of behavior**. Crc Press, 1999.

ALUJA, M. et al. Behavior of flies in the genus *Anastrepha* (Trypetinae: Toxotrypanini). In: ALUJA, M. e NORRBOM, A. L. (Ed.). **Fruit flies (Tephritidae): phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 15, p.375-406.

ARNEGARD, M. E. et al. Genetics of ecological divergence during speciation. **Nature,** v. 511, p. 307, 2014.

ASHBURNER, M.; GOLIC, K.; HAWLEY, R. **Drosophila: A Laboratory Handbook**. New York: Cold Spring Harbor Laboratory Press, 2005.

AVISE, J. C. **Phylogeography: the history and formation of species**. USA: Harvard University Press, 2000.

BARBOSA, M. C.; TOVAR, F. J.; BRESSAN-NASCIMENTO, S. Morphological and molecular characterization of three species of *Anastrepha* Schiner and of *Ceratitis capitata* (Wiedemann) (Diptera: Tephritidae). **Neotropical Entomology,** v. 34, p. 917-925, 2005.

BARR, N. B.; CUI, L.; MCPHERON, B. A. Molecular systematics of nuclear gene *period* in genus *Anastrepha* (Tephritidae). **Annals of the Entomological Society of America,** v. 98, n. 2, p. 173-180, 2005.

BOLNICK, D. I.; FITZPATRICK, B. M. Sympatric speciation: Models and empirical evidence. **Annual Review of Ecology, Evolution, and Systematics,** v. 38, n. 1, p. 459-487, 2007.

BOYKIN, L. M. et al. Analysis of host preference and geographical distribution of *Anastrepha suspensa* (Diptera: Tephritidae) using phylogenetic analyses of mitochondrial cytochrome oxidase I DNA sequence data. **Bulletin of Entomological Research,** v. 96, n. 05, p. 457-469, 2006.

BUSH, G. L. The taxonomy, cytology, and evolution of the genus *Rhagoletis* in North America (Diptera, Tephritidae). **Bulletin of the Museum of Comparative Zoology at Harvard College.,** v. 134, p. 431-562, 1966.

CÁCERES, C. et al. Incipient speciation revealed in *Anastrepha fraterculus* (Diptera; Tephritidae) by studies on mating compatibility, sex pheromones, hybridization, and cytology. **Biological Journal of the Linnean Society,** v. 97, n. 1, p. 152-165, 2009.

CANAL, N. A. et al. Morphometric study of third-instar larvae from five morphotypes of the *Anastrepha fraterculus* cryptic species complex (Diptera, Tephritidae). **ZooKeys,** v. 540, 2015.

CHIFMAN, J.; KUBATKO, L. Quartet inference from SNP data under the coalescent model. **Bioinformatics,** v. 30, n. 23, p. 3317-3324, 2014.

CRUICKSHANK, T. E.; HAHN, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. **Molecular Ecology,** v. 23, n. 13, p. 3133-3157, 2014.

DEGNAN, J. H.; ROSENBERG, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. **Trends in Ecology & Evolution,** v. 24, n. 6, p. 332-340, 2009.

DUARTE, A. L.; MALAVASI, A. Tratamentos quarentenários. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000. p.187-200.

FEDER, J. L. et al. Allopatric genetic origins for sympatric host-plant shifts and race formation in *Rhagoletis*. **Proceedings of the National Academy of Sciences,** v. 100, n. 18, p. 10314-10319, 2003.

FEDER, J. L.; CHILCOTE, C. A.; BUSH, G. L. The geographic pattern of genetic differentiation between host associated populations of *Rhagoletis pomonella* (Diptera: Tephritidae) in the eastern United States and Canada. **Evolution,** v. 44, n. 3, p. 570-594, 1990.

FEDER, J. L.; HUNT, T. A.; BUSH, L. The effects of climate, host plant phenology and host fidelity on the genetics of apple and hawthorn infesting races of *Rhagoletis pomonella*. **Entomologia Experimentalis et Applicata,** v. 69, n. 2, p. 117-135, 1993.

FRÍAS, D.; SELIVON, D.; HERNÁNDEZ-ORTIZ, V. Taxonomy of immature stages: new morphological characters for Tephritidae larvae identification. Fruit Flies of Economic Importance: From Basic to Applied Knowledge 2006, Salvador, Brazil. p.29-44.

GARRISON, N. L. et al. Spider phylogenomics: untangling the Spider Tree of Life. **PeerJ,** v. 4, p. e1719, 2016.

HAILER, F. et al. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. **Science,** v. 336, n. 6079, p. 344-347, 2012.

HELED, J.; DRUMMOND, A. J. Bayesian inference of species trees from multilocus data. **Molecular Biology and Evolution,** v. 27, n. 3, p. 570-580, 2010.

HENRIQUEZ, C. L. et al. Phylogenomics of the plant family Araceae. **Molecular Phylogenetics and Evolution,** v. 75, n. Supplement C, p. 91-102, 2014.

HERNÁNDEZ-ORTIZ, V. et al. Cryptic species of the *Anastrepha fraterculus* complex (Diptera: Tephritidae): a multivariate approach for the recognition of South American morphotypes. **Annals of the Entomological Society of America,** v. 105, n. 2, p. 305-318, 2012.

HICKERSON, M. J. et al. Phylogeography's past, present, and future: 10 years after Avise, 2000. **Molecular Phylogenetics and Evolution,** v. 54, n. 1, p. 291-301, 2010.

ILVES, K. L.; LÓPEZ-FERNÁNDEZ, H. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. **Molecular Ecology Resources,** v. 14, n. 4, p. 802-811, 2014.

KAWAHARA, A. Y.; BREINHOLT, J. W. Phylogenomics provides strong evidence for relationships of butterflies and moths. **Proceedings of the Royal Society B: Biological Sciences,** v. 281, n. 1788, 2014.

LARGET, B. R. et al. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. **Bioinformatics,** v. 26, n. 22, p. 2910-2911, 2010.

LEACHÉ, A. D. et al. The influence of gene flow on species tree estimation: A simulation study. **Systematic Biology,** v. 63, n. 1, p. 17-30, 2014.

LIU, L. BEST: Bayesian estimation of species trees under the coalescent model. **Bioinformatics,** v. 24, n. 21, p. 2542-2543, 2008.

MALAVASI, A.; ZUCCHI, R. A.; SUGAYAMA, R. L. Biogeografia. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000. p.93-98.

MALLET, J. Hybrid speciation. **Nature,** v. 446, n. 7133, p. 279-283, 2007.

MARIE CURIE SPECIATION NETWORK. What do we need to know about speciation? **Trends in Ecology & Evolution,** v. 27, n. 1, p. 27-39, 2012.

MCCORMACK, J. E. et al. Applications of next-generation sequencing to phylogeography and phylogenetics. **Molecular Phylogenetics and Evolution,** v. 66, n. 2, p. 526-538, 2013.

MCPHERON, B. A. et al. Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotrypanini) based upon 16S rRNA mitochondrial DNA. In: ALUJA, M. e NORRBOM, A. (Ed.). **Fruit flies (Tephritidae): Phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 13, p.343-362.

MENGUAL, X. et al. Phylogenetic relationships of the tribe Toxotrypanini (Diptera: Tephritidae) based on molecular characters. **Molecular Phylogenetics and Evolution,** v. 113, p. 84-112, 2017.

MICHEL, A. P. et al. Widespread genomic divergence during sympatric speciation. **Proceedings of the National Academy of Sciences,** v. 107, n. 21, p. 9724-9729, 2010.

MIRARAB, S.; WARNOW, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. **Bioinformatics,** v. 31, n. 12, p. i44-i52, 2015.

NAKHLEH, L. Computational approaches to species phylogeny inference and gene tree reconciliation. **Trends in Ecology & Evolution,** v. 28, n. 12, p. 719-728, 2013.

NATION, J. The role of pheromones in the mating system of *Anastrepha* fruit flies. In: ROBINSON, A. e HOOPER, G. (Ed.). **Fruit flies, their biology, natural enemies and control**. Amsterdam (Holanda): University of Amsterdam, 1989. p.189–205.

NOOR, M. A. F.; BENNETT, S. M. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. **Heredity,** v. 103, n. 6, p. 439-444, 2009.

NORRBOM, A. **Updates to biosystematic database of world Diptera for Tephritidae through 1999.** Washington, DC: US Dep. Agric: Diptera Data Dissemination Disk, CD-ROM 2004.

NORRBOM, A. L. et al. ***Anastrepha* and *Toxotrypana*: descriptions, illustrations, and interactive keys**. Version: 28th September 2013 2012 onwards.

NORRBOM, A. L.; ZUCCHI, R. A.; HERNÁNDEZ-ORTIZ, V. Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotrypanini) based on morphology. In: ALUJA, M. e NORRBOM, A. L. (Ed.). **Fruit flies (Tephritidae): phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 12, p.299-342.

NOSIL, P.; FUNK, D. J.; ORTIZ-BARRIENTOS, D. Divergent selection and heterogeneous genomic divergence. **Molecular Ecology,** v. 18, n. 3, p. 375-402, 2009.

PAMILO, P.; NEI, M. Relationships between gene trees and species trees. **Molecular Biology and Evolution,** v. 5, n. 5, p. 568-583, 1988.

PERRE, P. et al. Morphometric differentiation of fruit fly pest species of the *Anastrepha fraterculus* group (Diptera: Tephritidae). **Annals of the Entomological Society of America,** v. 107, n. 2, p. 490-495, 2014.

RANNALA, B.; YANG, Z. Phylogenetic inference using whole genomes. **Annual Review of Genomics and Human Genetics,** v. 9, n. 1, p. 217-231, 2008.

RUIZ-ARCE, R. et al. Phylogeography of *Anastrepha obliqua* Inferred with mtDNA Sequencing. **Journal of Economic Entomology,** v. 105, n. 6, p. 2147-2160, 2012.

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

SCALLY, M. et al. Resolution of inter and intra-species relationships of the West Indian fruit fly *Anastrepha obliqua*. **Molecular Phylogenetics and Evolution,** v. 101, p. 286-293, 2016.

SHOKRALLA, S. et al. Next-generation sequencing technologies for environmental DNA research. **Molecular Ecology,** v. 21, n. 8, p. 1794-1805, 2012.

SMITH-CALDAS, M. R. B. et al. Phylogenetic relationships among species of the *fraterculus* group (*Anastrepha*: Diptera: Tephritidae) inferred from DNA sequences of mitochondrial cytochrome oxidase I. **Neotropical Entomology,** v. 30, n. 4, p. 565-573, 2001.

SOBRINHO, I. S.; DE BRITO, R. A. Evidence for positive selection in the gene *fruitless* in *Anastrepha* fruit flies. **BMC Evolutionary Biology,** v. 10, n. 1, p. 293, 2010.

______. Positive and purifying selection influence the evolution of *doublesex* in the *Anastrepha fraterculus* species group. **PLoS ONE,** v. 7, n. 3, p. e33446-e33446, 2012.

SONG, S. et al. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. **Proceedings of the National Academy of Sciences,** v. 109, n. 37, p. 14942-14947, 2012.

VACHASPATI, P.; WARNOW, T. ASTRID: Accurate Species TRees from Internode Distances. **BMC Genomics,** v. 16, n. 10, p. S3, 2015.

VANÍČKOVÁ, L. et al. Current knowledge of the species complex *Anastrepha fraterculus* (Diptera, Tephritidae) in Brazil. **ZooKeys,** v. 540, 2015.

WEN, D.; YU, Y.; NAKHLEH, L. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. **PLOS Genetics,** v. 12, n. 5, p. e1006006, 2016.

WU, C.-I. The genic view of the process of speciation. **Journal of Evolutionary Biology,** v. 14, n. 6, p. 851-865, 2001.

WU, C.-I.; TING, C.-T. Genes and speciation. **Nature Reviews Genetics,** v. 5, n. 2, p. 114-122, 2004.

YU, Y.; BARNETT, R. M.; NAKHLEH, L. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. **Systematic Biology,** v. 62, n. 5, p. 738-751, 2013.

YU, Y.; DEGNAN, J. H.; NAKHLEH, L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. **PLOS Genetics,** v. 8, n. 4, p. e1002660, 2012.

YU, Y. et al. Maximum likelihood inference of reticulate evolutionary histories. **Proceedings of the National Academy of Sciences,** v. 111, n. 46, p. 16448-16453, 2014.

YU, Y.; NAKHLEH, L. A maximum pseudo-likelihood approach for phylogenetic networks. **BMC Genomics,** v. 16, n. 10, p. S10, 2015.

YU, Y. et al. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. **Systematic Biology,** v. 60, n. 2, p. 138-149, 2011.

ZUCCHI, R. A. Espécies de *Anastrepha*, sinonímias, plantas hospedeiras e parasitóides. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000a. p.41-48.

______. Taxonomia. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000b. p.1-24.

CHAPTER II

EVIDENCE OF ADAPTIVE EVOLUTION AND RELAXED CONSTRAINTS IN SEX-BIASED GENES OF SOUTH AMERICAN AND WEST INDIES FRUIT FLIES (DIPTERA: TEPHRITIDAE)

This chapter was accepted for publication in the journal Genome Biology and Evolution.

CHAPTER II – EVIDENCE OF ADAPTIVE EVOLUTION AND RELAXED CONSTRAINTS IN SEX-BIASED GENES OF SOUTH AMERICAN AND WEST INDIES FRUIT FLIES (DIPTERA: TEPHRITIDAE)

2.1. Abstract

Several studies have demonstrated that genes differentially expressed between sexes (sex-biased genes) tend to evolve faster than unbiased genes, particularly in males. The reason for this accelerated evolution is not clear, but several explanations have involved adaptive and non-adaptive mechanisms. Furthermore, the differences of sex-biased expression patterns of closely related species are also little explored out of *Drosophila*. To address the evolutionary processes involved with sex-biased expression in species with incipient differentiation, we analyzed male and female transcriptomes of *Anastrepha fraterculus* and *A. obliqua*, a pair of species that have diverged recently, likely in the presence of gene flow. Using this data, we inferred differentiation indexes, evolutionary rates and tested for signal of selection in thousands of genes expressed in their head and reproductive transcriptomes. Our results indicate that sex-biased and reproductive-biased genes evolve faster than unbiased genes in both species, which is due both, to adaptive pressure to some genes, as well as relaxed constraints to others. Furthermore, among some of the male-biased genes evolving under positive selection, we identified some related to sexual functions such as courtship behavior and fertility. These findings suggest that sex-biased genes may have played important roles in the establishment of reproductive isolation between these species, due to a combination of selection and drift, and unveil a plethora of genetic markers useful for more studies in these species and their differentiation.

2.2. Resumo

Vários estudos têm demonstrado que genes com expressão diferencial entre sexos apresentam uma tendência para evoluir mais rápido que genes com expressão não enviesada entre sexos, fenômeno mais comum em genes com expressão enviesada em machos. A razão desta evolução acelerada não está clara, porém várias possíveis explicações têm sido estabelecidas evocando mecanismos adaptativos e não adaptativos. Além disso, as diferenças nos padrões de expressão diferencial entre sexos de espécies proximamente relacionadas têm sido pouco estudadas em espécies fora do gênero *Drosophila*. Para tal, foram analisados transcriptomas de machos e fêmeas das espécies de recente divergência com fluxo gênico *Anastrepha fraterculus* e *A. obliqua*, o que permitiu avaliar os processos evolulivos envolvidos na expressão diferencial entre sexos em espécies com incipiente especiação. Usando estes dados, foram inferidos índices de diferenciação, taxas de evolução e testes de seleção positiva de milhares de genes expressos em transcriptomas de tecidos cefálicos e reprodutivos de estas espécies. Os resultados indicaram que os genes com expressão diferencial entre sexos e expressão enviesada para o tecido reprodutivo evoluem a taxas mais rápidas em ambas as espécies analisadas, o que se deve a pressão seletiva em alguns genes e seleção relaxada em outros. Além disso, alguns dos genes com expressão enviesada em machos e com sinais de ter evoluído sob seleção positiva estão envolvidos com funções sexuais como comportamento durante o acasalamento e fertilidade. Estes achados sugerem que genes com expressão diferencial entre sexos poderiam ter papeis importantes durante o estabelecimento de barreiras de isolamento reprodutivo nestas espécies, pela combinação de seleção e deriva. Este trabalho também revelou uma infinidade de marcadores genéticos promissores para estudos de diferenciação de estas espécies.

2.3. Introduction

Understanding the evolutionary mechanisms underlying sexual dimorphism has been a very challenging task. In this regard, an important question is how two individuals of different sexes in a species may have conspicuous sexual variation, even when both sexes share practically the same genome. Transcriptome studies indicate that most morphological sex differences are caused by divergent patterns of gene expression between sexes (ELLEGREN; PARSCH, 2007). These are referred to as sex-biased genes, which have consistently shown rapid sequence evolutionary rates across taxa (MANK et al., 2007; MEISEL, 2011; HUYLMANS et al., 2016; YANG; ZHANG; HE, 2016; PAPA et al., 2017). In *Drosophila*, male-biased expressed genes evolve particularly fast (ELLEGREN; PARSCH, 2007), which is mainly caused by adaptive evolution (PRÖSCHEL; ZHANG; PARSCH, 2006).

Potential explanations for such phenomenon involve sperm competition, sexual selection and/or sexual conflict (SWANSON; VACQUIER, 2002). If this hypothesis is at least partially true, some of the products of these genes might elicit pre- or post-mating barriers which may, ultimately, play important roles reinforcing species boundaries (SNOOK et al., 2009; GAVRILETS, 2014). Indeed, accessory gland proteins secreted in males' seminal fluid in *Drosophila* influence the females' physiology and behavior (EBERHARD; CORDERO, 1995; RAM; WOLFNER, 2007) and tend to evolve under positive selection (SWANSON et al., 2001), which may reflect a role on reproductive isolation in the first stages of speciation. Furthermore, several female reproductive proteins from *Drosophila* have also been reported to evolve under positive selection (SWANSON et al., 2004; PANHUIS; SWANSON, 2006). Female proteins that interact with rapidly evolving male proteins may evolve faster because of co-evolution (HAERTY et al., 2007). In addition, proteins in the external layer of the eggshell (chorion) have been reported to evolve adaptively due to possible role in the sperm-egg and/or egg-environment interactions (JAGADEESHAN; SINGH, 2007).

Despite the evidence of the contribution of positive selection on sex-biased genes in *Drosophila*, there are alternative evolutionary explanations for rapid evolution on such genes. Studies have demonstrated that selection is weakened when trait (or gene) expression is limited to a fraction of individuals such as sex-biased genes, resulting in an increased segregation of slightly deleterious variation, which can reach fixation by genetic drift (VAN DYKEN; WADE, 2010; PURANDARE et al., 2014). As a consequence, not only polymorphism levels are increased on such genes, but also divergence rates. In fact, it has been demonstrated that relaxed constraints, genetic drift or/and an increased segregation of slightly deleterious variation have an important impact on the evolution of male-specific genes (GERSHONI; PIETROKOVSKI, 2014; HARRISON et al., 2015). Furthermore, sex-biased genes tend to have narrower expression pattern than unbiased genes, that may imply less pleiotropy and functional constraints, showing faster evolution because of relaxed purifying selection (MANK et al., 2008).

Such as in reproductive tissues, other tissues may also express reproductive-related proteins. Reproductive behavior is mainly controlled by sex pheromones (HOWARD; BLOMQUIST, 2005). Pheromones and other environmental olfactory cues are perceived as taste and olfactory stimuli and then processed by the chemosensory system in organs located mainly in the head (such as antennae) (KOHL; HUOVIALA; JEFFERIS, 2015). Among the genes involved in this process, there are sets of gene families that encode for proteins involved in ligand-binding (odorant binding proteins and chemosensory proteins) and receptor functions (odorant receptors, gustatory receptors, ionotropic receptors and sensory neuron membrane proteins) (JIN; HA; SMITH, 2008; SÁNCHEZ-GRACIA et al., 2011). The molecular evolution of these protein families has been widely studied in insects and has revealed that several of these genes evolve under positive selection under a birth-and-death process that leads to a rapid gene turnover (SANCHEZ-GRACIA; VIEIRA; ROZAS, 2009; BRAND et al., 2015; CAMPANINI; DE BRITO, 2016).

Here we investigate genes expressed in reproductive and head tissues of two closely related species, South American fruit flies (*Anastrepha fraterculus*) and West Indies fruit files (*A. obliqua*), which belong to the *fraterculus* group (NORRBOM; ZUCCHI; HERNÁNDEZ-ORTIZ, 1999). Taxonomic identification of some species within this group based only on morphology is difficult due to overlapping variation even in the aculeus, which is one of the key traits in the systematics of this group (ZUCCHI, 2000; PERRE et al., 2016). Molecular phylogeny of the *fraterculus* group based on the mitochondrial gene COI showed polyphyly for these two species (SMITH-CALDAS et al., 2001). However, phylogenetic analyses using nuclear *loci* revealed that *A. obliqua* is a monophyletic lineage (SCALLY et al., 2016) and not as closely related to *A. fraterculus* as other species in the group, though there is evidence of historical introgression between these lineages (SCALLY et al., 2016; DÍAZ et al., 2017, submitted). Furthermore, these species may produce viable hybrids in laboratory with descendants of some combinations obeying Haldane´s rule (DOS SANTOS; URAMOTO; MATIOLI, 2001), and since they are found in sympatry in several regions, it is possible that current introgression may still occur in nature. Therefore, it is possible that *A. fraterculus* and *A. obliqua* have diverged recently while retaining some gene flow, emphasizing the importance of identifying genomic regions that responded to selection and may have had a leading role on their differentiation as has been proposed for other organisms with similar speciation patterns (FEDER; EGAN; NOSIL, 2012).

We generated transcriptomes of reproductive tissues from *A. fraterculus* and *A. obliqua* and compared with RNA-seq data produced from head tissues of the same populations and species (REZENDE et al., 2016). In this study, we estimated differentiation indexes and evolutionary rates from pairwise comparisons between both species and among seven species of Tephritidae. Besides, we test for signals of natural selection and relaxed constraints. These results enabled us to identify which tissues, reproductive or cephalic, and sex, would show genes with higher evolutionary rates and whether this is due to positive selection or non-adaptive evolution. Answers to these questions contribute not only to the understanding of the evolutionary mechanisms affecting sex-biased genes, but also may offer clues to the differentiation process influencing these fruit flies even in the presence of gene flow.

2.4. Material and Methods

2.4.1. Sampling and laboratory procedures

Individuals of *A. fraterculus* were collected from the field from guava (Myrtaceae) fruits (22° 01′ 03″ S, 47° 53′ 27″ W) and *A. obliqua* from jocote (Anacardiaceae) fruits (16° 41′ 58″ S, 49° 16′ 35″ W). These populations were maintained in laboratory under the following controlled conditions: 26 ± 1°C of temperature, 60–90% of humidity and natural photoperiod. Reproductive tissue of virgin adult (8-12 days) males (testis, accessory glands and phallus) and females (ovaries, accessory glands, spermatheca, uterus and ovipositor) flies of *A. fraterculus* and *A. obliqua* were collected. Total RNA was extracted from a pool of reproductive tissues of five individuals following the protocol proposed by Chomczynski e Mackey (1995). After extraction, each sample was formed by an equimolar mix of two pools, totaling samples from 10 individuals in every mix. For each profile (species, sex and tissue), it was prepared a biological replicate totalizing eight samples. RNA-seq libraries were constructed using the TruSeq® RNA Sample Preparation Kit (Illumina®) protocol according to the manufacturer's instructions. Libraries of 2 x 100 bp paired-end reads were sequenced on Illumina HiSeq^{TM}2000 at the Laboratory of Functional Genomics Applied to Agriculture and Agri-energy, ESALQ-USP, Brazil.

2.4.2. Cleaning and assembly

Reads obtained from sequencing of reproductive tissues as well as published transcriptomes from head tissues of this pair of species (REZENDE et al., 2016) were trimmed using the program Trimmomatic v.0.33 (BOLGER; LOHSE; USADEL, 2014), setting the parameters LEADING:5 TRAILING:5 SLIDINGWINDOW:5:20 MINLEN:50. This program also searches and removes any remaining TrueSeq Illumina adapters in the reads. Unpaired reads were also discarded. After this censoring, reads from the same species were joined to produce two assemblies. Each group of reads was normalized by coverage and assembled using default parameters of Trinity v.2.4.0 (GRABHERR et al., 2011).

2.4.3. Unigene prediction and assessment of the quality of assemblies

We searched for potential coding sequences (CDSs) in all six open reading frames (ORFs) of each transcript using the software TransDecoder v.3.0.1 (http://transdecoder.github.io) following three steps. First, TransDecoder.LongOrfs was used to retain all potential CDSs coding peptides longer than 100 aa. In the second step, these peptides were submitted to the hmmscan tool included in the HMMER v.3.1b2 package (EDDY, 2011) to search for protein signatures in the Pfam-A database and BLASTP v.2.6 (CAMACHO et al., 2009) to search for similar sequences in the non-redundant database of the Genbank (nr) including only proteins of arthropods. In the third step, the program TransDecoder.Predict uses the information produced by the other steps to predict the CDSs. Redundancy of the obtained CDSs was reduced using Cd-hit-est (FU et al., 2012) with a similarity threshold of 0.99. To obtain the final set of putative unigenes, transcripts with these CDSs were filtered using the Trinity assembly information and only the isoform with the highest expression per trinity component was retained. For that, the reads from each species were mapped to the respective assembly using Bowtie2 (LANGMEAD; SALZBERG, 2012) and the abundance of each transcript was estimated by eXpress v.1.5.1 (ROBERTS; PACHTER, 2013). These steps were performed by the script align_and_estimate_abundance.pl included in the Trinity package, adding no bias correction option for the eXpress program. The completeness and redundancy level of the raw and filtered assemblies of each species was evaluated by BUSCO (Benchmarking Universal Single-Copy Orthologs) (SIMÃO et al., 2015) using the Arthropoda database as reference.

2.4.4. Functional annotation

Predicted CDSs of unigenes were compared against the GenBank nr protein database including only arthropod proteins, the *Drosophila melanogaster* protein database (r6.14) and the Eukaryotic Orthologous Groups of proteins database (KOG) (KOONIN et al., 2004) using NCBI BLASTP v.2.6 (CAMACHO et al., 2009). To all these analyses, we set an e-value threshold of 10^{-6}. We also searched for conserved protein domains using InterProScan 5.24-63.0 (JONES et al., 2014). Annotations against nr and conserved protein domains databases were submitted to Blast2GO (CONESA et al., 2005) to obtain a list of gene ontology (GO) terms associated with the annotated genes. Frequencies of GO terms at the level 2 were obtained using the program WEGO (YE et al., 2006) and their distributions were plotted using GO terms with frequencies greater than 1%.

2.4.5. Identifying sex- and tissue-biased unigenes

Expression analysis was performed by using the scripts align_and_estimate_abundance.pl, abundance_estimates_to_matrix.pl, PtR and analyze_diff_expr.pl provided by the Trinity package (GRABHERR et al., 2011). In the align_and_estimate_abundance.pl, we used Bowtie2 (LANGMEAD; SALZBERG, 2012) and eXpress v1.5.1 (ROBERTS; PACHTER, 2013) to map the reads back to each species' assemblies (set of unigenes) and to estimate abundances of each unigene, respectively. This script was run adding no bias correction option for the eXpress program and very-sensitive option to Bowtie2. The abundance_estimates_to_matrix.pl script put the abundances values estimated to each RNA-seq library in a matrix. The PtR program was used to verify the quality of the biological replicates using Spearman correlation and principal component analysis of the unigenes expression across samples measured as $\log_2$ transformed of counts per million (CPM). Differential gene expression analysis among sexes and tissues was performed in edgeR (ROBINSON; MCCARTHY; SMYTH, 2010) using the TMM (trimmed mean of M-values) normalized abundances. Expression values are shown in transcripts per million (TPM). Unigenes with fold-changes greater than four and a significance of FDR corrected p-values smaller than 0.05 was considered as differentially expressed.

2.4.6. SNP calling, differentiation indexes and the McDonald–Kreitman test (MKT)

A. fraterculus unigenes were used as reference for SNP calling. Filtered reads from each library were mapped to each assembly according to tissue using Bowtie2 (LANGMEAD; SALZBERG, 2012). Mapped reads were converted to mpileup format and filtered based on minimum mapping quality of 20 and minimum PHRED quality of 30 using mpileup tool provided by Samtools v.1.3.1 package (LI et al., 2009), minimum coverage of 20, minimum reads of 1 to call the variant and strand filter (removed variants with more than 90% supported by only reads of one strand) using the tool mpileup2snp included at VarScan v2.4.2 (KOBOLDT et al., 2012). We considered only SNPs found in at least two libraries, regardless of sex to further analysis.

Allelic frequency for each SNP was calculated as the average of the frequencies estimated by VarScan in each library. Hence, the frequency of each SNP was estimated based on 20 to 40 individuals depending on the number of samples that detected a particular SNP. We determined whether SNPs promoted synonymous or non-synonymous amino acid changes using the prediction of complete CDSs and a custom python script. Allele frequencies were used to calculate the index of interspecific differentiation (D) defined as the absolute value of the difference in allele frequencies of a SNP in *A. fraterculus* and *A. obliqua* (RENAUT; NOLTE; BERNATCHEZ, 2010; ANDRÉS et al., 2013). The statistical comparison of D distributions inferred for each type of SNP (synonymous, non-synonymous and non-coding) was performed by applying Kolmogorov-Smirnov tests. We also estimated the average D using all SNPs in each CDS ($\bar{D}_{CDS}$), using only synonymous SNPs ($\bar{D}_{S}$), and using only non-synonymous SNPs ($\bar{D}_{NS}$).

McDonald–Kreitman test (MCDONALD; KREITMAN, 1991) (MKT) was performed for each CDS by comparing the number of synonymous and non-synonymous substitutions of polymorphic and almost fixed SNPs ($D > 0.95$) using a custom python script. We removed variants with a frequency smaller than 0.05 in both species to avoid biases produced by segregation of slightly deleterious mutations (PARSCH; ZHANG; BAINES, 2009). Only genes that had a value of at least one in all four classes of SNPs were included in the analysis, and significant departures from neutrality were estimated by Fisher exact test (two-tail p-value < 0.05). CDSs with significant departure of Fisher exact test and neutrality index (NI) lower than 1 were considered to evolve under positive selection. The script also calculates the direction of selection (DoS), where a signature of positive selection is observed when DoS > 0 (STOLETZKI; EYRE-WALKER, 2011).

2.4.7. Calculating evolutionary rates

Complete CDSs were submitted to the reciprocal best hit strategy in BLASTn with an e-value threshold of 10^{-6} to obtain the potential pairs of ortholog CDSs between *A. fraterculus* and *A. obliqua*. This strategy seeks to obtain the pairs of genes that produce best hit scores in a bi-directional BLAST comparison (interchanging the CDSs of the species as query and subject). Pairs of sequences that showed a length difference greater than 5% were removed since there was a higher chance of being different isoforms or different genes with only similar domains. We aligned the DNA sequences of putatively orthologs from the two species by their amino acid translations using the MAFFT algorithm (KATOH; STANDLEY, 2013) and back converted to DNA implemented in the program TranslatorX (ABASCAL; ZARDOYA; TELFORD, 2010). Resulted alignments were submitted to KaKs_Calculator (ZHANG et al., 2006) to calculate the pairwise non-synonymous (Ka) to synonymous substitution rate (Ks) ratio of the *fraterculus* group lineage using the Model Selection framework (POSADA, 2003). To decrease the chance of poorly alignments or saturation, we removed pairs with outlier Ks values, defined as values greater than the average plus three times the standard deviation, which was 0.62. Moreover, all the alignments with Ka/Ks > 1 were visually checked.

We also calculated the evolutionary rates of ortholog genes in Tephritidae and tested for selection using a phylogenetic approach. For that, the CDSs of *Ceratitis capitata* (GCF_000347755.2) (PAPANICOLAOU et al., 2016), *Rhagoletis zephyria* (GCF_001687245.1), *Zeugodacus cucurbitae* (GCF_000806345.1) (SIM; GEIB, 2017), *Bactrocera dorsalis* (GCF_000789215.1) and *Bactrocera oleae* (GCF_001188975.1) were downloaded from Genbank. To avoid using miss-annotated and miss-assembled sequences, we removed CDSs with more than one stop codon and reduced the redundancy using Cd-hit-est (FU et al., 2012) with a similarity threshold of 0.99. The putative cluster of orthologs were predicted using reciprocal best hit strategy in BLASTn with an e-value threshold of 10^{-6} and the CDSs of *A. fraterculus* as reference. The complete clusters (seven sequences) were submitted to the filtering and alignments steps of POTION program (HONGO et al., 2015). This pipeline excludes the sequences based on relative sequence length and identity, then align the clusters, trims the alignments using trimAl v.1.2 (CAPELLA-GUTIÉRREZ; SILLA-MARTÍNEZ; GABALDÓN, 2009) and detects recombination using three methodologies (Phi, NSS and MaxChi2) implemented in PhiPack (BRUEN; PHILIPPE; BRYANT, 2006). All parameters used in POTION are available in Appendix 1. The maximum likelihood phylogenies were inferred for each remained complete cluster of orthologs using GTRCAT model and 200 bootstrap replicates in the program RAxML v.8.2.9 (STAMATAKIS, 2014).

We used trimmed alignments and the phylogenies to estimate the global nonsynonymous/synonymous rate (dN/dS) ratio (ω) and performed the strict branch-site test implemented by CODEML included in the PAML v. 4.9 package (YANG, 2007). The ω for the Tephritidae lineage was estimated using the M0 model (model = 0). We removed clusters of orthologs with dS higher than the average dS plus three times the standard deviation (dS > 7). The ancestral branch of *A. fraterculus* and *A. obliqua* of each phylogeny was set as foreground for the branch-site test. In order to statistically test whether a gene is evolving under positive selection, we compared the likelihoods of MA (model = 2, NSsites = 2) and MA1 (model = 2 , NSsites= 2, fix_omega = 1) models using likelihood-ratio tests (LRTs) (ZHANG; NIELSEN; YANG, 2005). After the LRTs, we used the χ2 distribution to obtain p-values. We also detected variation in selection strength across the cluster of orthologs using RELAX (WERTHEIM et al., 2015). This program compares Likelihood Ratio Test of alternative and null (*k = 1*) models, where *k* is selection intensity defined as $\omega_{foreground} = \omega_{background}^{k}$. Significant comparisons with *k* > 1 and *k* < 1 indicate selection intensification and relaxation, respectively. To perform the phylogenies and selection tests in parallel, we used a custom python script which uses the script raxml_bs_wrapper.py (YANG; SMITH, 2014) and functions of the ete3 module (HUERTA-CEPAS; SERRA; BORK, 2016).

2.4.8. Comparing sexes and tissues

Since *A. fraterculus* and *A. obliqua* are phylogenetically closely related and displayed similar patterns of gene expression, we compared the patterns of sequence evolution of sex-biased genes found in one and both species using ortholog information. This approach allows the evaluation of genes with ancestral expression control and generalize the results for both species and perhaps to other related species in the *fraterculus* group as well. Besides, the genes with species-specific expression enable the analysis of recent evolutionary patterns after the change in expression pattern. Comparisons consisted to evaluate statistical differences of evolutionary patterns of the CDSs of the groups (by sex and tissues) using the Wilcoxon rank-sum test corrected by Holm approach (HOLM, 1979) and Fisher´s exact test.

2.5. Results

2.5.1. Sequencing, cleaning and assessment of the quality of the assemblies

We produced 28,808,966 x 2 reads for *A. fraterculus* and 28,020,776 x 2 reads for *A. obliqua*, from males and females, two replicates each totalizing eight RNA libraries of reproductive tissue (approximately 7M x 2 reads per library). These libraries along with the previously sequenced samples of head tissue totalized 58,551,775 x 2 reads and 55,626,431 x 2 for *A. fraterculus* and *A. obliqua*, respectively. The cleaning step removed an average of 13.94% of reads for *A. fraterculus* and 14.71% for *A. obliqua*. Summary statistics for the two assemblies produced similar N50, mean, median and length distributions (Table S1, Appendix 2). Furthermore, from 1,066 conserved Arthropoda ortholog groups, BUSCO identified 95% as complete orthologs and around 3% as fragmented orthologs in the transcriptome assemblies of *A. fraterculus* and *A. obliqua* (Table S2, Appendix 2). Moreover, roughly 50% of the complete orthologs are duplicated in the raw assembly, however, the redundancy in filtered unigenes was almost zero (approximately 0.5%).

2.5.2. Functional annotation

Around 70% of the CDSs were annotated using *D. melanogaster* protein database and over 90% were matched with a protein from the GenBank non-redundant (nr) protein including only Arthropoda entries (Table S3, Appendix 2). The comparison against the nr database showed that most frequent top hits were to Tephritidae species (Figure S1, Appendix 2). Blast2GO successfully mapped approximately 67% of the CDSs. The distributions of the level 2 GO terms of both species and tissues were similar (Figure S2, Appendix 2). KOG functional classification also showed a similar representation of the categories in reproductive tissues of the two species (Figure S3, Appendix 2).

2.5.3. SNP calling, differentiation indexes and MKT

A total of 226,827 and 140,504 intra and interspecific SNPs were identified in reproductive and head transcriptomes. We found 109,828 SNPs (79,947 of them associated with synonymous and 29,881 with non-synonymous changes) in 3,662 coding regions expressed in reproductive tissues and 63,489 SNPs in 2,602 CDSs expressed in head tissues, of which 48,288 were synonymous and 15,201 non-synonymous. SNP Frequency distribution showed that more than 50% of the SNPs have rare alleles (Figure S4, Appendix 2). We rejected the hypothesis that frequency distributions of D for synonymous, non-synonymous and non-coding SNPs were drawn from the same distribution (Kolmogorov–Smirnov test, $p < 0.01$ for each of the three pairwise comparisons). These distributions also revealed that over 6% of the total of SNPs were fixed or almost fixed between species ($D > 0.95$) and the proportion of this type of SNPs is greater in non-synonymous variants in both tissues (Figure 1 A and B). The histograms of $\overline{D}_{CDS}$, $\overline{D}_{S}$ and $\overline{D}_{NS}$ showed that there is a greater proportion of highly differentiated unigenes using non-synonymous than synonymous SNPs (Figure 1 C and D). After excluding rare alleles, we retained 906 CDSs which met the minimum requisites to perform the MKT, that is, at least one synonymous and one non-synonymous fixed SNP and one synonymous and one non-synonymous polymorphic SNP. Fifty-one CDSs showed significant statistical departure from neutrality and $NI < 1$, thus were considered evolving under positive selection.

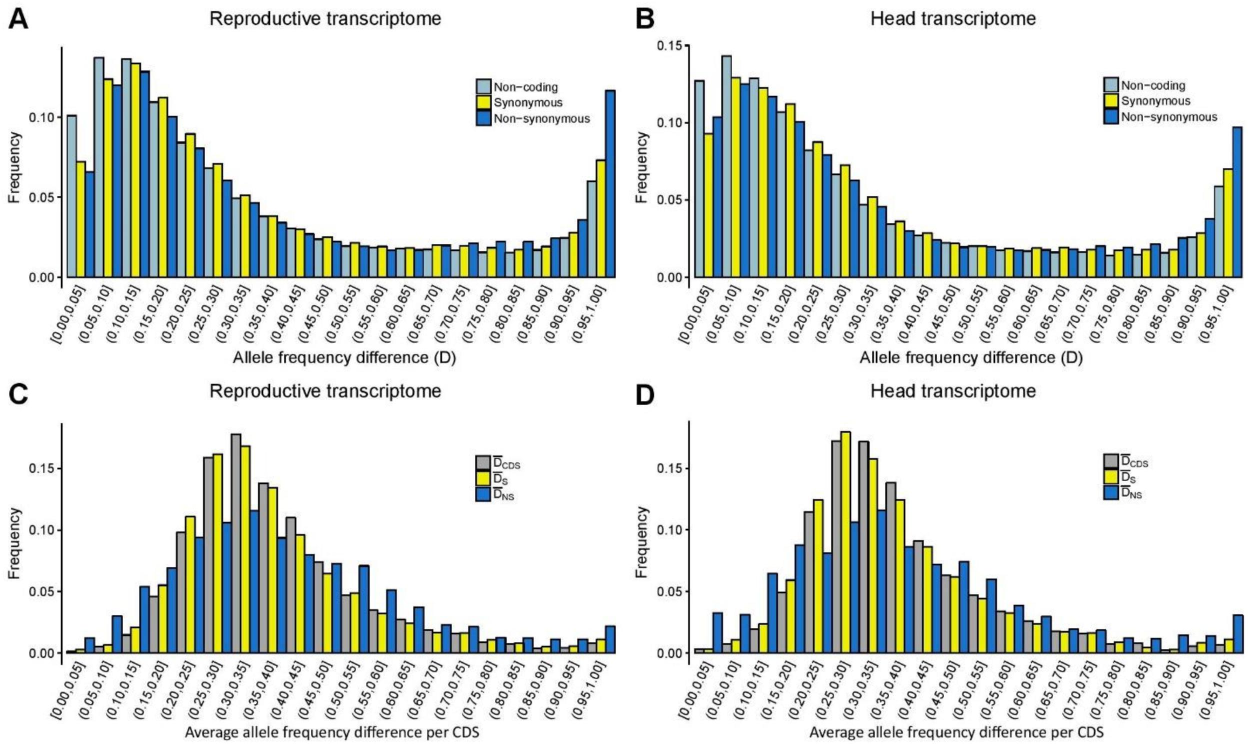


Figure 1. Frequency distributions of differentiation index. (A) and (B) Distributions of D (absolute allele frequency differences between *A. fraterculus* and *A. obliqua*) of SNPs found in reproductive and head transcriptomes, respectively. Light blue, yellow, and blue bars represent the distribution of non-coding, synonymous, and non-synonymous SNPs, respectively. (C) and (D) Distributions of average D per CDS using all SNPs ($\overline{D}_{CDS}$), using only synonymous ($\overline{D}_S$) and only non-synonymous ($\overline{D}_{NS}$) found in reproductive and head transcriptomes, respectively. Gray, yellow and blue bars represent $\overline{D}_{CDS}$, $\overline{D}_S$ and $\overline{D}_{NS,}$ respectively.

2.5.4. Patterns of gene expression across sexes and tissues

There were 12,887 and 13,605 unigenes expressed (TPM > 1) in the *A. fraterculus* transcriptome in head and reproductive tissues, respectively, with similar values also found in *A. obliqua*: 12,073 (head tissue) and 13,455 (reproductive tissue). Biological replicates are strongly correlated, with coefficients ranging from 0.96-0.98 and 0.95-0.98 for *A. fraterculus* and *A. obliqua*, respectively (Figure 2 A and B). Moreover, female and male samples of both species showed Spearman correlation coefficients higher than 0.96, establishing well-defined clusters in the principal component analysis (Figure 2 C and D). Patterns of differential gene expression across sexes and tissues were similar in both studied species (Figure 3). A total of

21.3% and 28.7% of the unigenes showed biased expression between sexes in *A. fraterculus* and *A. obliqua*, respectively (Table S4, Appendix 2). This difference is mainly affected by 6% more genes which are up-regulated in *A. obliqua* males. Interestingly, less than 1% of the genes in head transcriptomes are sex-biased in both species (Table S4 and figure S5, Appendix 2). In the comparison across tissues, approximately 27% of the genes were tissue-biased in male transcriptomes of both species (Table S5, Appendix 2). Female transcriptomes showed 22% of tissue-biased genes in *A. fraterculus* and 33% in *A. obliqua*. Besides, we also found variation in the magnitude of differential expression of biased expressed genes. Male-biased genes displayed greater fold-change average (measure by $\log_2$) than female-biased genes in both species (Wilcoxon rank sum test p-value < 0.01; Figure S6, Appendix 2). Tissue-biased genes expressed in males and females showed opposite patterns of magnitude of gene expression, while in males, the genes with greater differences in gene expression are reproductive-biased, in females, they are head-biased.

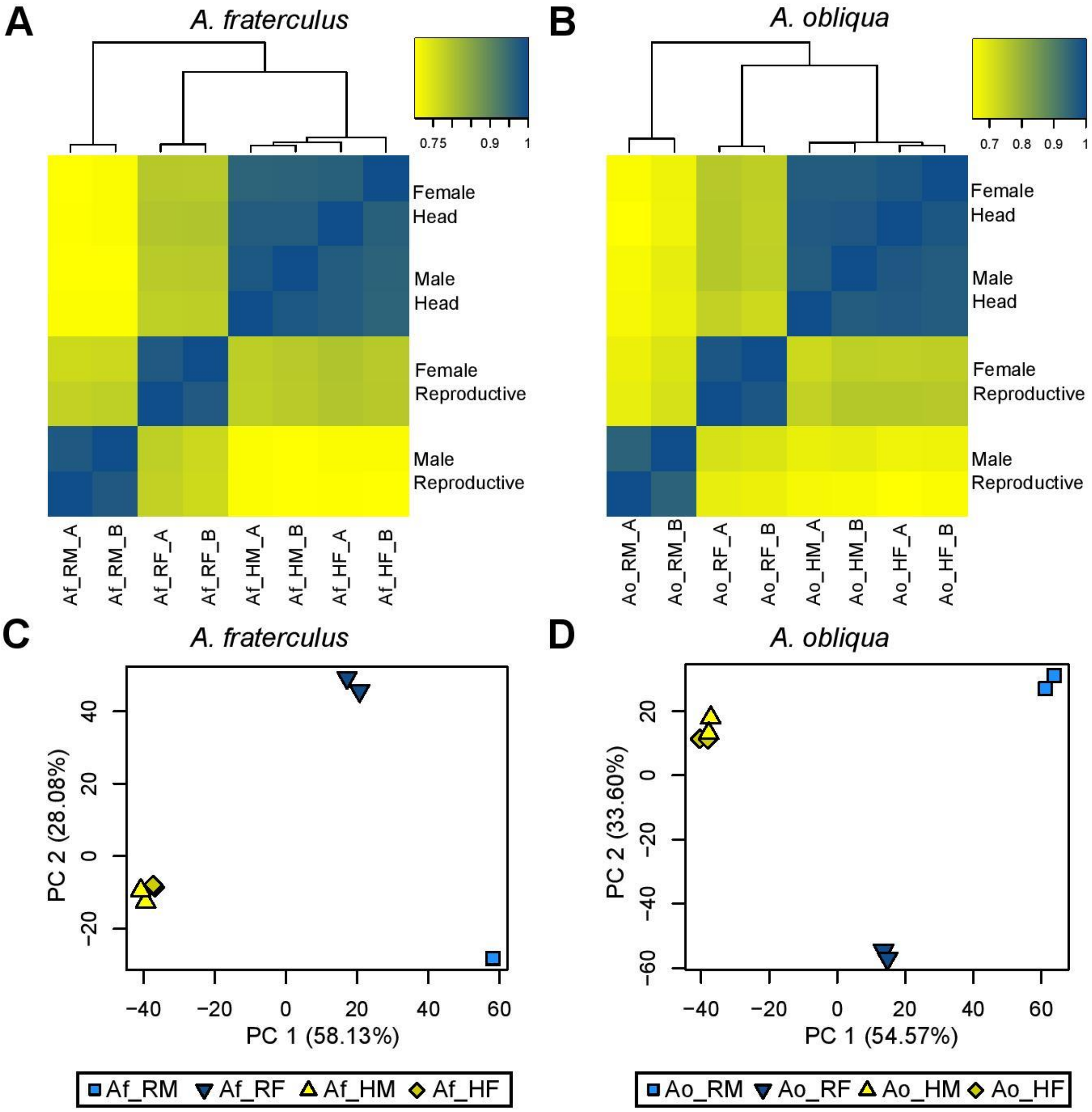


Figure 2. Analysis of expression of biological replicates. Heatmap of Spearman correlations and hierarchical cluster of samples from *A. fraterculus* (A) and *A. obliqua* (B). Principal component analysis of all samples from *A. fraterculus* (C) and *A. obliqua* (D). RM and RF: Samples from male and female reproductive transcriptomes. HM and HF: Samples from male and female head transcriptomes.

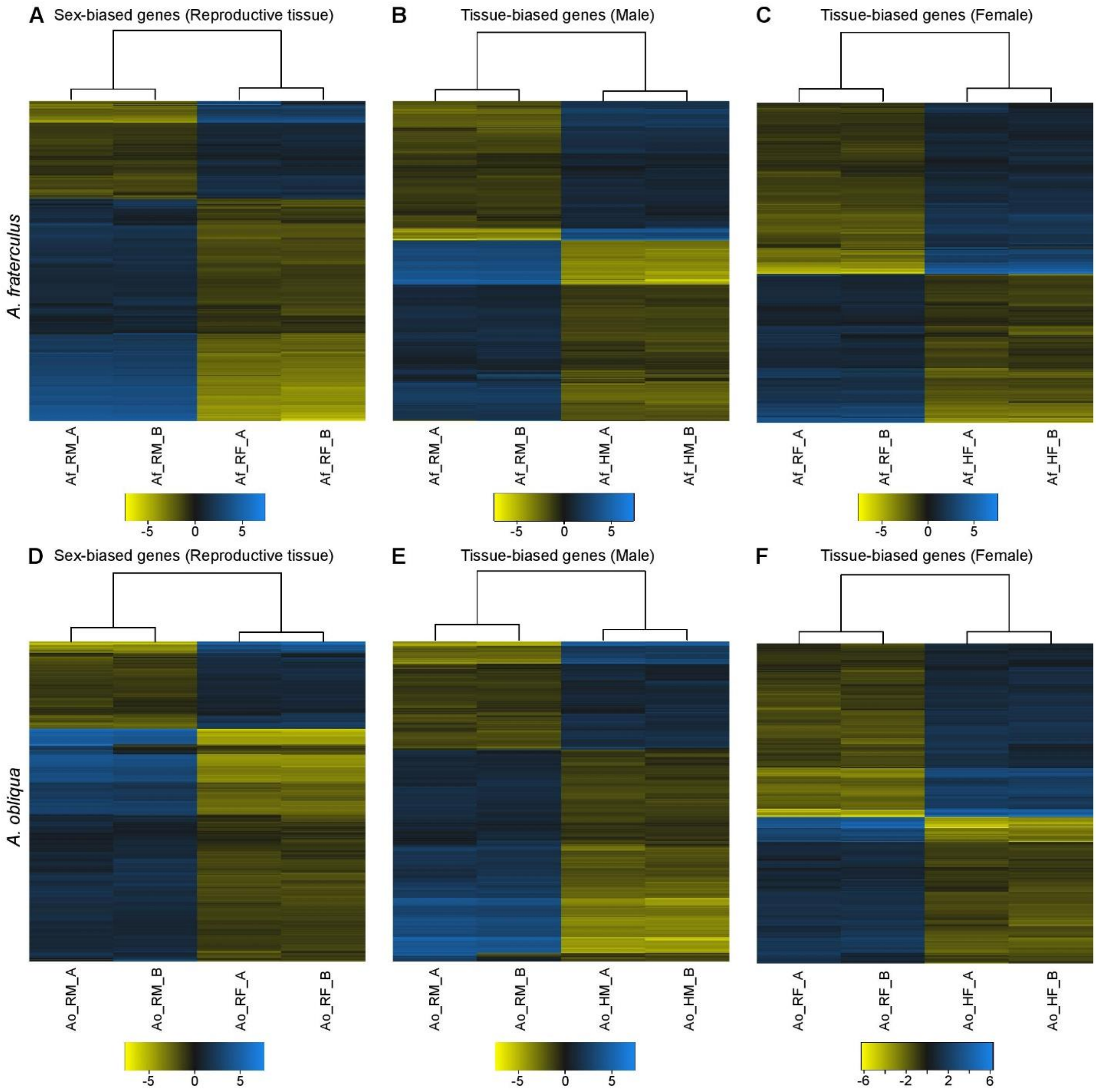


Figure 3. Heatmap and hierarchical clustering of differentially expressed genes on head and reproductive transcriptomes from *A. fraterculus* and *A. obliqua*. Differentially expressed genes between male and female reproductive transcriptomes from *A. fraterculus* (A) and *A. obliqua* (D). Differentially expressed genes between male reproductive and head transcriptomes from *A. fraterculus* (B) and *A. obliqua* (E). Differentially expressed genes between female reproductive and head transcriptomes from *A. fraterculus* (C) and *A. obliqua* (F). RM and RF: Samples from male and female reproductive transcriptomes. HM and HF: Samples from male and female head transcriptomes.

2.5.5. Evolutionary patterns of sex- and tissue-biased genes

Most comparisons between population differentiation index averages ($\bar{D}_{CDS}$, $\bar{D}_{NS}$ and $\bar{D}_{S}$) across sex and tissues failed to show significant differences between biased and unbiased genes, but the few that did, involved contrasts to non-synonymous mutations (Figure 4). Male-biased unigenes in both species displayed greater levels of differentiation than unbiased using the parameter $\overline{D}_{NS}$ (Wilcoxon rank sum test p-value < 0.05), whereas male reproductive-biased genes had the highest average $\overline{D}_{NS}$, which was also significantly different from male head-biased and unbiased genes (Wilcoxon rank sum test p-value < 0.01 in both comparisons). In contrast, female transcriptomes failed to show significant differences in any comparison (Figure S7, Appendix 2).

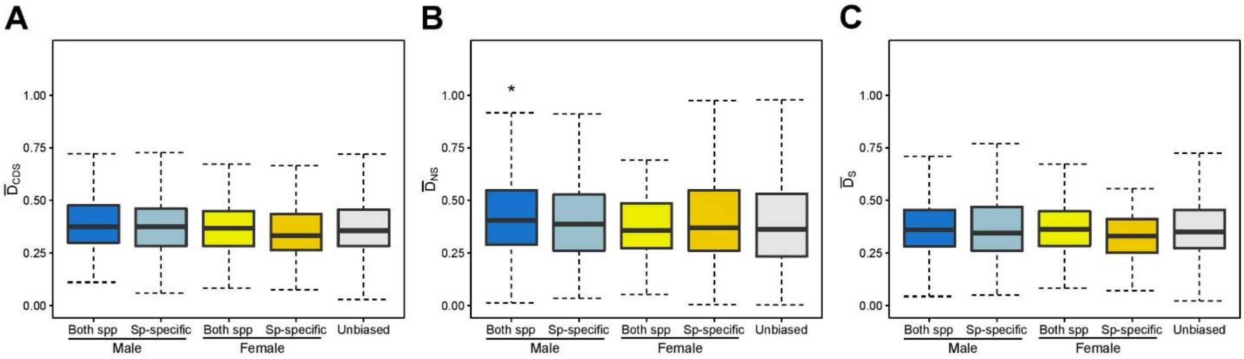


Figure 4. Boxplots of differentiation indexes measured among sex-biased and unbiased genes. Differentiation was estimated as average allele frequency differences between *A. fraterculus* and *A. obliqua* using all ($\overline{D}_{CDS}$), non-synonymous ($\overline{D}_{NS}$) and synonymous ($\overline{D}_{S}$) SNPs. Sex-biased genes are grouped in genes with the same expression pattern in both species (both spp) and biased expression detected in a particular species (sp-specific). Comparison of $\overline{D}_{CDS}$ (A), $\overline{D}_{NS}$ (B) and $\overline{D}_{S}$ (C) among male-, female-biased and unbiased genes expressed in reproductive tissues. * Holm corrected p-value of Wilcoxon rank sum test < 0.05. * just above the box indicates significant level in the comparison to unbiased genes.

Analysis of ~4,000 orthologs between *A. fraterculus* and *A. obliqua* and ~3,000 among seven Tephritidae species revealed that male-biased genes in both non-species-specific and species-specific groups have significantly higher evolutionary rates than unbiased genes (Figure 5 and Table S6, Appendix 2). Likewise, female-biased unigenes also showed significantly higher evolutionary rates than unbiased, though the comparison involving the species-specific expression genes failed to reject the null hypothesis (Figure 5). Moreover, comparisons between tissues revealed that reproductive-biased genes, be it male or female, displayed higher rates of evolution than unbiased in both species (Figure 5).

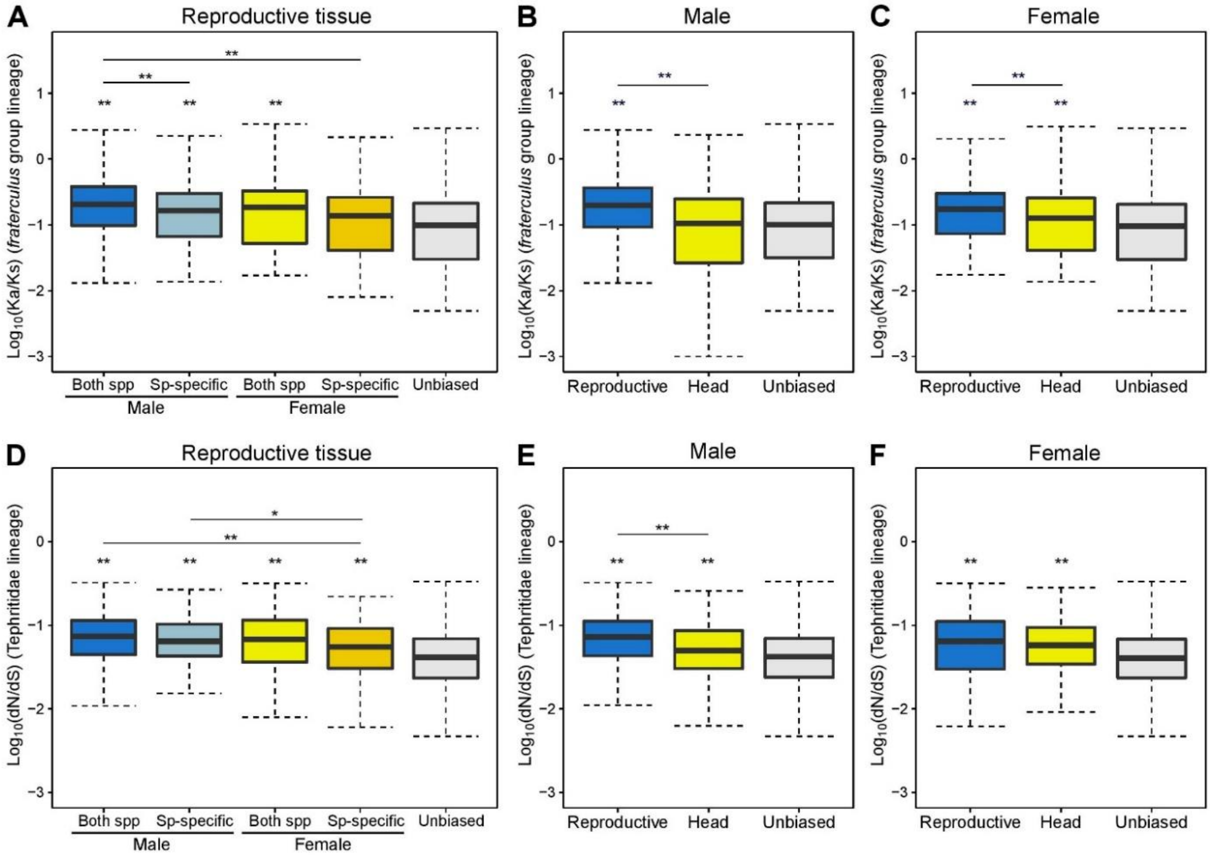


Figure 5. Evolutionary rates for sex and tissue-biased genes estimated based on pairwise comparison (*A. fraterculus* and *A. obliqua*) and seven Tephritidae species (*A. fraterculus*, *A. obliqua*, *Ceratitis capitata*, *Rhagoletis zephyria*, *Zeugodacus cucurbitae*, *Bactrocera dorsalis* and *Bactrocera oleae*). Sex-biased genes are grouped in genes with the same expression pattern in both species (both spp) and biased expression detected in a particular species (sp-specific). Boxplots of Log$_{10}$(Ka/Ks) from *A. fraterculus* and *A. obliqua* orthologs for sex-biased genes (A) and tissue-biased genes (B) and (C). Boxplots of Log$_{10}$(dN/dS) from seven Tephritidae species orthologs for sex-biased genes (D) and tissue-biased genes (E) and (F). * Holm corrected p-value of Wilcoxon rank sum test < 0.05. ** Holm corrected p-value of Wilcoxon rank sum test < 0.01. * or ** just above the box indicates significant level in comparisons with unbiased genes.

Male-biased and male reproductive-biased genes displayed significantly greater proportion of genes evolving under positive selection than unbiased as evaluated by MKT and pairwise Ka/Ks (Table 1). Moreover, these contrasts showed a positive mean DoS, suggesting adaptive evolution, even though these set of genes exhibited higher values than unbiased, this difference was not statistically significant (Figure S8, Appendix 2). These results diverge from what was found by the branch-site tests, which indicated similar proportion of genes evolving under positive selection among biased and unbiased genes. Interestingly, we noticed that 11% of the male-biased genes with signals of positive selection play important roles in *Drosophila*'s reproduction (Table 2), and it is possible that they may retain similar roles in *Anastrepha*. However, we also found a greater proportion of genes with signals of relaxed selection among male-biased and reproductive-biased (both sexes) in comparison to unbiased genes, suggesting differences in selective constraints in these contrasts (Table 1).

Table 1. Patterns of evolution for sex- and tissue-biased and unbiased genes.

	McDonald-Kreitman test		**Pairwise Ka/Ks**		**Branch-site test**		**RELAX**	
	N[a]	**p < 0.05** [b]	**N**[a]	**Ka/Ks>1**[c]	**N**[a]	**p < 0.05** [d]	**N**[a]	**p < 0.05** [e]
Reproductive								
Male-biased (both) [f]	155	**16****	488	**23***	272	17	272	**70****
Male-biased (specific) [g]	15	**4****	282	10	184	12	184	22
Female-biased (both) [f]	22	0	115	4	71	3	71	15
Female-biased (specific) [g]	36	1	195	**11***	136	6	136	**33****
Unbiased	407	15	3274	94	2481	120	2481	431
Male								
Reproductive-biased	174	**19***	527	**28****	303	14	303	**84****
Head-biased	25	0	257	9	161	6	161	26
Unbiased	402	17	3321	92	2506	127	2506	432
Female								
Reproductive-biased	40	3	354	13	140	6	140	**36***
Head-biased	37	1	234	10	217	14	217	29
Unbiased	414	17	3129	82	2406	111	2406	416

[a] N: Number of unigenes.
[b] Significant departure from non-synonymous and synonymous proportion of polymorphic and fixed SNPs using Fisher's exact test and NI < 1.
[c] Number of orthologs of *A. fraterculus* and *A. obliqua*.
[d] Number of orthologs with significant likelihood ratio tests between MA and MA1 using the *A. fraterculus* and *A. obliqua* ancestral branch as foreground.
[e] Number of orthologs with $k < 1$ (relaxed selection) and significant likelihood ratio tests between null ($k = 1$) and alternative using the *A. fraterculus* and *A. obliqua* ancestral branch as foreground.
[f] Both: Same expression pattern in of *A. fraterculus* and *A. obliqua*.
[g] Specific: Sex-biased expressed gene either in *A. fraterculus* or *A. obliqua*.
***** Fisher's exact test comparing biased with unbiased genes showing p <0.05.
****** Fisher's exact test comparing biased with unbiased genes showing p <0.01.

Table 2. Signals of positive selection in sex-biased expressed genes potentially associated with reproduction.

Annotation with *D. melanogaster* database	Expression pattern	Signal of selection	Role	Reference
Neural Lazarillo	male-biased[a]	Ka/Ks > 1	Fertility and courtship behavior	Ruiz et al. (2011)
takeout	male-biased[c]	Ka/Ks > 1	Courtship	Dauwalder et al. (2002)
CG15406	male-biased[b]	Bst[d]	Influence female´s remating	Sitnik et al. (2016)
kelch like family member 10	male-biased[c]	MKT[e]	Spermatogenesis	Arama et al. (2007)
Dynein intermediate chain at 61B	male-biased[c]	Bst[d]	Spermatogenesis	Fatima (2011)
hedgehog	male-biased[b]	Bst[d]	Male´s germ line maintenance	Zhang et al. (2013)
male fertility factor kl5	male-biased[c]	MKT[e]	Sperm motility	Carvalho; Lazzaro e Clark (2000)
lost boys	male-biased[c]	MKT[e]	Sperm motility	Yang et al. (2011)
Tektin A	male-biased[c]	MKT[e]	Sperm motility	Dorus et al. (2006)
Egg-derived tyrosine phosphatase	female-biased[c]	Bst[d]	Oogenesis and embryogenesis	Yamaguchi et al. (2005)
CG14645	female-biased[b]	Ka/Ks > 1	Courtship	Immonen e Ritchie (2012)
CG14187	female-biased[a]	Ka/Ks > 1	Chorion protein	Tootle et al. (2011)

[a] Sex-biased expressed gene only in *A. fraterculus*.
[b] Sex-biased expressed gene only in *A. obliqua*.
[c] Sex-biased expressed gene in both species.
[d] Bst: Branch-site test.
[e] MKT: McDonald-Kreitman test.

2.6. Discussion

The RNA-seq data generated high quality *de novo* assemblies of *A. fraterculus* and *A. obliqua* as evaluated by length distribution and gene content metrics. The N50 values of around 1,800 bp of these assemblies were in line with equivalent transcriptomes of other available tephritids (HSU et al., 2012; MORROW et al., 2014; SALVEMINI et al., 2014). Additionally, we found almost all conserved Arthropoda ortholog clusters in these transcriptome assemblies, suggesting a significant representation and completeness for the panel of genes expressed in head and reproductive tissues of both species. A further indication of their completeness is that most of the CDSs were successfully annotated against proteins of the *D. melanogaster* database (~70%) and the nr Genbank database (~90%). Furthermore, functional annotation using the distribution of GO and KOG categories for reproductive tissues of *A. fraterculus* and *A. obliqua* showed similar distributions, akin to what has been described for head tissues from these species (REZENDE et al., 2016).

Transcriptome data here studied enabled identification of thousands of SNPs across *A. fraterculus* and *A. obliqua*. Even though we identified hundreds of SNPs fixed, or nearly fixed, in one or the other species, the most common pattern observed for *A. fraterculus* and *A. obliqua* transcriptomes indicates that the species have diverged recently, since a great number of SNPs show little allele frequency difference across species. However, these results should be interpreted with caution, since we estimated SNP allele frequency distributions from pools of individuals of a single population per species, and they may not represent the whole diversity across the species' geographic distributions. Microsatellite analyses across *A. fraterculus* Brazilian populations showed some evidence of differentiation, but over 90% of variation is intrapopulational (MANNI et al., 2015). Furthermore, there is evidence that these species differentiated with gene flow (SCALLY et al., 2016; DÍAZ et al., 2017, submitted), which would make variation in general common to several localities, rather than isolated, even across species boundaries, patterns that, if common across the species' distribution, might indicate that the diversity distribution here inferred for *A. fraterculus* and *A. obliqua* may hold for the majority of SNPs identified.

The allele frequency distributions are consistent with a scenario where the majority of the genome would be somewhat homogenous, interspersed by highly differentiated regions (MARTIN et al., 2013), such as what was found in two recently diverged species of *Gryllus* in the presence of gene flow (ANDRÉS et al., 2013), even if there would be no selective forces involved and only drift is driving the species apart. Here, despite the reduced number of contigs with large allele frequency differences across species, we still detected at least 5% of the SNPs nearly fixed for different alleles in different species. Interestingly, the distributions of D (allele frequency difference between species) inferred from synonymous, non-synonymous and non-coding SNPs are significantly different from one another. These differences seem to be at least in part due to adaptive evolution since there is a significantly greater proportion of fixed differences across species which are associated to non-synonymous substitutions, even when contrasted with non-coding substitutions. This increased proportion of non-synonymous substitutions in fixed differences between species also holds when we consider all substitutions present in a CDS. The unigenes with high $\overline{D}_{NS}$ values may potentially be “islands of divergence”, which are genomic regions that remain differentiated between species even in the presence of gene flow due to directional selection (NOSIL; FEDER, 2012). Even though there are other reasons why genomic islands of divergence may occur (NOOR; BENNETT, 2009; CRUICKSHANK; HAHN, 2014), many which cannot be tested for the data here presented because of the lack of a reference genome, these results point that at least a portion of the divergence between *A. fraterculus* and *A. obliqua* is due to regions affected by selection. We found some male-biased expressed genes with signals of positive selection involved with male courtship and fertility (Table 2), thus these genes may be related to the establishment of pre-zygotic barriers. This observation agrees with studies on morphotypes of *A. fraterculus* complex species which have suggested that their reproductive isolation is mainly due to prezygotic barriers (RULL et al., 2013; JUÁREZ et al., 2015).

The SNP allele frequency distributions allowed us not only to identify genes potentially involved with species differences, but also to investigate general patterns of evolution for genes expressed in reproductive tissues across the two closely related species. In general, SNPs in both species showed a large proportion of rare alleles (Figure S3, Appendix 2), which might be due to demographic expansion and/or selective sweeps or weak purifying selection in particular genes (FU, 1997; FAY; WYCKOFF; WU, 2001). We consider the former to be more likely considering that this pattern seems to be widespread across several genes, and that population expansion due to the increased distribution of host fruits with agriculture has been suggested to have happened to both studied species based on coalescent simulations (DÍAZ et al., 2017, submitted).

In general, the expression profiles were similar between *A. fraterculus* and *A. obliqua* showing 20-30% sex-biased unigenes, a pattern similar to what was found in *Drosophila* species (ZHANG et al., 2007). This may be due to stability of the gene expression control machinery because of evolutionary constraints (ZHANG et al., 2007; HE et al., 2011). Data from both species consistently indicate that the majority of sex-biased genes come from reproductive tissues, which agrees with what was reported for a comparison of somatic tissues and gonad transcriptomes in *D. melanogaster* (PARISI et al., 2004). Furthermore, the ~0.5% of sex-biased genes displayed in head tissues of both species of *Anastrepha* contrasts with the ~16% differentially expressed genes in *D. melanogaster* head (CHANG et al., 2011). However, in the latter there is a large difference in the genes with sexually diverged expression between central system and peripheral tissues (GOLDMAN; ARBEITMAN, 2007), so this variation in expression pattern across head organs and structures could obscure the expression of sex-biased genes in the whole head. Besides, our results show a higher number of up-regulated genes and larger magnitudes of their fold changes in males than in females, which could be due to the existence of more male-biased genes or the differences in expression of female-biased are too small, which would require larger statistical power to detect these genes (ASSIS; ZHOU; BACHTROG, 2012). The comparison between tissues reveals that most tissue-biased expressed genes are in male reproductive tissues, possibly because most tissue-specific genes are expressed in testis, as indicated for *D. melanogaster* (MEIKLEJOHN; PRESGRAVES, 2012).

Our results suggest that male-biased genes have higher ω such as it was found for several lineages (TORGERSON; KULATHINAL; SINGH, 2002; VACQUIER; SWANSON, 2011; HARRISON et al., 2015), such as *Drosophila* species (ZHANG; HAMBUCH; PARSCH, 2004), but female-biased genes also evolve significantly faster than unbiased genes in *Anastrepha* species, which has also been described for some animal taxa such as birds (MANK et al., 2007), mosquitoes (PAPA et al., 2017; WHITTLE; EXTAVOUR, 2017) and fishes (YANG et al., 2016). Sex-biased genes also tend to evolve more rapidly, particularly their non-synonymous rates, which, along with the higher rates of fixation of non-synonymous mutations ($\bar{D}_{NS}$) found in males, suggests that these genes may have been evolving under adaptive constraints in the *Anastrepha* species here studied. However, the faster evolution of sex-biased genes may be explained by other factors than sexual selection such as tissue-specific expression, which we observed here, genetic drift, turnovers in expression patterns and relaxed selective constraints (MANK et al., 2008; MEISEL, 2011; GERSHONI; PIETROKOVSKI, 2017; MANK, 2017). We found evidence of both selection and relaxed constraints in these genes. MKT and high rates of evolution (Ka/Ks > 1) exhibited greater proportion of genes that potentially evolved under positive selection in male-biased genes when compared to unbiased genes, even though we failed to find a greater proportion of biased expressed genes with significant branch-site tests in the *Anastrepha* branch. It is possible that this is a consequence of the reduced number of substitutions in the short branch between the two recently diverged species that failed to reach the significance level in the branch-site test, but this may also be caused by higher rates of evolution in male-biased genes which would complicate orthology assignment even for species in the same genus (ELLEGREN; PARSCH, 2007). In fact, ~50% of male-biased genes only showed orthologs in *Anastrepha* but not to other more distantly related species, preventing their analysis in the branch-site test and thus potentially producing a bias in the proportion of positively selected genes. Furthermore, we found greater proportions of genes with signals of relaxed selection in both species male-biased and species-specific female-biased genes than in unbiased genes, that contribute to explain the proportion of genes with high rates of evolution. In the case of male-biased, we found two genes (out of 23) with Ka/Ks > 1 and significant for relaxed selective constraints, but no one with significant branch-site test.

Comparison between head and reproductive transcriptomes from *A. fraterculus* and *A. obliqua* revealed that reproductive genes evolve faster than head genes in both sexes, showing similar patterns to *Drosophila* (JAGADEESHAN; SINGH, 2005). This outcome may be explained by broader patterns of expression in those genes, but since there is no available data for other tissues in these species, we are not able to estimate their actual specificity. Nevertheless, testis typically shows a greater proportion of tissue-specific genes (BAKER et al., 2011; EMIG; KACPROWSKI; ALBRECHT, 2011; MEIKLEJOHN; PRESGRAVES, 2012; YANG et al., 2016), thus it is likely that several reproductive-biased genes would be tissue-specific, particularly for males. These potentially tissue-specific genes would be more likely to evolve under positive or relaxed selection, while generally expressed genes seem to experience stronger evolutionary constraints, possibly due to pleiotropy (MANK et al., 2008; HAYGOOD et al., 2010; KRYUCHKOVA-MOSTACCI; ROBINSON-RECHAVI, 2015), or because genes that are expressed in different tissues across the organism are more likely to be part of the constitutive set of essential genes for cell function, thus harboring lower non-synonymous rates. Finally, gene duplication can reduce functional constraints, so distinct levels of paralogy may also produce this outcome. Nevertheless, the mammalian gene family content is equivalent between housekeeping and tissue-specific genes (ZHANG; LI, 2004), hence, if this pattern is also true in insects, gene redundancy may not be a plausible explanation.

Our analysis identified hundreds of SNPs associated to unigenes that showed fixed, or nearly fixed, differences between *A. fraterculus* and *A. obliqua*, which have been significantly more associated with non-synonymous substitutions than to other substitutions and point to an important role for selection in their differentiation. Even though we still lack a formal connection between sex-biased genes and speciation, the first “speciation gene“ identified in *Drosophila* was the male-biased *Odysseus site homeo-box*, associated with post-zygotic isolation mechanisms which produce sterility in male hybrids (TING et al., 1998). Furthermore, if a set of sex-biased genes evolves under sexual selection or sexual conflict which may lead to intraspecific intersexual divergence, these same differences may foster differentiation between species, especially when the genes involved are sex-biased in a species-specific manner.

2.7. Conclusions

Our work not only contributes to the current functional genomic knowledge on two of the most important fruit pests from the Neotropics by generating next-generation transcriptome data for reproductive tissues which have been hitherto unavailable, but also explored differences in expression patterns between sexes and tissues. Although several studies are available on this matter for a wide variety of animals, particularly *Drosophila* (MALONE; HAWKINS; MICHALAK, 2006; YANG et al., 2006; PERRY; HARRISON; MANK, 2014; HARRISON et al., 2015; DEAN et al., 2017), little had been known for Tephritidae species. In this matter, our findings indicate that head tissues of *A. fraterculus* and *A. obliqua* exhibit few genes with sex-biased expression. More importantly, sexual dimorphism in expression profiles of reproductive tissues revealed that sex-biased genes evolve faster than unbiased genes, especially in males, a pattern that was associated with signals of positive selection and relaxed constraints. Our results shed some light on the evolution of sex- and tissue-biased genes expressed in reproductive and head tissue of *A. fraterculus* and *A. obliqua* which should be valuable to other species as well. Furthermore, we found a set of sex-biased genes in reproductive tissues that may be candidates to be involved in the differentiation process of *A. fraterculus* and *A. obliqua*. However, further studies that evaluate the populational variation of these genes are necessary to corroborate their role in the differentiation of these and other species of the *fraterculus* group.

2.8. References

ABASCAL, F.; ZARDOYA, R.; TELFORD, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. **Nucleic Acids Research,** v. 38, p. W7-13, 2010.

ANDRÉS, J. A. et al. Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. **Genetics,** v. 193, n. 2, p. 501-513, 2013.

ARAMA, E. et al. A ubiquitin ligase complex regulates caspase activation during sperm differentiation in *Drosophila*. **PLOS Biology,** v. 5, n. 10, p. e251, 2007.

ASSIS, R.; ZHOU, Q.; BACHTROG, D. Sex-biased transcriptome evolution in *Drosophila*. **Genome Biology and Evolution,** v. 4, n. 11, p. 1189-1200, 2012.

BAKER, D. A. et al. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*. **BMC Genomics,** v. 12, n. 1, p. 296, 2011.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114–2120, 2014.

BRAND, P. et al. Rapid evolution of chemosensory receptor genes in a pair of sibling species of orchid bees (Apidae: Euglossini). **BMC Evolutionary Biology,** v. 15, p. 176, 2015.

BRUEN, T. C.; PHILIPPE, H.; BRYANT, D. A simple and robust statistical test for detecting the presence of recombination. **Genetics,** v. 172, n. 4, p. 2665-2681, 2006.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics,** v. 10, n. 1, p. 421, 2009.

CAMPANINI, E. B.; DE BRITO, R. A. Molecular evolution of Odorant-binding proteins gene family in two closely related *Anastrepha* fruit flies. **BMC Evolutionary Biology,** v. 16, n. 1, p. 198, 2016.

CAPELLA-GUTIÉRREZ, S.; SILLA-MARTÍNEZ, J. M.; GABALDÓN, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics,** v. 25, n. 15, p. 1972-1973, 2009.

CARVALHO, A. B.; LAZZARO, B. P.; CLARK, A. G. Y chromosomal fertility factors kl-2 and kl-3 of *Drosophila melanogaster* encode dynein heavy chain polypeptides. **Proceedings of the National Academy of Sciences,** v. 97, n. 24, p. 13239-13244, 2000.

CHANG, P. L. et al. Somatic sex-specific transcriptome differences in *Drosophila* revealed by whole transcriptome sequencing. **BMC Genomics,** v. 12, n. 1, p. 364, 2011.

CHOMCZYNSKI, P.; MACKEY, K. Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide-and proteoglycan-rich sources. **Biotechniques,** v. 19, n. 6, p. 942-945, 1995.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics,** v. 21, n. 18, p. 3674-3676, 2005.

CRUICKSHANK, T. E.; HAHN, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. **Molecular Ecology,** v. 23, n. 13, p. 3133-3157, 2014.

DAUWALDER, B. et al. The *Drosophila takeout* gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. **Genes & Development,** v. 16, n. 22, p. 2879-2892, 2002.

DEAN, R. et al. Sperm competition shapes gene expression and sequence evolution in the ocellated wrasse. **Molecular Ecology,** v. 26, n. 2, p. 505-518, 2017.

DÍAZ, F. et al. Impact of agricultural activities on introgression and population expansion of three species of the *Anastrepha fraterculus* group, a radiating species complex of fruit flies. **Evolutionary Applications**, 2017, submitted.

DORUS, S. et al. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. **Nat Genet,** v. 38, n. 12, p. 1440-1445, 2006.

DOS SANTOS, P.; URAMOTO, K.; MATIOLI, S. R. Experimental hybridization among *Anastrepha* species (Diptera: Tephritidae): Production and morphological characterization of F1 hybrids. **Annals of the Entomological Society of America,** v. 94, n. 5, p. 717-725, 2001.

EBERHARD, W. G.; CORDERO, C. Sexual selection by cryptic female choice on male seminal products-a new bridge between sexual selection and reproductive physiology. **Trends in Ecology & Evolution,** v. 10, n. 12, p. 493-496, 1995.

EDDY, S. R. Accelerated profile HMM searches. **PLOS Computational Biology,** v. 7, n. 10, p. e1002195, 2011.

ELLEGREN, H.; PARSCH, J. The evolution of sex-biased genes and sex-biased gene expression. **Nature Reviews Genetics,** v. 8, n. 9, p. 689-698, 2007.

EMIG, D.; KACPROWSKI, T.; ALBRECHT, M. Measuring and analyzing tissue specificity of human genes and protein complexes. **EURASIP Journal on Bioinformatics and Systems Biology,** v. 2011, n. 1, p. 5, 2011.

FATIMA, R. *Drosophila* dynein intermediate chain gene, *Dic61B*, is required for spermatogenesis. **PLOS ONE,** v. 6, n. 12, p. e27822, 2011.

FAY, J. C.; WYCKOFF, G. J.; WU, C.-I. Positive and negative selection on the human genome. **Genetics,** v. 158, n. 3, p. 1227-1234, 2001.

FEDER, J. L.; EGAN, S. P.; NOSIL, P. The genomics of speciation-with-gene-flow. **Trends in Genetics,** v. 28, n. 7, p. 342-350, 2012.

FU, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics,** v. 28, n. 23, p. 3150-3152, 2012.

FU, Y.-X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. **Genetics,** v. 147, n. 2, p. 915-925, 1997.

GAVRILETS, S. Is sexual conflict an “engine of speciation”? **Cold Spring Harbor Perspectives in Biology,** v. 6, n. 12, 2014.

GERSHONI, M.; PIETROKOVSKI, S. Reduced selection and accumulation of deleterious mutations in genes exclusively expressed in men. **Nature communications,** v. 5, p. 4438, 2014.

______. The landscape of sex-differential transcriptome and its consequent selection in human adults. **BMC Biology,** v. 15, n. 1, p. 7, 2017.

GOLDMAN, T. D.; ARBEITMAN, M. N. Genomic and functional studies of *Drosophila* sex hierarchy regulated gene expression in adult head and nervous system tissues. **PLOS Genetics,** v. 3, n. 11, p. e216, 2007.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature biotechnology,** v. 29, n. 7, p. 644-652, 2011.

HAERTY, W. et al. Evolution in the fast lane: Rapidly evolving sex-related genes in *Drosophila*. **Genetics,** v. 177, n. 3, p. 1321-1335, 2007.

HARRISON, P. W. et al. Sexual selection drives evolution and rapid turnover of male gene expression. **Proceedings of the National Academy of Sciences,** v. 112, n. 14, p. 4393-4398, 2015.

HAYGOOD, R. et al. Contrasts between adaptive coding and noncoding changes during human evolution. **Proceedings of the National Academy of Sciences,** v. 107, n. 17, p. 7853-7857, 2010.

HE, Q. et al. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. **Nat Genet,** v. 43, n. 5, p. 414-420, 2011.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian Journal of Statistics,** v. 6, n. 2, p. 65-70, 1979.

HONGO, J. A. et al. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. **BMC Genomics,** v. 16, n. 1, p. 567, 2015.

HOWARD, R. W.; BLOMQUIST, G. J. Ecological, behavioral, and biochemical aspects of insect hydrocarbons. **Annual Review of Entomology,** v. 50, n. 1, p. 371-393, 2005.

HSU, J.-C. et al. Discovery of genes related to insecticide resistance in *Bactrocera dorsalis* by functional genomic analysis of a *de novo* assembled transcriptome. **PLoS ONE,** v. 7, n. 8, p. e40950, 2012.

HUERTA-CEPAS, J.; SERRA, F.; BORK, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. **Molecular Biology and Evolution,** v. 33, n. 6, p. 1635-1638, 2016.

HUYLMANS, A. K. et al. *De novo* transcriptome assembly and sex-biased gene expression in the cyclical parthenogenetic *Daphnia galeata*. **Genome Biology and Evolution,** v. 8, n. 10, p. 3120-3139, 2016.

IMMONEN, E.; RITCHIE, M. G. The genomic response to courtship song stimulation in female *Drosophila melanogaster*. **Proceedings of the Royal Society B: Biological Sciences,** v. 279, n. 1732, p. 1359-1365, 2012.

JAGADEESHAN, S.; SINGH, R. S. Rapidly evolving genes of *Drosophila*: Differing levels of selective pressure in testis, ovary, and head tissues between sibling species. **Molecular Biology and Evolution,** v. 22, n. 9, p. 1793-1801, 2005.

______. Rapid evolution of outer egg membrane proteins in the *Drosophila melanogaster* subgroup: A case of ecologically driven evolution of female reproductive traits. **Molecular Biology and Evolution,** v. 24, n. 4, p. 929-938, 2007.

JIN, X.; HA, T. S.; SMITH, D. P. SNMP is a signaling component required for pheromone sensitivity in *Drosophila*. **Proceedings of the National Academy of Sciences,** v. 105, n. 31, p. 10996-11001, 2008.

JONES, P. et al. InterProScan 5: genome-scale protein function classification. **Bioinformatics,** v. 30, n. 9, p. 1236-1240, 2014.

JUÁREZ, M. L. et al. Evaluating mating compatibility within fruit fly cryptic species complexes and the potential role of sex pheromones in pre-mating isolation. **ZooKeys,** v. 540, p. 125-155, 2015.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. **Molecular Biology and Evolution,** v. 30, n. 4, p. 772-780, 2013.

KOBOLDT, D. C. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. **Genome Research,** v. 22, n. 3, p. 568-576, 2012.

KOHL, J.; HUOVIALA, P.; JEFFERIS, G. S. Pheromone processing in *Drosophila*. **Current Opinion in Neurobiology,** v. 34, p. 149-157, 2015.

KOONIN, E. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. **Genome Biology,** v. 5, n. 2, p. R7, 2004.

KRYUCHKOVA-MOSTACCI, N.; ROBINSON-RECHAVI, M. Tissue-specific evolution of protein coding genes in human and mouse. **PLOS ONE,** v. 10, n. 6, p. e0131673, 2015.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods,** v. 9, n. 4, p. 357-359, 2012.

LI, H. et al. The sequence alignment/map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, 2009.

MALONE, J. H.; HAWKINS, D. L.; MICHALAK, P. Sex-biased gene expression in a ZW sex determination system. **Journal of Molecular Evolution,** v. 63, n. 4, p. 427-436, 2006.

MANK, J. E. The transcriptional architecture of phenotypic dimorphism. **Nature Ecology & Evolution,** v. 1, p. 0006, 2017.

MANK, J. E. et al. Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. **Molecular Biology and Evolution,** v. 24, n. 12, p. 2698-2706, 2007.

MANK, J. E. et al. Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. **The American Naturalist,** v. 171, n. 1, p. 35-43, 2008.

MANNI, M. et al. Relevant genetic differentiation among Brazilian populations of *Anastrepha fraterculus* (Diptera, Tephritidae). **ZooKeys,** v. 540, 2015.

MARTIN, S. H. et al. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. **Genome Research,** v. 23, n. 11, p. 1817-1828, 2013.

MCDONALD, J. H.; KREITMAN, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. **Nature,** v. 351, n. 6328, p. 652-654, 1991.

MEIKLEJOHN, C. D.; PRESGRAVES, D. C. Little evidence for demasculinization of the *Drosophila* X chromosome among genes expressed in the male germline. **Genome Biology and Evolution,** v. 4, n. 10, p. 1007-1016, 2012.

MEISEL, R. P. Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution. **Molecular Biology and Evolution,** v. 28, n. 6, p. 1893-1900, 2011.

MORROW, J. et al. Comprehensive transcriptome analysis of early male and female *Bactrocera jarvisi* embryos. **BMC Genetics,** v. 15, n. Suppl 2, p. S7, 2014.

NOOR, M. A. F.; BENNETT, S. M. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. **Heredity,** v. 103, n. 6, p. 439-444, 2009.

NORRBOM, A. L.; ZUCCHI, R. A.; HERNÁNDEZ-ORTIZ, V. Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotrypanini) based on morphology. In: ALUJA, M. e NORRBOM, A. L. (Ed.). **Fruit flies (Tephritidae): phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 12, p.299-342.

NOSIL, P.; FEDER, J. L. Genomic divergence during speciation: causes and consequences. **Philosophical Transactions of the Royal Society of London B: Biological Sciences,** v. 367, n. 1587, p. 332-342, 2012.

PANHUIS, T. M.; SWANSON, W. J. Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*. **Genetics,** v. 173, n. 4, p. 2039-2047, 2006.

PAPA, F. et al. Rapid evolution of female-biased genes among four species of *Anopheles* malaria mosquitoes. **Genome Research,** v. 27, p. 1536-1548, 2017.

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 Biology,** v. 17, n. 1, p. 192, 2016.

PARISI, M. et al. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. **Genome Biology,** v. 5, n. 6, p. R40, 2004.

PARSCH, J.; ZHANG, Z.; BAINES, J. F. The influence of demography and weak selection on the McDonald–Kreitman test: An empirical study in *Drosophila*. **Molecular Biology and Evolution,** v. 26, n. 3, p. 691-698, 2009.

PERRE, P. et al. Toward an automated identification of *Anastrepha* fruit flies in the *fraterculus* group (Diptera, Tephritidae). **Neotropical Entomology,** v. 45, n. 5, p. 554–558, 2016.

PERRY, J. C.; HARRISON, P. W.; MANK, J. E. The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. **Molecular Biology and Evolution,** v. 31, n. 5, p. 1206-1219, 2014.

POSADA, D. Using MODELTEST and PAUP* to select a model of nucleotide substitution. In: BAXEVANIS, A.;DAVISON, D., *et al* (Ed.). **Current Protocols in Bioinformatics**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003.

PRÖSCHEL, M.; ZHANG, Z.; PARSCH, J. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. **Genetics,** v. 174, n. 2, p. 893-900, 2006.

PURANDARE, S. R. et al. Accelerated evolution of morph-biased Genes in pea aphids. **Molecular Biology and Evolution,** v. 31, n. 8, p. 2073-2083, 2014.

RAM, K. R.; WOLFNER, M. F. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. **Integrative and Comparative Biology,** v. 47, n. 3, p. 427-445, 2007.

RENAUT, S.; NOLTE, A. W.; BERNATCHEZ, L. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). **Molecular Ecology,** v. 19, p. 115-131, 2010.

REZENDE, V. B. et al. Head transcriptomes of two closely related species of fruit flies of the *Anastrepha fraterculus* group reveals divergent genes in species with extensive gene flow. **G3: Genes|Genomes|Genetics,** v. 6, n. 10, p. 3283-3295, 2016.

ROBERTS, A.; PACHTER, L. Streaming fragment assignment for real-time analysis of sequencing experiments. **Nature Methods,** v. 10, n. 1, p. 71-73, 2013.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics,** v. 26, n. 1, p. 139-140, 2010.

RUIZ, M. et al. Sex-dependent modulation of longevity by two *Drosophila* homologues of human Apolipoprotein D, GLaz and NLaz. **Experimental Gerontology,** v. 46, n. 7, p. 579-589, 2011.

RULL, J. et al. Evolution of pre-zygotic and post-zygotic barriers to gene flow among three cryptic species within the *Anastrepha fraterculus* complex. **Entomologia Experimentalis et Applicata,** v. 148, n. 3, p. 213-222, 2013.

SALVEMINI, M. et al. *De Novo a*ssembly and transcriptome analysis of the Mediterranean Fruit Fly *Ceratitis capitata e*arly embryos. **PLoS ONE,** v. 9, n. 12, p. e114191, 2014.

SÁNCHEZ-GRACIA, A. et al. Comparative genomics of the major chemosensory gene families in Arthropods. In: (Ed.). **Encyclopedia of Life Sciences**. Chichester (UK): John Wiley & Sons, Ltd, 2011.

SANCHEZ-GRACIA, A.; VIEIRA, F. G.; ROZAS, J. Molecular evolution of the major chemosensory gene families in insects. **Heredity,** v. 103, n. 3, p. 208-216, 2009.

SCALLY, M. et al. Resolution of inter and intra-species relationships of the West Indian fruit fly *Anastrepha obliqua*. **Molecular Phylogenetics and Evolution,** v. 101, p. 286-293, 2016.

SIM, S. B.; GEIB, S. M. A chromosome-scale ssembly of the *Bactrocera cucurbitae* genome provides insight to the genetic basis of *white pupae*. **G3: Genes|Genomes|Genetics**, 2017.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics,** v. 31, n. 19, p. 3210-3212, 2015.

SITNIK, J. L. et al. The female post-mating response requires genes expressed in the secondary cells of the male accessory gland in *Drosophila melanogaster*. **Genetics,** v. 202, n. 3, p. 1029-1041, 2016.

SMITH-CALDAS, M. R. B. et al. Phylogenetic relationships among species of the *fraterculus* group (*Anastrepha*: Diptera: Tephritidae) inferred from DNA sequences of mitochondrial cytochrome oxidase I. **Neotropical Entomology,** v. 30, n. 4, p. 565-573, 2001.

SNOOK, R. R. et al. Interactions between the sexes: new perspectives on sexual selection and reproductive isolation. **Evolutionary ecology,** v. 23, n. 1, p. 71-91, 2009.

STAMATAKIS, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics,** v. 30, n. 9, p. 1312-1313, 2014.

STOLETZKI, N.; EYRE-WALKER, A. Estimation of the neutrality index. **Molecular Biology and Evolution,** v. 28, n. 1, p. 63-70, 2011.

SWANSON, W. J. et al. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. **Proceedings of the National Academy of Sciences,** v. 98, n. 13, p. 7375-7379, 2001.

SWANSON, W. J.; VACQUIER, V. D. The rapid evolution of reproductive proteins. **Nature Reviews Genetics,** v. 3, n. 2, p. 137-144, 2002.

SWANSON, W. J. et al. Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. **Genetics,** v. 168, n. 3, p. 1457-1465, 2004.

TING, C.-T. et al. A rapidly evolving homeobox at the site of a hybrid sterility gene. **Science,** v. 282, n. 5393, p. 1501-1504, 1998.

TOOTLE, T. L. et al. *Drosophila* eggshell production: Identification of new genes and coordination by Pxt. **PLOS ONE,** v. 6, n. 5, p. e19943, 2011.

TORGERSON, D. G.; KULATHINAL, R. J.; SINGH, R. S. Mammalian sperm proteins are rapidly evolving: Evidence of positive selection in functionally diverse genes. **Molecular Biology and Evolution,** v. 19, n. 11, p. 1973-1980, 2002.

VACQUIER, V. D.; SWANSON, W. J. Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. **Cold Spring Harbor Perspectives in Biology,** v. 3, n. 11, p. a002931, 2011.

VAN DYKEN, J. D.; WADE, M. J. The genetic signature of conditional expression. **Genetics,** v. 184, n. 2, p. 557-570, 2010.

WERTHEIM, J. O. et al. RELAX: Detecting relaxed selection in a phylogenetic framework. **Molecular Biology and Evolution,** v. 32, n. 3, p. 820-832, 2015.

WHITTLE, C. A.; EXTAVOUR, C. G. Rapid evolution of ovarian-biased genes in the yellow fever mosquito *Aedes aegypti*. **Genetics,** v. 206, n. 4, p. 2119-2137, 2017.

YAMAGUCHI, S. et al. Involvement of EDTP, an egg-derived tyrosine phosphatase, in the early development of Drosophila melanogaster. **Journal of Biochemistry,** v. 138, n. 6, p. 721-728, 2005.

YANG, L.; ZHANG, Z.; HE, S. Both male-biased and female-biased genes evolve faster in fish genomes. **Genome Biology and Evolution,** v. 8, n. 11, p. 3433-3445, 2016.

YANG, X. et al. Tissue-specific expression and regulation of sexually dimorphic genes in mice. **Genome Research,** v. 16, n. 8, p. 995-1004, 2006.

YANG, Y. et al. Regulation of flagellar motility by the conserved flagellar protein CG34110/Ccdc135/FAP50. **Molecular Biology of the Cell,** v. 22, n. 7, p. 976-987, 2011.

YANG, Y.; SMITH, S. A. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. **Molecular Biology and Evolution,** v. 31, n. 11, p. 3081-3092, 2014.

YANG, Z. PAML 4: phylogenetic analysis by maximum likelihood. **Molecular biology and evolution,** v. 24, n. 8, p. 1586-1591, 2007.

YE, J. et al. WEGO: a web tool for plotting GO annotations. **Nucleic Acids Research,** v. 34, n. suppl 2, p. W293-W297, 2006.

ZHANG, J.; NIELSEN, R.; YANG, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. **Molecular Biology and Evolution,** v. 22, n. 12, p. 2472-2479, 2005.

ZHANG, L.; LI, W.-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. **Molecular Biology and Evolution,** v. 21, n. 2, p. 236-239, 2004.

ZHANG, Y. et al. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. **Nature,** v. 450, n. 7167, p. 233-237, 2007.

ZHANG, Z.; HAMBUCH, T. M.; PARSCH, J. Molecular evolution of sex-biased genes in *Drosophila*. **Molecular Biology and Evolution,** v. 21, n. 11, p. 2130-2139, 2004.

ZHANG, Z. et al. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. **Genomics, Proteomics & Bioinformatics,** v. 4, n. 4, p. 259-263, 2006.

ZHANG, Z. et al. Dual roles of Hh signaling in the regulation of somatic stem cell self-renewal and germline stem cell maintenance in *Drosophila* testis. **Cell Res,** v. 23, n. 4, p. 573-576, 2013.

ZUCCHI, R. A. Taxonomia. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000. p.1-24.

Appendix 1

```
#Config file including all parameters used in POTION.

##############PROJECT PARAMETERS#####################

mode = site                    # main analysis mode. Currently POTION supports only site-models analysis.

CDS_dir_path = /labev/carlos/transcriptome/xsp_3/selection/potion/sequences/              #path to folder containing CDS data

homology_file_path = /labev/carlos/transcriptome/xsp_3/selection/potion/groups/ortholog_clusters        # path to the ORTHOMCL 1.4 main output file

project_dir_path = /labev/carlos/transcriptome/xsp_3/selection/potion/results_50          # path to the main directory where results will be created.Parent directory must exist.

max_processors = 22            # number of processors to be used in parallelized steps of POTION

remove_identical = yes         # "yes" to remove 100% identical nucleotide groups at the very beginning of

                          # analysis, "no" otherwise

verbose = 1                    # 1 to print nice log messages telling you what is going on. 0 otherwise

###########SEQUENCE/GROUP PARAMETERS###############

groups_to_process = all        # Defines which lines of the cluster file (ortholog groups) will be processed.

                          # Use "all" to process every group, "-" to set groups between two given lines

                          # (including the said lines).

                          # Use "!" to not process a specific line, can be used with "-" to specify a

                          # set to not be processed. Useful if groups are taking too long to finish.

                               # Use "," or ";" to set distinct sets

                               # Examples: 1;4-10;12  will process groups 1, 4 to 10 and group 12

                               #      all;!3     will process all groups, except group 3

                               #      all;!3-5   will process all groups, except groups 3 to 5

behavior_about_bad_clusters = 1     # what should POTION do if it finds a cluster with a sequence removed

                          # due to any filter? Possible options are:

                          # 0 - does not filter any sequence (not recommended)

                          # 1 - removal of any flagged sequence

                          # 2 - removal of any group with flagged sequences
```

```
homology_filter = 1            # this variable controls for what POTION will do if a group with paralogous
                               # genes is found. Possible options are:
                               # 0 - analyze all sequences within group
                               # 1 - remove all paralogous within group, analyzing only single-copy genes
                               # 2 - remove groups with paralogous genes
                               # 3 - remove single-copy genes, analyzing all paralogous within group together
                               # 4 - remove single-copy genes and split remaining paralogous into individual
                               # species, evaluating each subgroup individually
validation_criteria = 3            # quality criteria to remove sequences. Possible values are:
                               # 1 - checks for valid start codons
                               # 2 - checks for valid stop codons
                               # 3 - checks for sequence size multiple of 3
                               # 4 - checks for nucleotides outside ATCG
                               # 'all' applies every verification
additional_start_codons = ()          # these codons, plus the ones specified in codon table, will be the valid start
                               # codons for validation purposes
additional_stop_codons = ()           # same as start codons
codon_table = 1                    # codon table id (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi)
absolute_min_sequence_size = 150     # minimum sequence length cutoff for sequence/group further evaluation
absolute_max_sequence_size = 40000 # maximum sequence length cutoff for sequence/group further evaluation
relative_min_sequence_size = 0.70      # sequences smaller than mean|meadian times this value will be filtered
relative_max_sequence_size = 1.3       # sequences greater than mean|meadian times this value will be filtered
sequence_size_average_metric = mean     # which average metric will be calculated to determine the
                               # minimum/maximum relative lengths ranges for sequence removal
                               # Possible values are "mean" and "median"
min_group_identity = 50              # mean minimum group identity cutoff in pairwise sequence alignments
max_group_identity = 99.9              # mean maximum group identity cutoff in pairwise sequence alignments
group_identity_comparison = aa        # the kind of sequence that will be used when computing mean group
identity
                              # possible values are "nt" or "aa"
```

```
min_sequence_identity = 50          # minimum (mean/median) sequence identity cutoff in pairwise sequence
alignments

max_sequence_identity = 99.99     # maximum (mean/median) sequence identity cutoff in pairwise sequence
alignemnts

sequence_identity_average_metric = mean # would you like to use mean or median to measure sequence
identity?

                              # possible values are "mean" and "median"

sequence_identity_comparison = aa      # the kind of sequence that will be used when computing sequence
identity

                              # possible values are "nt" and "aa"

min_gene_number_per_cluster = 7        # minimum # genes in group after all filtering steps

max_gene_number_per_cluster = 7       # maximum # genes in group after all filtering steps

min_specie_number_per_cluster = 7      # minimum # species in group after all filtering steps

max_specie_number_per_cluster = 7     # maximum # species in group after all filtering steps

reference_genome_file =               # genome reference name, leave blank for none (same name used in fasta
file)

###########THIRD-PARTY SOFTWARE CONFIGURATION###############

multiple_alignment = mafft            # program used for multiple sequence alignment. Possible values are

                              # muscle, mafft and prank

bootstrap = 100                      # number of bootstraps in phylogenetic analysis

phylogenetic_tree_speed = fast         # fast or slow analysis? Used in phylip dnaml or proml only

phylogenetic_tree = phyml_nt          # program used for phylogenetic tree reconstruction. Possible values are

                                            # proml dnaml, phyml_aa and phyml_nt

recombination_qvalue = 0.05           # q-value for recombination detection. Must occur for all the specified tests

rec_minimum_confirmations = 2          # minimum number of significant recombination tests positives

rec_mandatory_tests = phi NSS maxchi2   # any combination of the three test names, separated by spaces, or
# N.A. to use any test

remove_gaps = strictplus      # numeric values between 0 and 1 will remove columns with that percentage of

                               # gaps. Values of "strict" or "strictplus" will use respectively these

                               # filters to remove unreliable regions (described in trimal article)

PAML_models = m12 m78                # codeml models to be generated. "m12" and/or "m78" values acceptable.

pvalue = 0.05                       # p-values for positive selection detection

qvalue = 0.05                       # q-values for positive selection detection
```

Appendix 2

Table S1. Sequencing and *de novo* assembly attributes from *A. fraterculus* and *A. obliqua* transcriptomes.

	A. fraterculus	***A. obliqua***
Cleaned pair-end reads	50,404,832	47,448,324
Total bases assembled	127,880,853	116,985,482
Contigs	133,069	116,688
N50	1,777	1,899
Trinity unigenes	62,247	56,400
Filtered unigenes*	16,020	15,477
Contigs longer than 1000	37,153	34,646
Contigs longer than 2000	16,439	15,977
Contigs longer than 10000	170	164
Median (bp)	483	491
Average (bp)	961.01	1,002.55
Longest contig	27,197	27,455

*Filtered unigenes: Trinity unigenes with non-redundant CDSs and filtered by expression. See Material and Methods for details.

Table S2. Quality assessment summary of raw assemblies and filtered unigenes from *A. fraterculus* and *A. obliqua*.

	A. fraterculus		***A. obliqua***	
	Raw assembly	**Unigenes**	**Raw assembly**	**Unigenes**
Complete BUSCOs	1,015 (95.2)	991 (93%)	1013 (95%)	996 (93.4%)
Complete single-copy BUSCOs	482 (45.2%)	986 (92.5%)	507 (47.6%)	990 (92.9%)
Complete duplicated BUSCOs	533 (50%)	5 (0.5%)	506 (47.5%)	6 (0.6%)
Fragmented BUSCOs	36 (3.4%)	35 (3.3%)	32 (3%)	32 (3%)
Missing BUSCOs	15 (1.4%)	40 (3.8%)	21 (2%)	38 (3.6%)

* BUSCO: Single-copy sequence that represents an ortholog group of Arthropoda.

Table S3. Annotation summary of *A. fraterculus* and *A. obliqua* transcriptomes.

	A. fraterculus	***A. obliqua***
Total number of CDSs	17,136	16,624
Annotated CDSs against *D. melanogaster*	12,048	11,848
Annotated CDSs against nr (Arthropoda)	15,592	15,166
Annotated CDSs with Interproscan	17,136	16,624
CDSs with GO term	11,438	11,249

Table S4. Sex-biased expressed genes in *A. fraterculus,* in *A. obliqua* and in both species.

	A. fraterculus		***A. obliqua***		**Both species***	
	Reproductive	**Cephalic**	**Reproductive**	**Cephalic**	**Reproductive**	**Cephalic**
Male-biased	1813	40	2495	25	488	0
Female-biased	797	36	931	27	115	1
Unbiased	9632	11742	8504	11255	3274	4097

*Both species: Includes only the ortholog pairs showing similar expression patterns.

Table S5. Tissue-biased expressed genes in *A. fraterculus*, in *A. obliqua* and in both species.

	A. fraterculus		***A. obliqua***		**Both species***	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Reproductive-biased	1969	1441	2269	1760	527	234
Head-biased	1520	1259	1157	2086	257	354
Unbiased	9368	9247	8941	7708	3321	3129

*Both species: Includes only the ortholog pairs showing similar expression patterns.

Table S6. Pairwise evolutionary rates for *A. fraterculus*, *A. obliqua* and global comparison of seven Tephritidae species.

	N[a]	**Ka**[b]	**Ks**[b]	**Ka/Ks**[b]	**N**[c]	**dN**[b]	**dS**[b]	**dN/dS**[b]
Reproductive								
Male-biased (both spp)[d]	488	0.0071**	0.0345**	0.2046**	272	0.1916**	2.6670	0.0733**
Male-biased (sp-specific)[e]	282	0.0053**	0.0341	0.1650**	184	0.1425**	2.3539**	0.0642**
Female-biased (both spp)[d]	115	0.0062**	0.0359	0.1871**	71	0.1591**	2.5230	0.0681**
Female-biased (sp-specific)[e]	195	0.0044	0.0313	0.1383	136	0.1120	2.4849*	0.0551**
Unbiased	3274	0.0032	0.0315	0.0983	2481	0.0996	2.6083	0.0412
Male								
Reproductive-biased	527	0.0068**	0.0346**	0.1990**	303	0.1954**	2.7554**	0.0725**
Head-biased	257	0.0036	0.0294*	0.1054	161	0.1095	2.0692**	0.0498**
Unbiased	3321	0.0033	0.0320	0.1004	2506	0.1017	2.6019	0.0419
Female								
Reproductive-biased	234	0.0057**	0.0341	0.1730**	140	0.1734**	2.6986	0.0642**
Head-biased	354	0.0043	0.0308	0.1262**	217	0.1268**	2.1814**	0.0579**
Unbiased	3129	0.0031	0.0319	0.0957	2406	0.0986	2.6117	0.0403

[a] Number of pairs of orthologs of *A. fraterculus* and *A. obliqua*.

[b] Median of evoluationary rates.

[c] Number of cluster of orthologs in *A. fraterculus*, *A. obliqua*, *Ceratitis capitata*, *Rhagoletis zephyria*, *Zeugodacus cucurbitae*, *Bactrocera dorsalis* and *Bactrocera oleae* (Tephritidae).

[d] Both spp: Same expression pattern in *A. fraterculus* and *A. obliqua*.

[e] Sp-specific: Sex-biased expressed gene either in *A. fraterculus* or *A. obliqua*

* Holm corrected p-value of Wilcoxon rank sum test < 0.05 between sex-biased and unbiased genes.

** Holm corrected p-value of Wilcoxon rank sum test < 0.01 between sex-biased and unbiased genes.

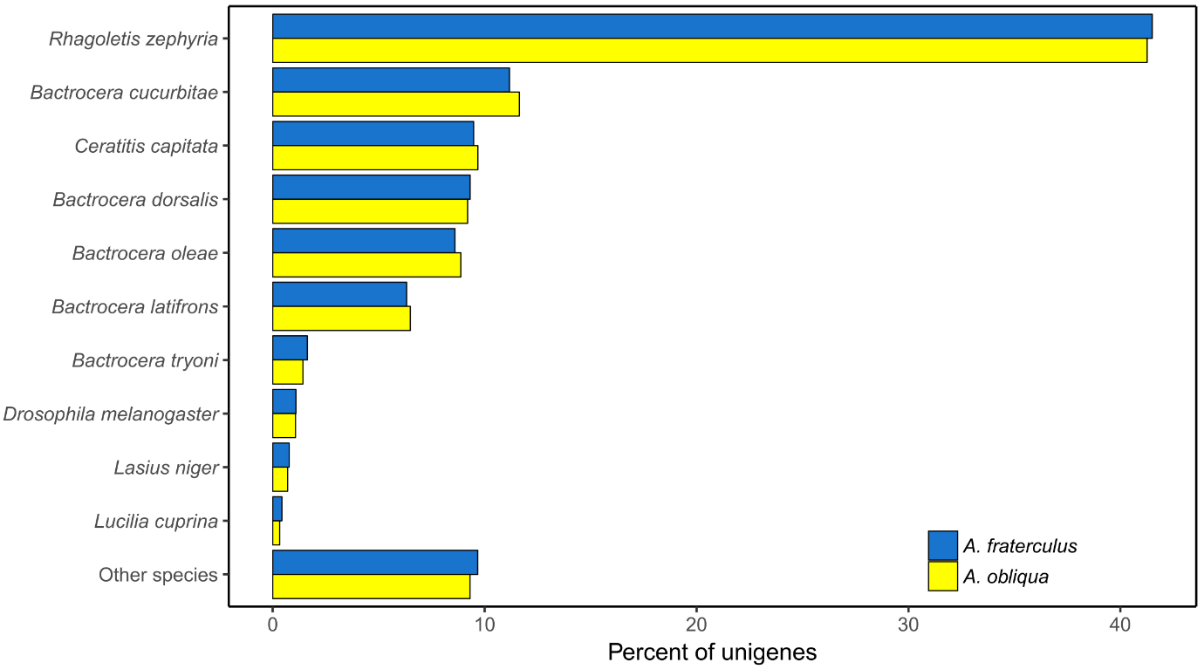


Figure S1.-Top BLAST hit species distribution based on the comparison of CDSs of transcriptomes from *A. fraterculus* and *A. obliqua* against the Genbank's non-redundant (nr) protein database of Arthropoda. Only the 10 most represented species are shown separately, and the rest were included in the category "Other species".

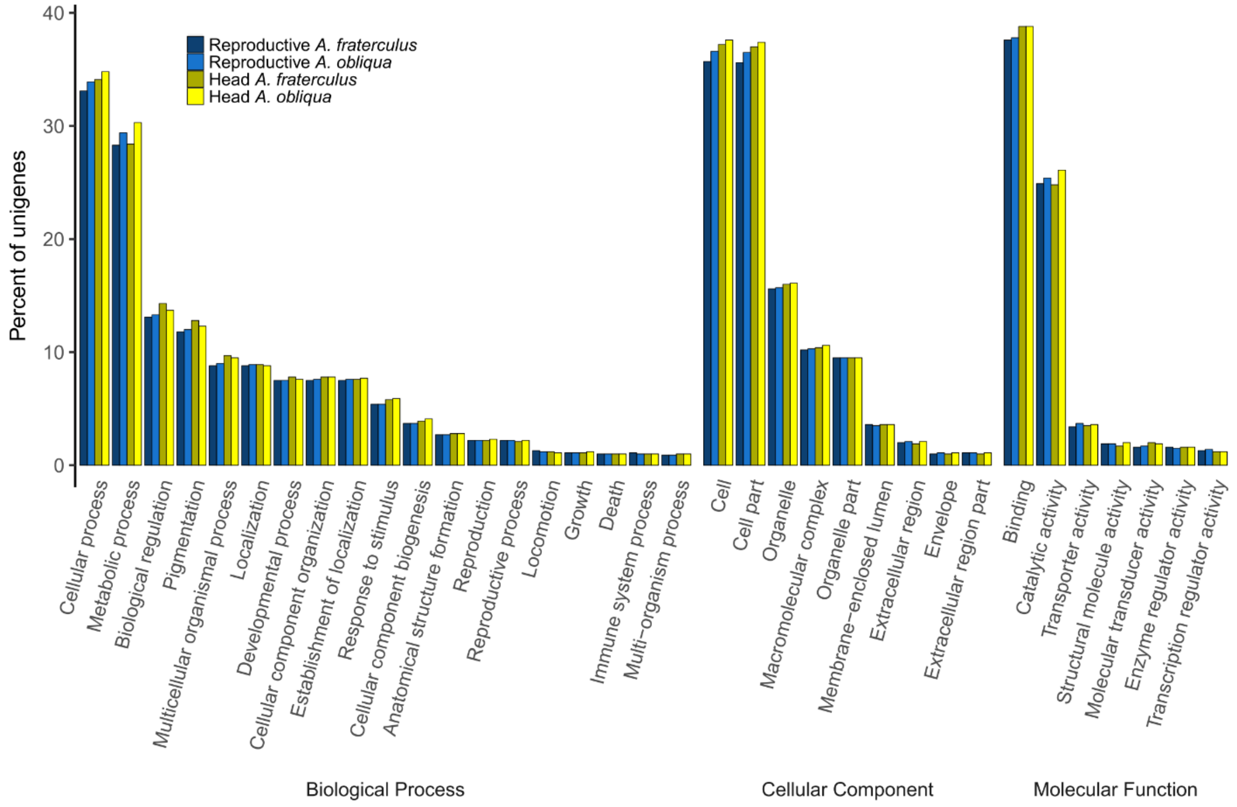


Figure S2.-Gene ontology classification of unigenes from reproductive and head transcriptomes from *A. fraterculus* and *A. obliqua*. The x-axis indicates level 2 GO terms with percentages greater than 1%. The y-axis indicates the percentage of unigenes.

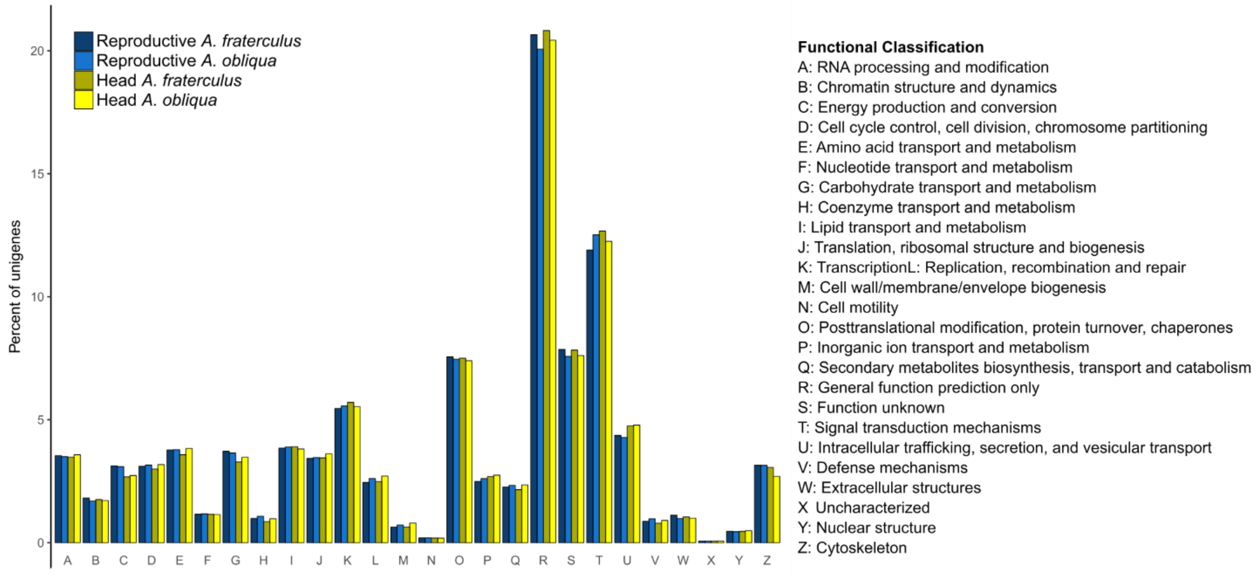


Figure S3.-KOG classification histogram of unigenes of reproductive and head transcriptomes from *A. obliqua* and *A. fraterculus*. The x-axis indicates functional category of groups of KOG. The y-axis indicates the percentage of unigenes.

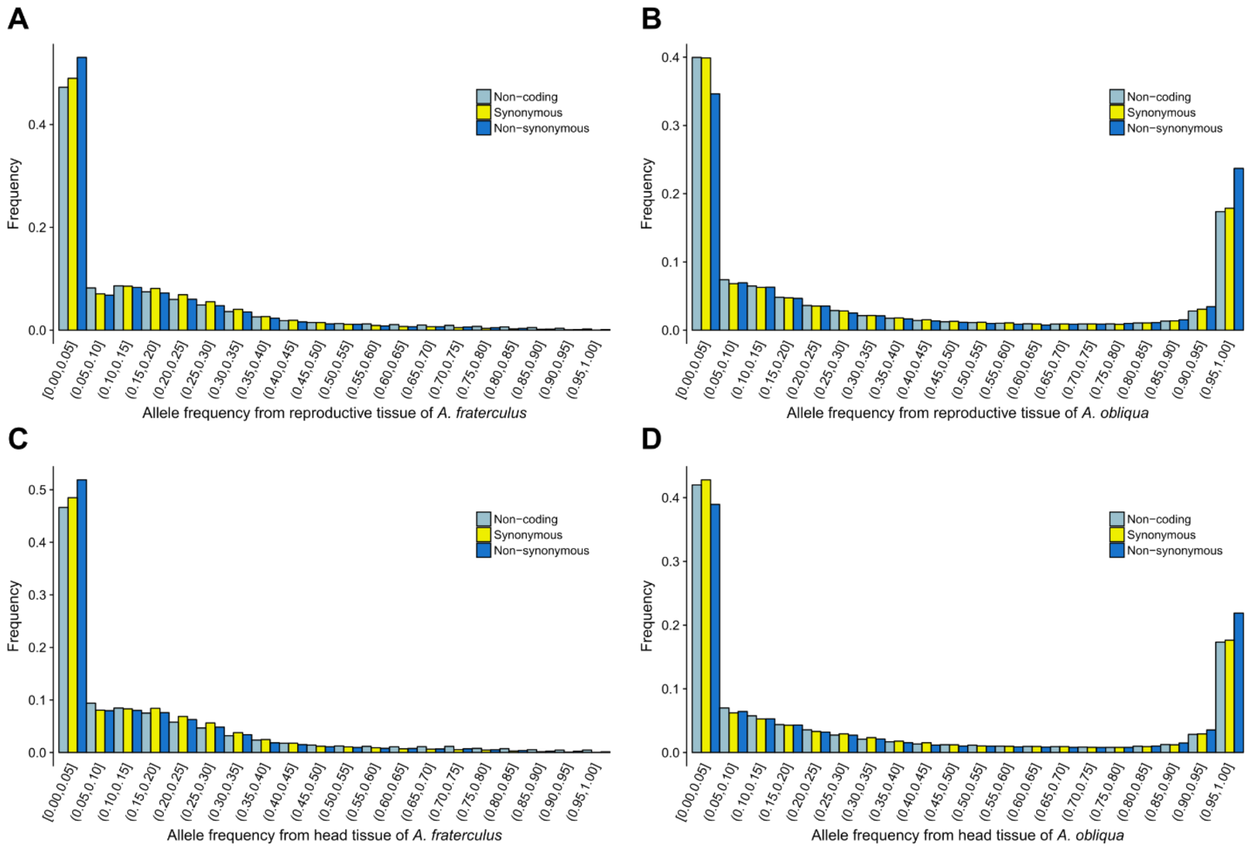


Figure S4.-Histogram of intra- and inter-specific SNP allele frequencies detected in reproductive and head transcriptomes. (A) and (B) SNPs identified from reproductive transcriptomes of *A. fraterculus* and *A. obliqua*, respectively. (C) and (D) SNPs identified from head transcriptomes of *A. fraterculus* and *A. obliqua*, respectively.

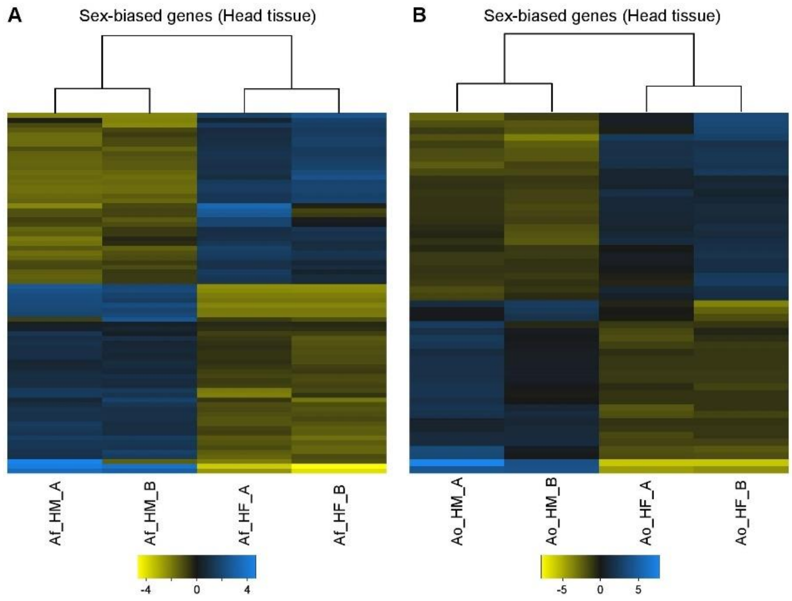


Figure S5.-Heatmap and hierarchical clustering of sex-biased genes from *A. fraterculus* and *A. obliqua* head transcriptomes. (A) Heatmap of 40 male-biased and 36 female-biased genes from *A. fraterculus*. (B) Heatmap of 25 male-biased and 27 female-biased genes from *A. obliqua*. HM and HF: Samples from male and female head tissues, respectively.

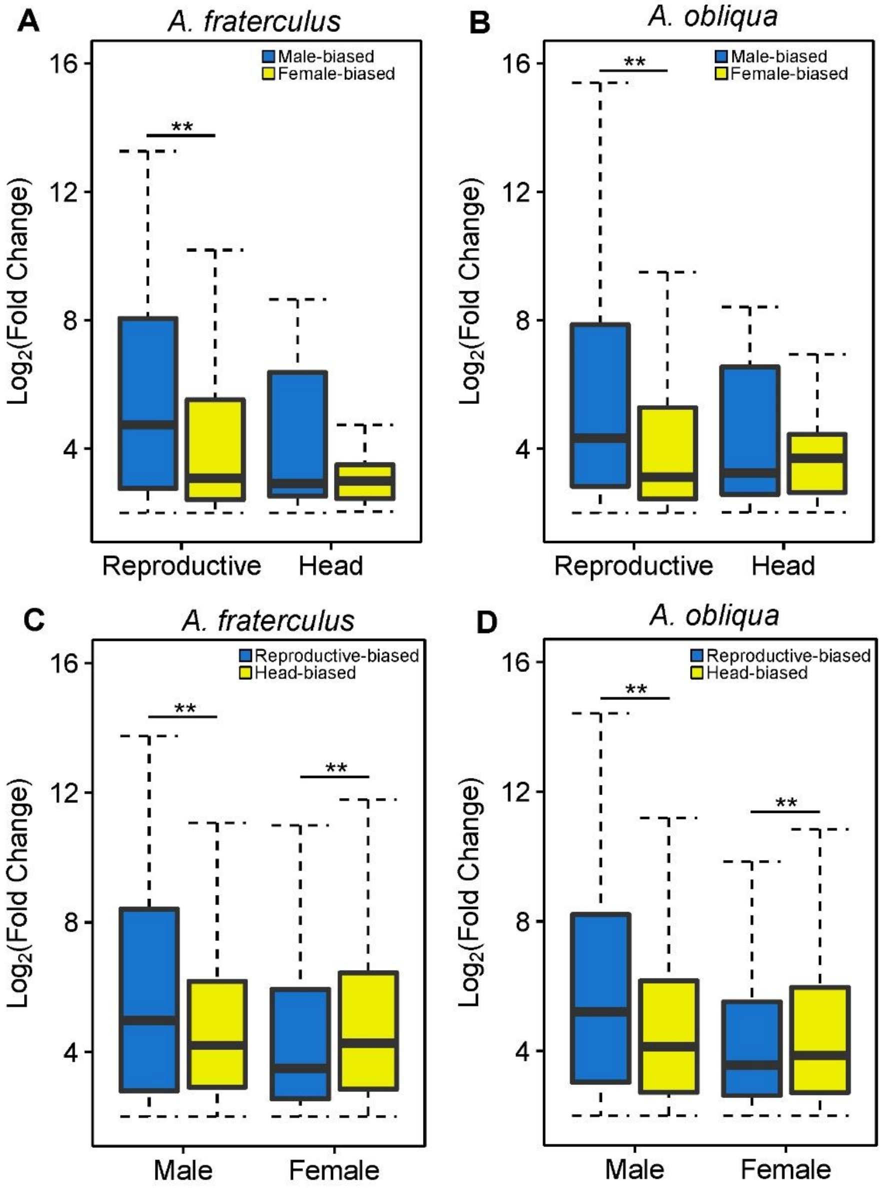


Figure S6.-Comparison of levels of differential expression measured in Log$_2$(fold change) of sex- and tissue-biased genes from *A. fraterculus* and *A. obliqua*. (A) and (B) Comparison of sex-biased genes expressed in reproductive and head transcriptomes from *A. fraterculus* and *A. obliqua*, respectively. (C) and (D) Comparison of tissue-biased genes expressed in males and females from *A. fraterculus* and *A. obliqua*, respectively. ** Wilcoxon rank sum test p-value < 0.01.

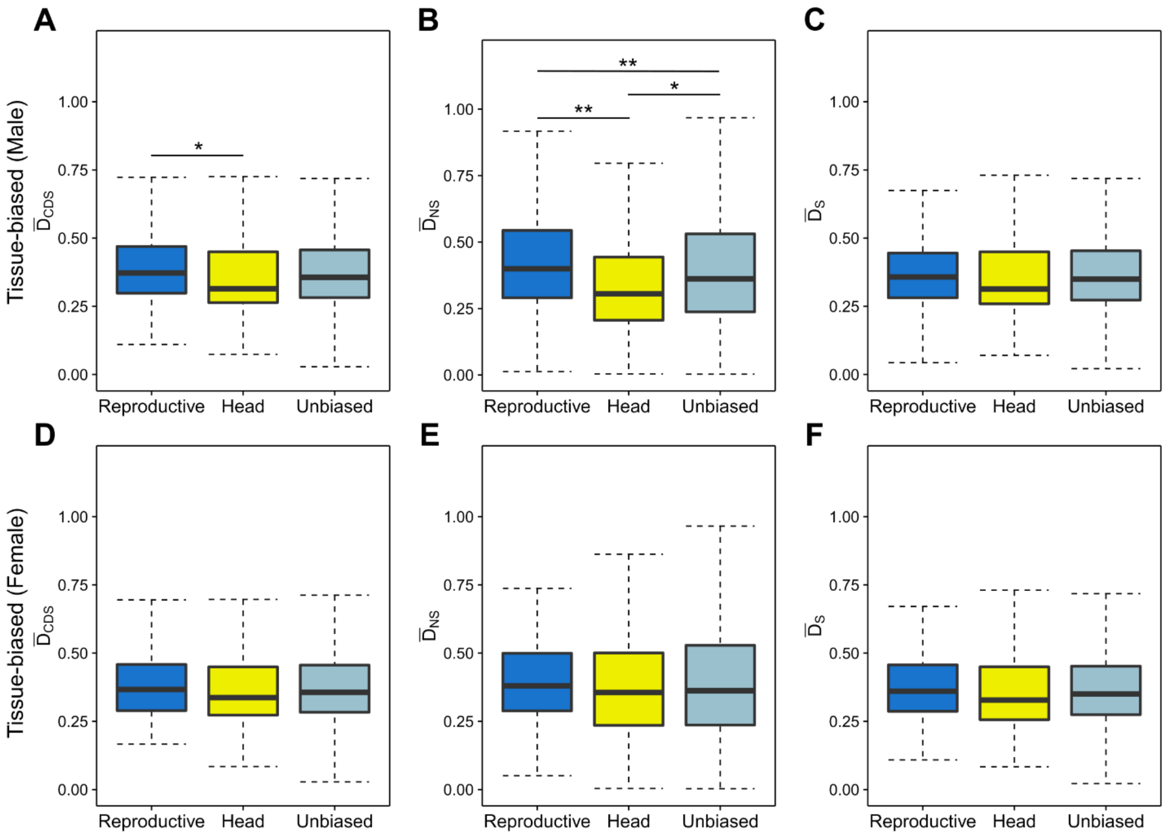


Figure S7.-Boxplots of differentiation indexes among tissue-biased and unbiased genes. Differentiation was estimated as average allele frequency differences between *A. fraterculus* and *A. obliqua* using all ($\overline{D}_{\mathrm{CDS}}$), non-synonymous ($\overline{D}_{\mathrm{NS}}$) and synonymous ($\overline{D}_{\mathrm{S}}$) SNPs. $\overline{D}_{\mathrm{CDS}}$ (A), $\overline{D}_{\mathrm{NS}}$ (B) and $\overline{D}_{\mathrm{S}}$ (C) comparison among reproductive-biased, head-biased and unbiased genes expressed in male profiles. Comparison of $\overline{D}_{\mathrm{CDS}}$ (D), $\overline{D}_{\mathrm{NS}}$ (E) and $\overline{D}_{\mathrm{S}}$ (F) among reproductive-biased, head-biased and unbiased genes expressed in female profiles. * Holm corrected p-value of Wilcoxon rank sum test < 0.05. ** Holm corrected p-value of Wilcoxon rank sum test < 0.01.

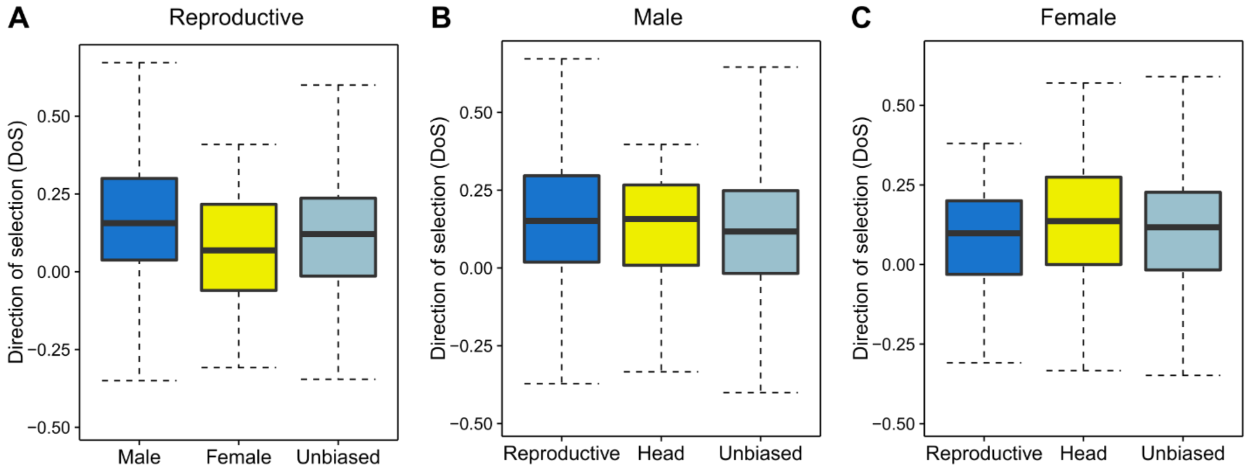


Figure S8.-Direction of selection (DoS) among sex-biased, tissue-biased and unbiased expressed unigenes estimated from *A. fraterculus* and *A. obliqua* SNPs. (A) DoS comparison among male-biased, female-biased and unbiased genes expressed in reproductive tissues. (B) DoS comparison among reproductive-biased, head-biased and unbiased genes expressed in male profiles. (C) DoS comparison among reproductive-biased, head-biased and unbiased genes expressed in female profiles.

CHAPTER III

PHYLOGENOMIC APPROACH REVEALS SIGNATURES OF INTROGRESSION AMONG NEOTROPICAL TRUE FRUIT FILES (*ANASTREPHA*: TEPHRITIDAE)

CHAPTER III – PHYLOGENOMIC APPROACH REVEALS SIGNATURES OF INTROGRESSION AMONG NEOTROPICAL TRUE FRUIT FILES (*ANASTREPHA*: TEPHRITIDAE)

3.1. Abstract

New sequencing techniques have allowed us to explore the variation of thousands of genes and elucidate the evolutionary forces involved in the differentiation process even in complex scenarios, such as in process of rapid radiation. That seems to be the case of *Anastrepha* species, which is a genus with notably species diversity and wide geographical distribution, which was divided into 21 species groups, being one of them *fraterculus* group. This species group includes several lineages that have diverged recently and likely in the presence of gene flow. Our main aim is to infer phylogenetic relationships among key lineages in the *fraterculus* group and their relationship to other closely related species groups. For that, we analyzed 20 RNA-seq libraries of female reproductive tissues from 10 different *Anastrepha* lineages. We used these transcriptomes to infer high-quality ortholog clusters in a phylogenetic framework to infer accurate phylogenies using multispecies coalescent methods. The phylogenetic analysis indicated that incomplete lineage sorting and introgression were important sources of gene tree discordance. Moreover, ABBA-BABA tests and species network analyses revealed an extensive pattern of gene flow among *fraterculus* group lineages, being most of them probably vestiges of ancestral introgression. Our findings help establish relationships among the most important *Anastrepha* species groups, as well as agrees with the hypothesis that the diversification of *fraterculus* group lineages such as *A. fraterculus* complex was influenced by interspecific gene flow.

3.2. Resumo

Tecnologias de sequenciamento de próxima geração têm permitido explorar a variação de milhares de genes e dilucidar as forças evolutivas envolvidas na diferenciação de espécies, mesmo em cenários complexos como acontece durante radiações rápidas. Este parece ser o caso das espécies do gênero *Anastrepha*, que apresenta uma notável diversidade de espécies e ampla distribuição geográfica que foram divididas em 21 grupos de espécies, sendo um deles o grupo *fraterculus*. Este grupo inclui linhagens que têm divergido recentemente num provável cenário com fluxo gênico. O objetivo principal é inferir as relações filogenéticas de linhagens chave do grupo *fraterculus* contextualizados com outras linhagens de grupos de espécies proximamente relacionados. Para tal, 20 bibliotecas de RNA-seq de tecido reprodutivo de fêmeas de 10 linhagens diferentes foram analisadas. Estes transcriptomas foram usados para inferir grupos de ortólogos de alta qualidade usando uma abordagem filogenética, os que foram utilizados como matéria prima para produzir acuradas filogenias usando *multispecies coalescent methods*. As análises filogenéticas indicaram presença de sorteamento incompleto de linheagens e hibridação como fontes de incongruência de árvores de genes. Além disso, testes ABBA-BABA e *network* de espécies revelaram sinais de extenso fluxo gênico entre linhagens do grupo *fraterculus*, sendo a maioria vestígios de introgressão ancestral. Nossos achados são concordantes com a hipótese de que a diversificação das linhagens do grupo *fraterculus* como por exemplo as do complexo *A. fraterculus* foram influenciadas por fluxo gênico interespecífico.

3.3. Introduction

Over the last three decades, evolutionary biologists have focused on trying to understand the speciation process in a biogeographical context, where genetic drift and physical barriers to gene flow are essential during population differentiation (MALLET, 2010). In this scenario, hybridization in secondary contact may act as a reinforcement of the differentiation favoring to break down gene flow (BARTON, 1979; ABBOTT et al., 2013). However, gene exchange between species that diverged in parapatry or sympatry can produce new allele combinations and introduce variation much faster than mutation (MALLET, 2007). If this evolutionary novelty has high adaptive value, because of divergent selection and/or sexual selection, they may foster differentiation (SERVEDIO, 2016). Furthermore, this mechanism is notably important during rapid adaptive radiation, in which introgression favors diversification (SEEHAUSEN, 2004). Some of the more iconic examples are host races in *Rhagoletis* fruit flies (FEDER et al., 2005), mimicry pattern of *Heliconius* butterflies (THE *HELICONIUS* GENOME CONSORTIUM, 2012), beak shape diversification of Darwin's finches (LAMICHHANEY et al., 2015) and vision pattern of cichlid fishes (MEIER et al., 2017).

Rapid adaptive radiations are characterized by retention of shared ancestral variation and possible interspecific gene flow (BERNER; SALZBURGER, 2015), which may cause two major sources of gene tree discordance: incomplete lineage sorting (ILS) and introgression (DEGNAN; ROSENBERG, 2009), which limit our ability to infer the most likely underlying species-tree to explain this discordance. New phylogenetic approaches such as the multispecies coalescent methods have been developed to accurately estimate species-trees using genomic data in very reasonable computational time (LIU et al., 2015). Most of these methods are very robust to deal with ILS, but extensive gene flow can generate inconsistent results (SOLÍS-LEMUS; YANG; ANÉ, 2016). In such cases, introgression events can be modeled in a species network as reticulated edges and associated with inheritance probability, that is the proportion of genes from the parental species transferred through the reticulation (YU; DEGNAN; NAKHLEH, 2012). However, phylogenetic network reconstruction under maximum likelihood may be computationally prohibitive for large-scale datasets, but recent methods based on pseudo-maximum likelihood have been shown to be computationally less expensive and produce robust results (YU; NAKHLEH, 2015; SOLÍS-LEMUS; ANÉ, 2016).

We explored these sources of tree discordance in *Anastrepha* Schiner (Diptera: Tephritidae) lineages, because it is a species-rich genus of fruit flies which includes several recently diverged species that makes it an interesting model for studies of speciation. This is the largest genus in tribe Toxotrypanini with almost 300 described species, which are distributed in tropical and subtropical regions of America (ALUJA, 1994; NORRBOM, Allen L. et al., 1999; NORRBOM; ZUCCHI; HERNÁNDEZ-ORTIZ, 1999; NORRBOM; KORYTKOWSKI, 2009; 2011; 2012; NORRBOM et al., 2015). *Anastrepha* has been divided into 21 species groups based on morphology and number of chromosomes (NORRBOM, Allen L. et al., 1999; NORRBOM et al., 2012 onwards). Although, clear morphological and behavioral (e.g. host preference and sexual activity) divergence among some species groups (ALUJA et al., 1999; NORRBOM, Allen L et al., 1999), phylogenetic analysis using mitochondrial 16S rDNA and nuclear *period* failed to establish robust relationships of these groups (MCPHERON et al., 1999; BARR; CUI; MCPHERON, 2005). In an unprecedented effort, Mengual et al. (2017) have recently analyzed six genes from 146 *Anastrepha* species, but identified that only seven species groups were monophyletic and most of the relationships among groups were poorly supported. These findings suggest that *Anastrepha* species may have experienced episodes of rapid radiations, coupled with large population sizes, which would explain this tremendous diversification with low phylogenetic resolution.

Among *Anastrepha* groups, we focus on *fraterculus* group with 34 described species, because it includes several recent diverged species such as the *A. fraterculus* complex (NORRBOM et al., 2012 onwards) . Closely related species such as *A. turpiniae*, *A. fraterculus* (*s.l.*) and *A. sororcula* are hard to identify based on morphology due to overlapping variation even in the aculeus (structure inside the ovipositor), which is a key trait in the systematics of this group (ZUCCHI, 2000b; PERRE et al., 2014). mtDNA phylogenies showed that most species in this group are not monophyletic, except *Anastrepha suspensa* (SMITH-CALDAS et al., 2001; BOYKIN et al., 2006; SCALLY et al., 2016). However, a phylogenetic inference based on nuclear markers indicated that *A. obliqua* is a monophyletic group (SCALLY et al., 2016). The mito-nuclear discordance suggested that these species evolved under complex scenarios retaining shared ancestral polymorphism due to a rapid radiation causing ILS (SILVA; BARR, 2008) and introgression (SCALLY et al., 2016; RULL et al., 2017; DÍAZ et al., 2017, submitted). Crosses in laboratory have demonstrated that closely related species of this group may generate viable and fertile hybrids (DOS SANTOS; URAMOTO; MATIOLI, 2001; RULL et al., 2017), suggesting that hybridization in nature is possible, since these species in many instances are found in sympatry.

An extreme example of the complexity of this group is *A. fraterculus*, which because of its large genetic and morphological variation and wide geographical distribution has been considered a species complex (STONE, 1942; SELIVON et al., 2004; SELIVON; PERONDINI; MORGANTE, 2005). Nowadays, morphological studies using samples distributed across America identified eight morphotypes in the *A. fraterculus* complex (HERNÁNDEZ-ORTIZ et al., 2012; HERNÁNDEZ-ORTIZ et al., 2015). Several analyses using diverse data such as cytogenetics, genetics, morphometrics, pheromone composition and reproductive incompatibilities support at least three Brazilian lineages of *A. fraterculus* complex (SELIVON et al., 2004; SELIVON et al., 2005; HERNÁNDEZ-ORTIZ et al., 2012; RULL et al., 2012; BŘÍZOVÁ et al., 2013; MANNI et al., 2015; DIAS et al., 2016), but differences among them are not so clear as for other lineages of the complex (HENDRICHS et al., 2015; VANÍČKOVÁ et al., 2015). Possibly because it is difficult to differentiate between intra- and inter-lineage variation in this complex, phylogenetic relationships among Brazilian lineages are still obscure.

In order to investigate phylogenetic relationships among some of the more relevant species groups in *Anastrepha*, we generated female transcriptomes from some of the main pest species present in Brazil, with a special focus on the *fraterculus* group. This dataset was used to generate the first phylogenomic inference among species from this genus, based on multispecies coalescent methods and investigate for gene tree discordance in terms of ILS and hybridization to contributes to a better understanding of the diversification patterns of this rich-species genus.

3.4. Material and Methods

3.4.1. Sampling and laboratory procedures

We established a sampling design aimed at collecting different fruits from a wide geographic area to sample key *Anastrepha* species from the most economically important species group and as many *fraterculus* group species as possible. For that we selected specific fruits that have previously been reported to be hosts of specific species such as *Inga* sp. to collect *A. distincta*, *Curcubita spp.* (pumpkin) to collect *A. grandis*, passiflora to collect *A. pseudoparallela*, sapote to collect *A. serpentina*, as well as several fleshy fruits which are reported host to a wide array of species such as guava (ZUCCHI, 2000a). Despite that, we collected a limited number of species (~15% of the known species) (Figure S1, Table S1 in Appendix 4), due to the fact that several *fraterculus* group species are collected mostly from traps, and have been found at very low local frequencies when sampled from fresh fruits. Since we can only use fresh specimens to perform RNA-seq experiments, the use of flies from traps or previous collections would be inadequate.

Infested fruits were collected from the field and in some cases individuals that emerged were used to establish populations in the laboratory, which were reared in the same conditions as in Rezende et al. (2016). Pupae from each fruit were separated in individual cages until eclosion when females were isolated and identified following the taxonomic key published by Arias et al. (2014). The aculeus and reproductive tissues (ovaries, accessory glands, spermatheca, uterus and ovipositor) were carefully removed from females of 8-15 days old and tissues were stored in 1.5ml tubes with TRIzol® Reagent at -80°C. RNA was isolated following a TRIzol/chloroform protocol (CHOMCZYNSKI; MACKEY, 1995). RNA-seq libraries were prepared using the TruSeq® Stranded mRNA Sample Prep LS Protocol (Illumina®) according to the manufacturer's instructions. Sequencing was performed using the HiSeq SBS v4 High Output Kit on Illumina HiSeq^{TM}2500 setting 2 x 125 bp (1/12 of lane per sample) or 2 x 100 bp (1/10 of lane per sample) paired-end reads cycles at the Laboratory of Functional Genomics Applied to Agriculture and Agri-energy, ESALQ-USP, Brazil. We also used one *A. fraterculus* transcriptome [AfraTUAR01 (SRR4026776)] and the *Ceratitis capitata* genome [GCF_000347755.2 (PAPANICOLAOU et al., 2016)], which were downloaded from Genbank.

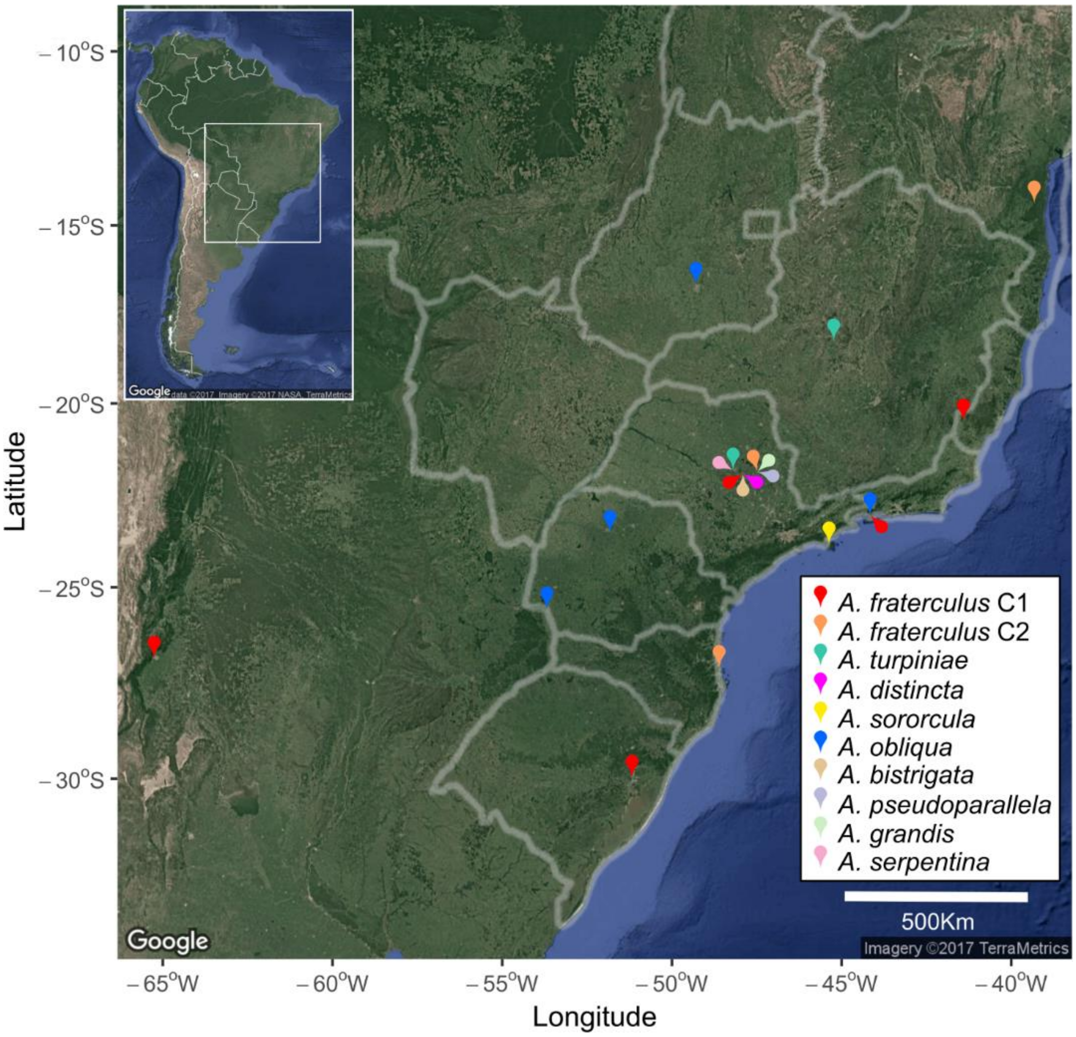


Figure 1. Distribution information of sampled *Anastrepha*.

3.4.2. Cleaning, *de novo* assembly and unigene prediction

The reads from each library were trimmed based on quality and removed any remaining TrueSeq Illumina adapters using the program Trimmomatic v. 0.35 (BOLGER; LOHSE; USADEL, 2014) (see parameters at Appendix 1). Filtered pair-end reads were normalized by k-mer counting and assembled setting -SS_lib_type RF for stranded samples and default parameters on Trinity package v. 2.3.2 (GRABHERR et al., 2011).

The putative coding sequences (CDSs) for each library were predicted using TransDecoder v. 3.0.1 (GRABHERR et al., 2011). We inferred the longest CDSs using TransDecoder.LongOrfs, which were annotated against Pfam and Unitprot90 databases using

HMMER v. 3.1b2 hmmscan (EDDY, 2011) and NCBI BLASTP (2.6.0) (CAMACHO et al., 2009), respectively. The final prediction of CDSs was performed by TransDecoder.Predict program. Redundancy of the obtained CDSs were reduced using a similarity threshold to 0.99 in the CD-Hit program (FU et al., 2012). Transcripts containing this set of CDSs were filtered based on the highest expressed isoform per trinity component. For that, the reads were mapped to the respective assembly using Bowtie2 (LANGMEAD; SALZBERG, 2012) and the abundances were estimated by eXpress v.1.5.1 (ROBERTS; PACHTER, 2013). These runs were carried out using the script align_and_estimate_abundance.pl included in the Trinity package, setting --very-sensitive option for Bowtie2 (LANGMEAD; SALZBERG, 2012), no bias correction for the eXpress program (ROBERTS; PACHTER, 2013) and default settings for the remaining parameters. The quality and completeness of the assemblies were assessed by BUSCO2 (Benchmarking Universal Single-Copy Orthologs) (SIMÃO et al., 2015) using 1,066 single-copy orthologs of Arthropoda from OrthoDB (ZDOBNOV et al., 2017).

3.4.3. Functional annotation

We used NCBI BLASTP (2.6.0) to align the translated putative CDSs against the non-redundant protein database of Arthropoda from Genbank (nr), the *Drosophila melanogaster* protein database (r6.14) and the Eukaryotic Orthologous Groups of proteins database (KOG) (KOONIN et al., 2004). Hits with e-value lower than 10^{-6} were considered as significant. Conserved protein domains were searched using InterProScan v. 5.23-62.0 (JONES et al., 2014). Blast2GO program (CONESA et al., 2005) was used to associate each annotated transcript with gene ontology (GO) terms.

3.4.4. Orthology inference and phylogenetic analysis

The orthology prediction of the putative CDSs was performed in a pipeline that uses a phylogenetic-based approach (YANG; SMITH, 2014). For that, The CDSs was submitted to all-by-all BLAST using NCBI BLASTN (2.6.0) (CAMACHO et al., 2009) setting relaxed parameters and filtered by MLC program (VAN DONGEN, 2000). These clusters of putative homologs were aligned using MAFFT v.7.305b (KATOH; STANDLEY, 2013), filtered by Phyutility v.2.2.6 (SMITH; DUNN, 2008) and then used to infer phylogenetic trees using RAxML v.8.2.9 (STAMATAKIS, 2014). The phylogenetic trees were filtered by pruning tip and internal taxa with long branches using the scripts trim_tips.py and cut_long_internal_branches.py available on https://bitbucket.org/yangya/phylogenomic_dataset_construction. A maximum inclusion approach was used to infer the set of orthologs (DUNN; HOWISON; ZAPATA, 2013). The row sequences of ortholog groups were divided into two sets: i) including *Anastrepha* sequences and using *C. capitata* as an outgroup; ii) including *fraterculus* group sequences, but using *A. bistrigata* as an outgroup. These two data sets were independently submitted to the POTION pipeline (HONGO et al., 2015) considering different parameters (see Appendix 2 and Appendix 3). This pipeline was used to re-align the sequences based on amino acid using MAFFT (KATOH; STANDLEY, 2013) and then remove proteins with high variation in length and percent of identity. Moreover, the alignments were trimmed using the option strictplus in trimAl v.1.2 (CAPELLA-GUTIÉRREZ; SILLA-MARTÍNEZ; GABALDÓN, 2009). Set of orthologs with signals of recombination detected by Phi, NSS or MaxChi2 methodologies implemented in PhiPack (BRUEN; PHILIPPE; BRYANT, 2006) were also removed. All parameters used in this pipeline are available in Appendix 1. The maximum likelihood (ML) phylogenies of each cluster of ortholog were inferred using GTRCAT model and 200 bootstrap replicates in the program RAxML v.8.2.9 (STAMATAKIS, 2014).

We produced a Densitree plot to visualize phylogenetic discordance among unigenes of species of the *fraterculus* group. For that, the ML trees were rooted using *A. bistrigata* as outgroup and converted to ultrametric trees using the APE module (PARADIS; CLAUDE;

STRIMMER, 2004) in R v.3.2.0 (R CORE TEAM, 2015). These trees were submitted to DensiTree (BOUCKAERT, 2010) setting the option star tree to produce the density tree.

We also inferred the phylogenetic relationships among *Anastrepha* samples using mtDNA Cytochrome c oxidase subunit I (COI), Cytochrome b (cytB), NADH dehydrogenase subunit 4 (ND4) and NADH dehydrogenase subunit 5 (ND5) and among specimens from the *fraterculus* group using internal transcribed spacer 1 (ITS1). BLASTn was used to compare the set of unigenes of the transcriptomes against the sequences of the complete mitochondrial genome of *A. fraterculus* Andean morphotype (NC_034912) (ISAZA; ALZATE; CANAL, 2017) and the Andean ITS1 lineages (SUTTON et al., 2015), with sequences having been selected based on the best hit criterium. These sequences were aligned using MAFFT (KATOH; STANDLEY, 2013) and manually trimmed. The ML trees were inferred using the same option in the program RAxML v.8.2.9 (STAMATAKIS, 2014).

3.4.5. Species-tree inference

We used two coalescent-based approaches to estimate the species tree of *Anastrepha* taxa using *C. capitata* as outgroup: ASTRAL-II v.4.10.12 (MIRARAB; WARNOW, 2015) and ASTRID (VACHASPATI; WARNOW, 2015). ASTRAL-II uses the quartet trees of the maximum likelihood phylogenies of each gene to produce the topology of the species tree and calculates the quartet support, which is the percentage of quartets that agree with a specific branch in the species-tree. ASTRID also takes the gene trees to estimate internode distances and infer the species tree. Our focus is the species of the *fraterculus* group, so we inferred the species coalescent tree of this group with *A. bistrigata* as outgroup using three methodologies: ASTRAL-II v.4.10.12 (MIRARAB; WARNOW, 2015), ASTRID (VACHASPATI; WARNOW, 2015) and SVDquartets (CHIFMAN; KUBATKO, 2014). The latter is implemented in PAUP* v4.0a156 and uses the information of unlinked *loci*, such as SNPs. For that, the reads of each sample were mapped to the transcriptome of *A. bistrigata* using --nofw and --very-sensitive options of Bowtie2 (LANGMEAD; SALZBERG, 2012). PCR duplicated reads were removed using MarkDuplicates tool included in Picard package (http://broadinstitute.github.io/picard/). The SNP calling was performed by mpileup tool of Samtools v.1.3.1 package (LI et al., 2009) and mpileup2snp tool of VarScan v2.4.2 (KOBOLDT et al., 2012). For that, we set the minimum mapping quality of 20, minimum PHRED average quality of 30, minimum coverage of 6 and strand filter (removed variants with more than 90% supported by only one strand). Bi-allelic SNPs detected in all of the samples (excluding sites with missing data) were used to build the SNP´s matrix as the input of SVDquartets+PAUP*. The tree was reconstructed using 100,000 quartets and 1,000 bootstrap replicates.

3.4.6. Introgression detection

To test for hybridization signals among the *Anastrepha* lineages, we performed ABBA-BABA tests (PATTERSON et al., 2012) using two data sets, which included all samples and a subset with only *A. fraterculus*' group samples, using *A. serpentina* and *A. bistrigata*, respectively, as outgroup and mapping reference. We performed these tests in Abbababa2 tool implemented by ANGSD (KORNELIUSSEN; ALBRECHTSEN; NIELSEN, 2014). Abbababa2 is a variant of the classic test that considers the possibility of more than one sample for each population in a dataset. ABBA-BABA test compares segregation of biallelic polymorphism of four samples (H1, H2, H3 and H4), being H1, H2 and H3 three ingroups and H4 an outgroup. Evidence of introgression was tested using the *D*-statistic, which is calculated as the deviations of proportions of shared alleles between H2 and H3 (ABBA), and H1 and H3 (BABA) following the formula (nABBA-nBABA)/(nABBA+nBABA) (DURAND et al., 2011). We also used this program to perform a jack-knife bootstrap approach to calculate a bias-corrected *D*-statistic, *D*-statistic standard error and Z-test (SORAGGI; WIUF; ALBRECHTSEN, 2017). For the latter test, p-values were corrected for multiple tests using the false discovery rate approach (BENJAMINI; HOCHBERG, 1995).

Since we found evidence of interspecific gene flow among *Anastrepha* lineages, we inferred phylogenetic networks to detect possible events of hybridization, which are visualized as reticulation in the network (HUSON; SCORNAVACCA, 2011; YU et al., 2012). The phylogenetic networks were inferred from 3,220 *Anastrepha* orthologs and 3,045 *A. fraterculus* groups and *A. bistrigata* orthologs using pseudo-maximum likelihood approach (YU; NAKHLEH, 2015) implemented in PHYLONET v. 3.6.1 (THAN; RUTHS; NAKHLEH, 2008). This program estimates the network topology and inheritance probability (γ), which is the hybridization probability considering the two parent population (YU et al., 2012). We built six networks varying the number of reticulation from 0 to 5 and setting the taxa association using the species-tree topology. The optimal network of each run was selected based on the highest likelihood after 500 searches. The likelihood of the best network for each run was also compared to choose the optimum number of reticulations in the network.

3.5. Results

3.5.1. Sequencing and assembly

The analyzed RNA samples had an average of 21.6 (± 2.5) x 2 million of raw reads. The filtering step removed an average of 6.7% of the reads, retaining an average of 20.3 (± 2.4) millions of pair-end reads (Table S2 in Appendix 4). The *de novo* assemblies produced more than 55,000 contigs per sample with N50 ranging from 1,466 to 2,524 nt with an average of 1,968 nt (Table S3 in Appendix 4). Furthermore, the quality assessment revealed that ~97% of Arthropoda ortholog clusters in BUSCO were found to be complete in the assemblies, but with high levels of redundancy (31% of duplicates, on average) (Table S4 in Appendix 4). In contrast, unigenes showed almost the same rate of complete orthologs, but the duplicated rates were drastically reduced (~1% of duplicates) (Table S4 in Appendix 4).

BLAST searches showed a high percent of unigenes with significant hits; depending on the species assembly at least 75% were annotated with the *D. melanogaster* protein database and 92% with nr Genbank Arthropoda database (Table S5 in Appendix 4). Likewise, on average 91% (ranging from 87 and 93%) and 71% (with a range of 67-73%) of the unigene CDSs were annotated using Interproscan and Blast2GO, respectively (Table S5 in Appendix 4). Furthermore, the functional annotation revealed similar distribution of GO terms and KOG categories for all studied *Anastrepha* lineages (Figure S1 and Figure S2 in Appendix 4).

3.5.2. Phylogenomic inference

mtDNA maximum likelihood phylogeny showed few clades with robust support, among them the *fraterculus* group lineage and *A. pseudoparallela* as the most closely related species to the *fraterculus* group (Figure 2A). Three different clades were identified in the *fraterculus* group; the first included specimens from *A. fraterculus*, *A. obliqua*, *A. sororcula* and *A. turpiniae*, the second included only individuals from *A. fraterculus*, and the third with *A. distincta*. In contrast, a species-tree based on 1,658 genes showed all nodes with high bootstrap values (>90%) and *A. bistrigata* from the *striata* group as sister lineage to the *fraterculus* group (Figure 2B). The relationship among *fraterculus* group specimens including all samples is the same as the one produced only with *fraterculus* group species using *A. bistrigata* as outgroup based on 3,045 ortholog clusters (ASTRAL-II and ASTRID) and 827,256 SNPs (SVDquartets+PAUP*) (Figure 3A). This species-tree revealed that *A. obliqua* and *A. turpiniae* formed individual lineages and there were also two strongly supported *A. fraterculus* lineages, which hereafter will be referred to as *A. fraterculus* C1, for the lineage that has ITS1 lineage TI/Tia (Figure S3), and *A. fraterculus* C2, which formed a separate group related to ITS1 lineage TII (Sutton et al., 2015). ASTRAL-II and ASTRID approaches showed exactly the same topology, but as SVDquartets+PAUP* methodology was performed using biallelic data, relationships inside the lineages were not comparable.

ASTRAL produced the final topology based on 5,029,080 quartets and 69% of them were found in the inferred species tree. Furthermore, incongruency levels between gene trees and species trees measured by quartets support and density tree plot (Figure 2B and Figure 3), indicated robust support for some species clades, such as *A. obliqua* and *A. turpiniae*, as well as great support for *A. fraterculus* complex. In contrast, other clades have lower support, such as the separation between *A. grandis* and *A. pseudoparalella*, or the separation between *A. fraterculus* C2, *A. turpiniae* and *A. distincta* from *A. fraterculus* C1. Densitree plot also evidenced strong signal for gene tree discordance among genes (Figure 3B).

3.5.3. Detection of introgression among *Anastrepha* lineages

ABBA-BABA tests for all possible triplet combinations among *fraterculus* lineages concordant with the species-tree topology showed signals for introgression in almost all combinations (Table 1). Only two ingroups failed to be significant in both references, both of which involved *A. distincta* ([*A. distincta*, *A. turpiniae*], *A. obliqua* and [*A. distincta*, *A. fraterculus* C1], *A. sororcula*). Furthermore, *D*-statistic calculated using these two outgroups were strongly correlated (Pearson´s correlation: $p < 2.2e^{-16}$, R = 0.997). Likewise, comparisons among lineages from different *Anastrepha* species groups revealed that all possible combinations displayed statistically significant ABBA-BABA tests, which may indicate introgression even between distantly related lineages (Table 2). Moreover, ABBA-BABA tests involving two lineages of *fraterculus* group (as H1 and H2) and other from other species group displayed 14 out of 45 (31.1%) combinations with signals of introgression (Table S6 in Appendix 4).

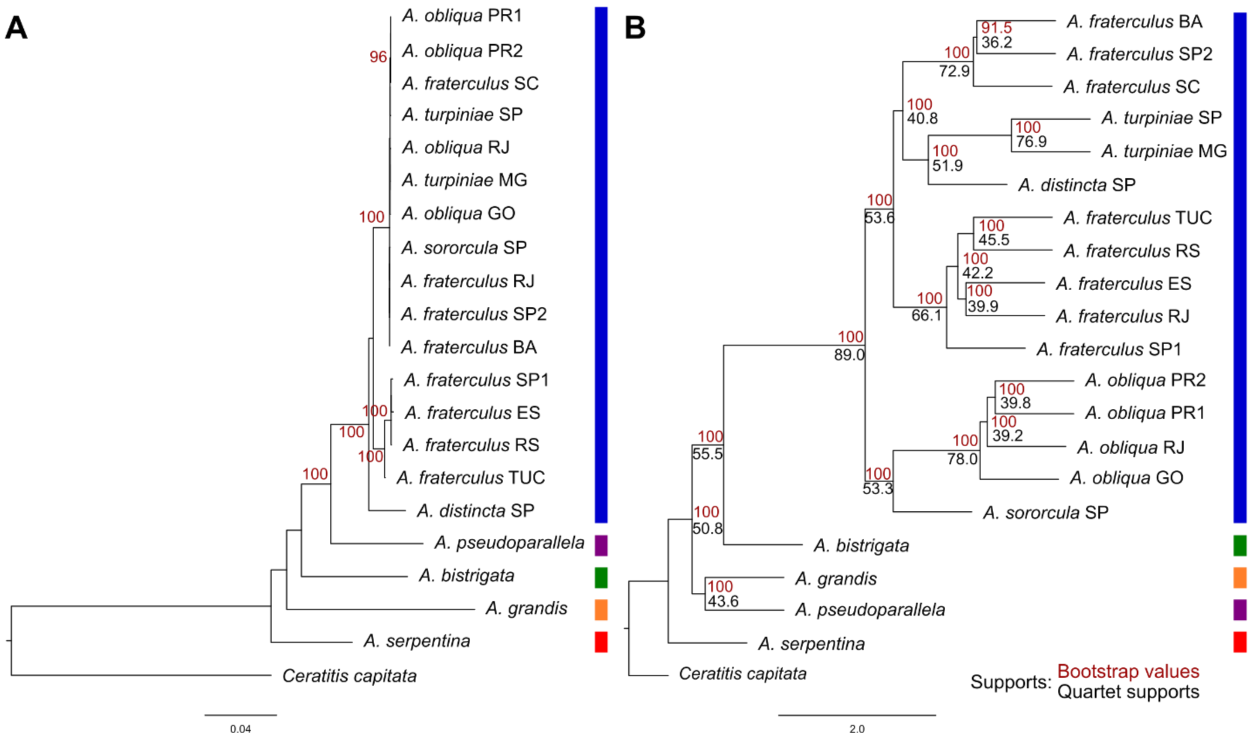


Figure 2. Phylogenetic inference of *Anastrepha* species using mtDNA and nuclear genes. (A) Phylogenetic tree estimated by RAxML for four partial sequences of mtDNA genes (COI, Cytb, ND4 and ND5). (B) Species-tree inference performed by ASTRAL-II and ASTRID for 1,658 nuclear genes. Bootstrap supports higher than 90% are shown in red. Quartet supports estimated by ASTRAL-II are shown below to bootstrap supports in the species-tree. Each color indicates samples from different *Anastrepha* species group: Blue, *fraterculus* group; green, *striata* group; orange, *grandis* group; purple, *pseudoparallela* group; and red, *serpentina* group. Sampling information is available at Table S1 in Appendix 4.

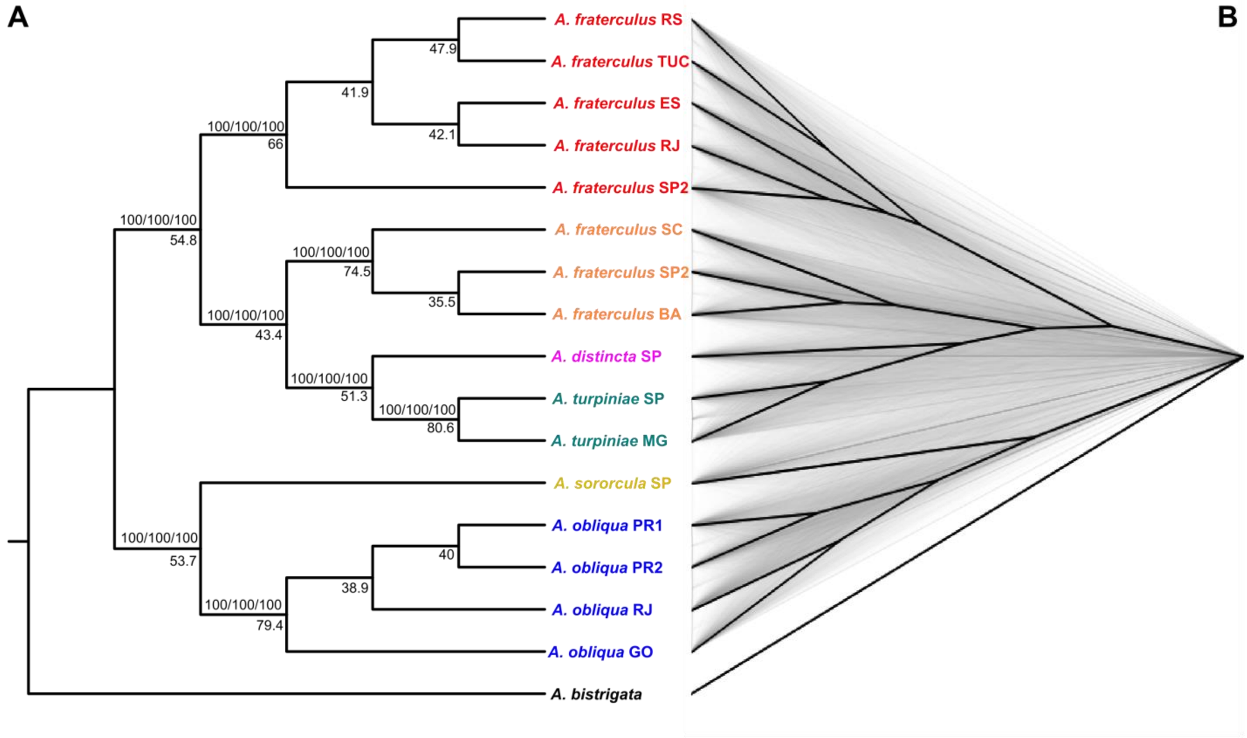


Figure 3. Species-tree inference and density tree plot of *fraterculus* group lineages using *A. bistrigata* as outgroup. (A) ASTRAL-II and ASTRID approaches were performed based on 3,045 nuclear genes and SVDquartets+PAUP* based on 827,256 SNPs. Bootstrap values of concordant topologies among methodologies are shown above the nodes. Quartet supports estimated by ASTRAL-II are shown below the nodes in the species-tree. (B) Densitree plot for 3,045 nuclear genes using *A. bistrigata* as outgroup. Each color indicates samples from different *fraterculus* group lineage: Red, *A. fraterculus* C1; orange, *A. fraterculus* C2; pink, *A. distincta*; green, *A. turpiniae*; yellow, *A. sororcula*; and blue, *A. obliqua*. Sampling information is available at Table S1 in Appendix 4.

Table 1. ABBA-BABA tests among *fraterculus* group lineages.

			A. bistrigata[a]			***A. serpentina***[a]		
H1[b]	**H2**[b]	**H3**[b]	***D*-statistic ±SE**[c]	**Z**	**q-value**	***D*-statistic ±SE**[c]	**z**	**q-value**
Adis	Atur	Afra2	0.040079 ±0.000035	6.78	0.000000	0.030842 ±0.000033	5.36	0.000000
Adis	Atur	Afra1	0.030345 ±0.000029	5.62	0.000000	0.021661 ±0.000029	4.00	0.000289
Adis	Atur	Aobl	0.012198 ±0.00004	1.93	0.055796	0.003682 ±0.000039	0.59	1.000000
Adis	Atur	Asor	0.058349 ±0.000043	8.88	0.000000	0.048293 ±0.000043	7.36	0.000000
Afra2	Atur	Afra1	0.049709 ±0.000037	8.22	0.000000	0.051297 ±0.000034	8.74	0.000000
Afra2	Atur	Aobl	0.174788 ±0.000038	28.18	0.000000	0.164693 ±0.000036	27.32	0.000000
Afra2	Atur	Asor	0.16163 ±0.000048	23.40	0.000000	0.157949 ±0.000044	23.72	0.000000
Afra2	Adis	Afra1	0.029532 ±0.000042	4.56	0.000006	0.038848 ±0.000039	6.21	0.000000
Afra2	Adis	Aobl	0.167544 ±0.000045	24.88	0.000000	0.163459 ±0.000042	25.22	0.000000
Afra2	Adis	Asor	0.12016 ±0.000056	16.08	0.000000	0.122249 ±0.000051	17.13	0.000000
Atur	Afra1	Aobl	-0.03514 ±0.000028	-6.66	0.000000	-0.032212 ±0.000026	-6.33	0.000000
Atur	Afra1	Asor	-0.05071 ±0.000035	-8.52	0.000000	-0.051052 ±0.000034	-8.71	0.000000
Adis	Afra1	Aobl	-0.02701 ±0.000035	-4.56	0.000006	-0.026373 ±0.000033	-4.61	0.000020
Adis	Afra1	Asor	-0.00611 ±0.000042	-0.95	0.344336	-0.011993 ±0.000039	-1.93	0.237601
Afra2	Afra1	Aobl	0.12969 ±0.000033	22.74	0.000000	0.127718 ±0.00003	23.44	0.000000
Afra2	Afra1	Asor	0.108472 ±0.000037	17.92	0.000000	0.104207 ±0.000034	17.94	0.000000
Aobl	Asor	Afra2	0.190585 ±0.000042	29.34	0.000000	0.191922 ±0.000039	30.66	0.000000
Aobl	Asor	Adis	0.141442 ±0.000053	19.43	0.000000	0.150838 ±0.000051	21.04	0.000000
Aobl	Asor	Atur	0.169981 ±0.000043	25.92	0.000000	0.178491 ±0.000042	27.55	0.000000
Aobl	Asor	Afra1	0.160009 ±0.000032	28.07	0.000000	0.163878 ±0.000031	29.34	0.000000

[a] *A. bistrigata* and *A. serpentina* were used as outgroups and mapping references.

[b] Lineage name abbreviations are as follows: Adis, *A. distincta*; Atur, A. turpiniae; Afra1, *A. fraterculus* C1; Afra2, *A. fraterculus* C2; Asor, *A. sororcula*; and Aobl, *A. obliqua*.

[c] Average *D*-statistic and standard error (SE) were calculated by m-delete blocked Jackknife approach in ANGSD.

Table 2. ABBA-BABA tests among *Anastrepha* species groups.

			***A. serpentina*[a]**		
H1[b]	**H2[b]**	**H3[b]**	***D*-statistic ±SE[c]**	**z**	**q-value**
Agra	Apse	Abis	0.026291 ±0.000036	4.35	0.000057
Agra	Apse	Aobl	0.038824 ±0.000034	6.66	0.000000
Agra	Apse	Asor	0.031314 ±0.000036	5.25	0.000000
Agra	Apse	Afra2	0.030461 ±0.000033	5.32	0.000000
Agra	Apse	Adis	0.032705 ±0.000035	5.51	0.000000
Agra	Apse	Atur	0.031468 ±0.000033	5.47	0.000000
Agra	Apse	Afra1	0.031688 ±0.000032	5.59	0.000000
Abis	Aobl	Agra	-0.04706 ±0.00003	-8.60	0.000000
Abis	Aobl	Apse	-0.02992 ±0.000031	-5.34	0.000000
Abis	Asor	Agra	-0.041887 ±0.000034	-7.20	0.000000
Abis	Asor	Apse	-0.035152 ±0.000036	-5.87	0.000000
Abis	Afra2	Agra	-0.048885 ±0.00003	-8.97	0.000000
Abis	Afra2	Apse	-0.044043 ±0.000031	-7.93	0.000000
Abis	Adis	Agra	-0.043353 ±0.000034	-7.43	0.000000
Abis	Adis	Apse	-0.033086 ±0.000034	-5.66	0.000000
Abis	Atur	Agra	-0.044576 ±0.000031	-7.97	0.000000
Abis	Atur	Apse	-0.036517 ±0.000032	-6.47	0.000000
Abis	Afra1	Agra	-0.047699 ±0.000029	-8.82	0.000000
Abis	Afra1	Apse	-0.039878 ±0.000029	-7.38	0.000000

[a] *A. serpentina* was used as outgroup and mapping reference.
[b] Lineage name abbreviations are as follows: Agra, *A. grandis*; Apse, *A. pseudoparallela*; Abis, *A. bistrigata*; Adis, *A. distincta*; Atur, *A. turpiniae*; Afra1, *A. fraterculus* C1; Afra2, *A. fraterculus* C2; Asor, *A. sororcula*; and Aobl, *A. obliqua*.
[c] Average *D*-statistic and standard error (SE) were calculated by m-delete blocked Jackknife approach in ANGSD.

The optimum number of reticulations in species networks for all *Anastrepha* lineages and for *fraterculus* complex lineages (*A. bistrigata* as outgroup) datasets inferred by PHYLONET showed that higher number of reticulations produced networks with more log pseudo-likelihood, but exhibited a plateau starting at three reticulations so we chose this value as optimal (Figure S4 in Appendix 4). Networks with zero reticulations displayed the same topologies as the inferred species-trees (Figures 2B, Figure 3A and Figure S5 in Appendix 4). Optimum species-networks showed that most reticulate edges have high inheritance probability (~0.3) except the reticulation of *A. fraterculus* C2 and the ancestral branch of *fraterculus* group and *A. bistrigata* (0.058) (Figure 4). Moreover, most reticulations involved one internal branch.

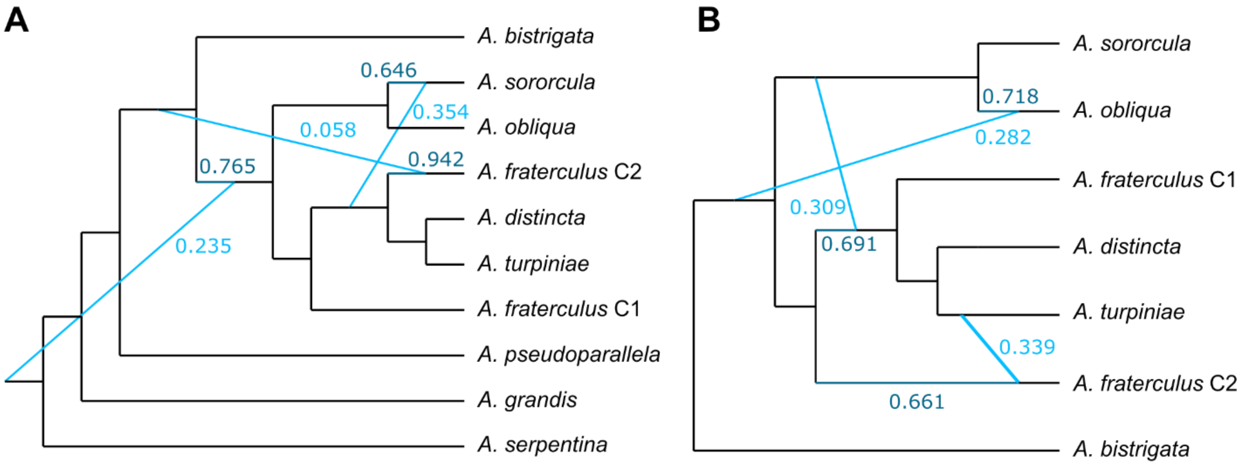


Figure 4. Optimum species networks from *Anastrepha* lineages inferred by pseudo-maximum likelihood approach. Networks were inferred allowing for three reticulations based on 3,220 nuclear genes from all *Anastrepha* lineages (A) and 3,045 nuclear genes from *fraterculus* group specimens and *A. bistrigata* as an outgroup (B). Inheritance probabilities (γ) are shown in sky-blue.

3.6. Discussion

The female reproductive transcriptomes of several *Anastrepha* species here produced have very similar quality, measured by standard metrics that report length distribution in *de novo* assemblies, such as average, median and N50, and comparable to what has been described in transcriptomes from *A. obliqua* and *A. fraterculus* (Rezende et al., 2016; Congrains et al., 2017, submitted) and other Tephritidae species (Salvemini et al., 2014; Wang; Xiong; Liu, 2016; Zheng et al., 2016). However, length distribution metrics like N50 may give a wrong idea of completeness, especially in transcriptome data, so we used BUSCO to evaluate the gene content of the assemblies. This program revealed that a very low percentage of conserved Arthropoda ortholog clusters were missing (<2.5%) or fragmented (<2.3%) in the raw transcriptomes. Furthermore, BUSCO was also useful to evaluate the efficiency of methodology to obtain unigenes, which produced low percentages of duplicates conserved orthologs (< 1.35%) without greatly increasing missing orthologs (>4%). Since BUSCO was designed for genomic analyses, such a low rate of missing orthologs on transcriptomes, along with the high percentages of annotated CDSs, should be a good indication of the high quality and completeness of the data and assembly here produced.

The quality of the reproductive produced transcriptomes allowed for the definition of a cluster with thousands of orthologs across several species of *Anastrepha* which enabled us to infer their evolutionary histories using mitochondrial and nuclear genes. mtDNA phylogeny based on four genes displayed poorly resolved relationships among *Anastrepha* species groups, which is the same pattern as previously published phylogenies using mtDNA or a very limited number of nuclear markers (McPheron et al., 1999; Barr et al., 2005; Silva; Barr, 2008; Mengual et al., 2017). Moreover, this phylogeny also failed to resolve relationships among *fraterculus* group lineages, showing a large clade that included almost all lineages, except four individuals from *A. fraterculus* C1 (putatively *A. fraterculus* Brazil 1) and *A. distincta,* which is a similar pattern found in previous *fraterculus* group mtDNA phylogenetic inferences (Smith-Caldas et al., 2001; Ludeña; Bayas; Pintaud, 2010; Scally et al., 2016). In contrast, our genome-wide approach showed strong support of different species groups and their relationships, as well as different lineages that represent separate species in the *fraterculus* group. In addition, the phylogenetic positions of *Anastrepha* groups such as the *pseudoparallela* group that was closely related to the *grandis* group and this clade as sister of the *serpentina* group agrees with a previously published phylogeny (Mengual et al., 2017). Furthermore, when the phylogeny of *fraterculus* groups is compared with the inference published by Scally et al. (2016), both showed well resolved tree for *fraterculus* group lineages, but we found that *A. distincta* diverged recently, being closely related to *A. turpiniae*, different from what was previously reported (Scally et al., 2016), which may be due to the high level of phylogenetic discordance among genes. Considering that in this study, we applied multispecies coalescent methods that take into account ILS (Mirarab; Warnow, 2015) and introgression (Yu; Nakhleh, 2015), and we also used thousands of genes, we believe that the presented phylogeny would probably better reflect the phylogenetic relationships among these *Anastrepha* lineages.

Within *A. fraterculus* group lineages, we found at least two separate clades of *A. fraterculus,* which agrees with the existence of cryptic species in the *A. fraterculus* complex, three of which are supposed to be present in Brazil (reviewed in VANÍČKOVÁ et al., 2015). Even though there may be some markers that could be effective at separating these species, it is more likely that only by applying integrative taxonomy across several populations and separate morphological attributes might we be able to effectively distinguishing these species (SCHUTZE et al., 2017). For that reason, it may not be trivial to assign individuals collected in the field to one of these cryptic species. Since it has been suggested that variation in ITS1 might be informative to separate some taxa in the *A. fraterculus* complex (SUTTON et al., 2015), we inferred phylogenetic relationships among ITS1 sequences from our samples along with other sequences deposited in the Genbank, including the characteristic sequences of lineages reported by Sutton et al. (2015). This phylogeny allowed us to confirm that *A. fraterculus* C1 is probably *A. fraterculus* Brazil 1. Though *A. fraterculus* C2 position was more complicated, since it is more closely related to the TIII group, which seems to be a mélange of different populations, including TIV group (SUTTON et al., 2015). Since the ITS1 phylogeny displayed two different *A. fraterculus* clades, which corresponded to the species-tree inferred by thousands of genes, it is possible that this marker could be useful to separate some lineages in the *A. fraterculus* complex, as suggested by Sutton et al. (2015), though care should be exercised in this endeavor. Previous studies have suggested that the genetic structure of *A. fraterculus* lineages, even among Brazilian lineages, would be influenced by different selective pressures affecting populations at different altitudes (SMITH-CALDAS et al., 2001; MANNI et al., 2015). However, both lineages here reported are widely and sympatrically distributed across

Brazil without clear association to differences in altitudes, or hosts, for that matter. Thus, the differentiation of these lineages may be caused by other factors than ecological divergence.

ABBA-BABA tests among *fraterculus* group lineages suggest extensive interspecific gene flow. However, statistically significances on this test should be carefully interpreted, because alternative explanations to ongoing gene flow are also possible, particularly when species diverged under complex scenarios. For instance, this test may result in false positives when introgression involves species which were not sampled or are already extinct, and some related species could retain the signal of introgression (DURAND et al., 2011; EATON et al., 2015). Furthermore, ancestral subdivision may also produce significant outcome evaluated by ABBA-BABA test, so the test cannot differentiate between introgression and ancestral subdivision (DURAND et al., 2011). Nevertheless, our results showed robust signals for introgression, producing *D*-statistics values greater than 0.1 in ten combinations tested. These values are relatively high when compared to other insect genera with interspecific gene flow such as *Heliconius* (ZHANG et al., 2016) or *Papilio* (ZHANG; KUNTE; KRONFORST, 2013) butterflies, which may indicate that the signals are not only a reminiscent of other hybridization events. Moreover, extensive gene flow has also been detected in a recent study focusing on *A. fraterculus*, *A. obliqua* and *A. sororcula* (DÍAZ et al., 2017, submitted). Species network analyses also detected signals of introgression as visualized in the reticulations. Mito-nuclear discordance agrees with some level of gene flow among *A. fraterculus*, *A. obliqua*, *A. turpiniae* and *A. sororcula*. This pattern was found in previous mitochondrial phylogenies and possible gene flow even among more distantly related species such as *A. schultzi* was also reported (SCALLY et al., 2016). These genetic findings are also supported by laboratory mating experiments that demonstrated possible hybrids among some *A. fraterculus* complex lineages (VERA et al., 2006; CÁCERES et al., 2009; RULL et al., 2013; DEVESCOVI et al., 2014; DIAS et al., 2016; RORIZ; JAPYASSÚ; JOACHIM-BRAVO, 2017) and even among more divergent species such as *A. fraterculus* lineages with *A. obliqua* and *A. sororcula* (DOS SANTOS et al., 2001; RULL et al., 2017). Even though these are forced crosses limited by space (cages) and they show various levels of assortative mating (reviewed in JUÁREZ et al., 2015) and reproductive isolation, it is probable that hybridization would occur in nature. Therefore, all these data suggest extensive signals of interspecific gene flow among *fraterculus* lineages.

Another possibility is that *D*-statistics fails to discriminate between recent and ancestral pattern of introgression (DURAND et al., 2011; OTTENBURGHS et al., 2017). Both optimum species networks showed that most signals for introgression involved ancestral lineages, except for the reticulation between *A. turpiniae* and *A. fraterculus* C2. Differences in the phylogenetic positions of *A. distincta* in mtDNA and nuclear phylogenies may indicate ancestral introgression or asymmetrical gene flow, being the former the most likely explanation due to differences in host preferences and morphology (aculeus tip) (ZUCCHI, 2000b; a). In contrast, the pattern found in mtDNA from *A. fraterculus* C2, *A. obliqua*, *A. turpiniae* and *A. sororcula* agrees with recurrent gene flow among these lineages.

Contrasting with the gene flow pattern among *fraterculus* group lineages, results of ABBA-BABA tests among more distantly related species, such as lineages from distinct species groups are considered less robust. Since recurrent mutations are more likely to occur among more distantly related species, this bias in the mutation rate may result in high rates of false positive inferences of introgression in this test (DURAND et al., 2011). Furthermore, species from different *Anastrepha* groups are morphologically and ecologically distinct (NORRBOM, Allen L et al., 1999) and there is no evidence of possible hybrids between them, which makes the hypothesis for ongoing gene flow among groups unlikely. However, the network inferred on all samples showed a reticulation between the ancestral branches of *fraterculus* group and all samples from *Anastrepha* species with an inheritance probability of 0.23, which may indicate a signal for ancient introgression among groups.

The phylogenetic history of *Anastrepha* species shows signals of rapid radiations in several clades (MENGUAL et al., 2017), which according to our results may have been diverging with gene flow. Speciation with gene flow has been characterized in the apple maggot *Rhagoletis pomonella* (Tephritidae), with their races diverging rapidly mainly due to host shifts (FEDER et al., 2003). An experimental study indicated that differences in allele frequencies across the genome in response to divergent selection (host shifts) are detectable in only one generation (EGAN et al., 2015). This model of rapid diversification with gene flow due to ecological divergence might explain the ancestral divergence among more distantly related species that adapted to different hosts such as *A. grandis*, *A. pseudoparallela* and *A. serpentina*, which prefer plants from Cucurbitaceae, Passifloraceae and Sapotaceae, respectively (ZUCCHI, 2000a). However, several lineages from the *fraterculus* group, even though may show some host preference, are generalists (ZUCCHI, 2000a), so the host-races model would be an unlikely explanation of its diversification. Alternatively, differentiation fostered by sexual selection can drive to speciation even when the divergent lineages are in sympatry and without strong ecological differences (M'GONIGLE et al., 2012). Studies of molecular evolution in *Anastrepha* found signals of positive selection in genes related to reproduction, which may indicate sexual selection acting on these genes (SOBRINHO; DE BRITO, 2010; 2012; CONGRAINS et al., 2017, submitted). Moreover, cross experiments between morphotypes of the *A. fraterculus* complex showed that hybrid females prefer to mate with hybrid male (SEGURA et al., 2011), which is another evidence for the key roles of introgression and sexual selection in the speciation of these taxa. However, this is a possible explanation for the diversification of the generalist *A. fraterculus* complex, but other closely related species with more specialist habits such as *A. distincta* may be differentiated due to host shifts that is boosted by divergent ecological selection.

3.7. Conclusions

Here we produced reproductive transcriptomes for 10 evolutionary lineages of *Anastrepha*, the majority of which are considered as major agricultural pests from South America. Our approach produced a high-quality cluster of orthologs which was used to accurately reconstruct phylogenetic relationships among these *Anastrepha* lineages applying multispecies coalescent methods. Our genome-wide approach not only generated a very well-supported species-tree, but also explored whether ILS and introgression could be likely sources of species-tree vs. gene-tree discordance. We detected extensive signals for introgression especially among *fraterculus* lineages, which may involve the diversification of generalist species such as *A. fraterculus* complex.

3.8. References

ABBOTT, R. et al. Hybridization and speciation. **Journal of Evolutionary Biology,** v. 26, n. 2, p. 229-246, 2013.

ALUJA, M. Bionomics and management of *Anastrepha*. **Annual review of entomology,** v. 39, n. 1, p. 155-178, 1994.

ALUJA, M. et al. Behavior of flies in the genus *Anastrepha* (Trypetinae: Toxotrypanini). In: ALUJA, M. e NORRBOM, A. L. (Ed.). **Fruit flies (Tephritidae): phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 15, p.375-406.

ARIAS, O. R. et al. Fruit flies of the genus *Anastrepha* (Diptera: Tephritidae) from some localities of Paraguay: New records, checklist, and illustrated key. **Journal of Insect Science,** v. 14, n. 1, p. 224-224, 2014.

BARR, N. B.; CUI, L.; MCPHERON, B. A. Molecular systematics of nuclear gene *period* in genus *Anastrepha* (Tephritidae). **Annals of the Entomological Society of America,** v. 98, n. 2, p. 173-180, 2005.

BARTON, N. H. The dynamics of hybrid zones. **Heredity,** v. 43, p. 341, 1979.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society. Series B (Methodological),** v. 57, n. 1, p. 289-300, 1995.

BERNER, D.; SALZBURGER, W. The genomics of organismal diversification illuminated by adaptive radiations. **Trends in Genetics,** v. 31, n. 9, p. 491-499, 2015.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114–2120, 2014.

BOUCKAERT, R. R. DensiTree: making sense of sets of phylogenetic trees. **Bioinformatics,** v. 26, n. 10, p. 1372-1373, 2010.

BOYKIN, L. M. et al. Analysis of host preference and geographical distribution of *Anastrepha suspensa* (Diptera: Tephritidae) using phylogenetic analyses of mitochondrial cytochrome oxidase I DNA sequence data. **Bulletin of Entomological Research,** v. 96, n. 05, p. 457-469, 2006.

BŘÍZOVÁ, R. et al. Pheromone analyses of the *Anastrepha fraterculus* (Diptera: Tephritidae) cryptic species complex. **Florida Entomologist,** v. 96, n. 3, p. 1107-1115, 2013.

BRUEN, T. C.; PHILIPPE, H.; BRYANT, D. A simple and robust statistical test for detecting the presence of recombination. **Genetics,** v. 172, n. 4, p. 2665-2681, 2006.

CÁCERES, C. et al. Incipient speciation revealed in *Anastrepha fraterculus* (Diptera; Tephritidae) by studies on mating compatibility, sex pheromones, hybridization, and cytology. **Biological Journal of the Linnean Society,** v. 97, n. 1, p. 152-165, 2009.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics,** v. 10, n. 1, p. 421, 2009.

CAPELLA-GUTIÉRREZ, S.; SILLA-MARTÍNEZ, J. M.; GABALDÓN, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics,** v. 25, n. 15, p. 1972-1973, 2009.

CHIFMAN, J.; KUBATKO, L. Quartet inference from SNP data under the coalescent model. **Bioinformatics,** v. 30, n. 23, p. 3317-3324, 2014.

CHOMCZYNSKI, P.; MACKEY, K. Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide-and proteoglycan-rich sources. **Biotechniques,** v. 19, n. 6, p. 942-945, 1995.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics,** v. 21, n. 18, p. 3674-3676, 2005.

CONGRAINS, C. et al. Evidence of adaptive evolution and relaxed constraints in sex-biased genes of South American and West Indies fruit flies (Diptera: Tephritidae). **Genome Biology and Evolution**, 2017, submitted.

DEGNAN, J. H.; ROSENBERG, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. **Trends in Ecology & Evolution,** v. 24, n. 6, p. 332-340, 2009.

DEVESCOVI, F. et al. Ongoing speciation within the *Anastrepha fraterculus* cryptic species complex: the case of the Andean morphotype. **Entomologia Experimentalis et Applicata,** v. 152, n. 3, p. 238-247, 2014.

DIAS, V. S. et al. An integrative multidisciplinary approach to understanding cryptic divergence in Brazilian species of the *Anastrepha fraterculus* complex (Diptera: Tephritidae). **Biological Journal of the Linnean Society,** v. 117, n. 4, p. 725-746, 2016.

DÍAZ, F. et al. Impact of agricultural activities on introgression and population expansion of three species of the *Anastrepha fraterculus* group, a radiating species complex of fruit flies. **Evolutionary Applications**, 2017, submitted.

DOS SANTOS, P.; URAMOTO, K.; MATIOLI, S. R. Experimental hybridization among *Anastrepha* species (Diptera: Tephritidae): Production and morphological characterization of F1 hybrids. **Annals of the Entomological Society of America,** v. 94, n. 5, p. 717-725, 2001.

DUNN, C.; HOWISON, M.; ZAPATA, F. Agalma: an automated phylogenomics workflow. **BMC Bioinformatics,** v. 14, n. 1, p. 330, 2013.

DURAND, E. Y. et al. Testing for ancient admixture between closely related populations. **Molecular Biology and Evolution,** v. 28, n. 8, p. 2239-2252, 2011.

EATON, D. A. R. et al. Historical introgression among the American live oaks and the comparative nature of tests for introgression. **Evolution,** v. 69, n. 10, p. 2587-2601, 2015.

EDDY, S. R. Accelerated profile HMM searches. **PLOS Computational Biology,** v. 7, n. 10, p. e1002195, 2011.

EGAN, S. P. et al. Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. **Ecology Letters,** v. 18, n. 8, p. 817-825, 2015.

FEDER, J. L. et al. Allopatric genetic origins for sympatric host-plant shifts and race formation in *Rhagoletis*. **Proceedings of the National Academy of Sciences,** v. 100, n. 18, p. 10314-10319, 2003.

FEDER, J. L. et al. Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. **Proceedings of the National Academy of Sciences,** v. 102, n. suppl 1, p. 6573-6580, 2005.

FU, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics,** v. 28, n. 23, p. 3150-3152, 2012.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature biotechnology,** v. 29, n. 7, p. 644-652, 2011.

HENDRICHS, J. et al. Resolving cryptic species complexes of major tephritid pests. **ZooKeys,** v. 540, 2015.

HERNÁNDEZ-ORTIZ, V. et al. Taxonomy and phenotypic relationships of the *Anastrepha fraterculus* complex in the Mesoamerican and Pacific Neotropical dominions (Diptera, Tephritidae). **ZooKeys,** v. 540, 2015.

HERNÁNDEZ-ORTIZ, V. et al. Cryptic species of the *Anastrepha fraterculus* complex (Diptera: Tephritidae): a multivariate approach for the recognition of South American morphotypes. **Annals of the Entomological Society of America,** v. 105, n. 2, p. 305-318, 2012.

HONGO, J. A. et al. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. **BMC Genomics,** v. 16, n. 1, p. 567, 2015.

HUSON, D. H.; SCORNAVACCA, C. A survey of combinatorial methods for phylogenetic networks. **Genome Biology and Evolution,** v. 3, p. 23-35, 2011.

ISAZA, J. P.; ALZATE, J. F.; CANAL, N. A. Complete mitochondrial genome of the Andean morphotype of *Anastrepha fraterculus* (Wiedemann) (Diptera: Tephritidae). **Mitochondrial DNA Part B,** v. 2, n. 1, p. 210-211, 2017.

JONES, P. et al. InterProScan 5: genome-scale protein function classification. **Bioinformatics,** v. 30, n. 9, p. 1236-1240, 2014.

JUÁREZ, M. L. et al. Evaluating mating compatibility within fruit fly cryptic species complexes and the potential role of sex pheromones in pre-mating isolation. **ZooKeys,** v. 540, p. 125-155, 2015.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. **Molecular Biology and Evolution,** v. 30, n. 4, p. 772-780, 2013.

KOBOLDT, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. **Genome Res,** v. 22, n. 3, p. 568-76, 2012.

KOONIN, E. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. **Genome Biology,** v. 5, n. 2, p. R7, 2004.

KORNELIUSSEN, T. S.; ALBRECHTSEN, A.; NIELSEN, R. ANGSD: Analysis of next generation sequencing data. **BMC Bioinformatics,** v. 15, n. 1, p. 356, 2014.

LAMICHHANEY, S. et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. **Nature,** v. 518, n. 7539, p. 371-375, 2015.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods,** v. 9, n. 4, p. 357-359, 2012.

LI, H. et al. The sequence alignment/map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, 2009.

LIU, L. et al. Estimating phylogenetic trees from genome-scale data. **Annals of the New York Academy of Sciences,** v. 1360, n. 1, p. 36-53, 2015.

LUDEÑA, B.; BAYAS, R.; PINTAUD, J.-C. Phylogenetic relationships of Andean-Ecuadorian populations of *Anastrepha fraterculus* (Wiedemann 1830) (Diptera: Tephritidae) inferred from COI and COII gene sequences. **Annales de la Société Entomologique de France,** v. 46, n. 3-4, p. 344-350, 2010.

M'GONIGLE, L. K. et al. Sexual selection enables long-term coexistence despite ecological equivalence. **Nature,** v. 484, n. 7395, p. 506-509, 2012.

MALLET, J. Hybrid speciation. **Nature,** v. 446, n. 7133, p. 279-283, 2007.

______. Why was Darwin's view of species rejected by twentieth century biologists? **Biology & Philosophy,** v. 25, n. 4, p. 497-527, 2010.

MANNI, M. et al. Relevant genetic differentiation among Brazilian populations of *Anastrepha fraterculus* (Diptera, Tephritidae). **ZooKeys,** v. 540, 2015.

MCPHERON, B. A. et al. Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotrypanini) based upon 16S rRNA mitochondrial DNA. In: ALUJA, M. e NORRBOM, A. (Ed.). **Fruit flies (Tephritidae): Phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 13, p.343-362.

MEIER, J. I. et al. Ancient hybridization fuels rapid cichlid fish adaptive radiations. **Nature Communications,** v. 8, p. 14363, 2017.

MENGUAL, X. et al. Phylogenetic relationships of the tribe Toxotrypanini (Diptera: Tephritidae) based on molecular characters. **Molecular Phylogenetics and Evolution,** v. 113, p. 84-112, 2017.

MIRARAB, S.; WARNOW, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. **Bioinformatics,** v. 31, n. 12, p. i44-i52, 2015.

NORRBOM, A. L. et al. Systematic database of names. In: THOMPSON, F. C. (Ed.). **Fruit fly expert identification system and systematic information database: a resource for identification and information on fruit flies and maggots, with information on their classification, distribution and documentation**. Leiden, Netherlands: Backhuys Publisher for the North American Dipterists' Society, v.Volume 9, 1999. p.65-251.

NORRBOM, A. L.; KORYTKOWSKI, C. A. **A revision of the *Anastrepha robusta* species group (Diptera: Tephritidae)**. Auckland, N.Z.: Magnolia Press, 2009.

______. New species of and taxonomic notes on *Anastrepha* (Diptera: Tephritidae). **Zootaxa,** v. 2740, p. 1-23, 2011.

______. New species of *Anastrepha* (Diptera: Tephritidae), with a key for the species of the megacantha clade. **Zootaxa,** v. 3478, p. 510-552, 2012.

NORRBOM, A. L. et al. ***Anastrepha* and *Toxotrypana*: descriptions, illustrations, and interactive keys**. Version: 28th September 2013 2012 onwards.

NORRBOM, A. L. et al. New species and host plants of *Anastrepha* (Diptera: Tephritidae) primarily from Peru and Bolivia. **Zootaxa,** v. 4041, p. 1-94, 2015.

NORRBOM, A. L.; ZUCCHI, R. A.; HERNÁNDEZ-ORTIZ, V. Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotrypanini) based on morphology. In: ALUJA, M. e NORRBOM, A. L. (Ed.). **Fruit flies (Tephritidae): phylogeny and evolution of behavior**. Boca Ratón, Florida: CRC Press, 1999. cap. Chapter 12, p.299-342.

OTTENBURGHS, J. et al. A history of hybrids? Genomic patterns of introgression in the True Geese. **BMC Evolutionary Biology,** v. 17, n. 1, p. 201, 2017.

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 Biology,** v. 17, n. 1, p. 192, 2016.

PARADIS, E.; CLAUDE, J.; STRIMMER, K. APE: Analyses of phylogenetics and evolution in R language. **Bioinformatics,** v. 20, n. 2, p. 289-290, 2004.

PATTERSON, N. J. et al. Ancient admixture in human history. **Genetics**, 2012.

PERRE, P. et al. Morphometric differentiation of fruit fly pest species of the *Anastrepha fraterculus* group (Diptera: Tephritidae). **Annals of the Entomological Society of America,** v. 107, n. 2, p. 490-495, 2014.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing 2015.

REZENDE, V. B. et al. Head transcriptomes of two closely related species of fruit flies of the *Anastrepha fraterculus* group reveals divergent genes in species with extensive gene flow. **G3: Genes|Genomes|Genetics,** v. 6, n. 10, p. 3283-3295, 2016.

ROBERTS, A.; PACHTER, L. Streaming fragment assignment for real-time analysis of sequencing experiments. **Nature Methods,** v. 10, n. 1, p. 71-73, 2013.

RORIZ, A. K. P.; JAPYASSÚ, H. F.; JOACHIM-BRAVO, I. S. Incipient speciation in the *Anastrepha fraterculus* cryptic species complex: reproductive compatibility between *A.* sp.1 aff. *fraterculus* and *A.* sp.3 aff. *fraterculus*. **Entomologia Experimentalis et Applicata,** v. 162, n. 3, p. 346-357, 2017.

RULL, J. et al. Random mating and reproductive compatibility among Argentinean and southern Brazilian populations of *Anastrepha fraterculus* (Diptera: Tephritidae). **Bulletin of Entomological Research,** v. 102, n. 4, p. 435-443, 2012.

RULL, J. et al. Evolution of pre-zygotic and post-zygotic barriers to gene flow among three cryptic species within the *Anastrepha fraterculus* complex. **Entomologia Experimentalis et Applicata,** v. 148, n. 3, p. 213-222, 2013.

RULL, J. et al. Experimental hybridization and reproductive isolation between two sympatric species of tephritid fruit flies in the *Anastrepha fraterculus* species group. **Insect Science**, p. n/a-n/a, 2017.

SALVEMINI, M. et al. *De Novo a*ssembly and transcriptome analysis of the Mediterranean Fruit Fly *Ceratitis capitata e*arly embryos. **PLoS ONE,** v. 9, n. 12, p. e114191, 2014.

SCALLY, M. et al. Resolution of inter and intra-species relationships of the West Indian fruit fly *Anastrepha obliqua*. **Molecular Phylogenetics and Evolution,** v. 101, p. 286-293, 2016.

SCHUTZE, M. K. et al. Tephritid integrative taxonomy: Where we are now, with a focus on the resolution of three tropical fruit fly species complexes. **Annual Review of Entomology,** v. 62, n. 1, p. 147-164, 2017.

SEEHAUSEN, O. Hybridization and adaptive radiation. **Trends in Ecology & Evolution,** v. 19, n. 4, p. 198-207, 2004.

SEGURA, D. F. et al. Assortative mating among *Anastrepha fraterculus* (Diptera: Tephritidae) hybrids as a possible route to radiation of the *fraterculus* cryptic species complex. **Biological Journal of the Linnean Society,** v. 102, n. 2, p. 346-354, 2011.

SELIVON, D.; PERONDINI, A. L. P.; MORGANTE, J. S. A genetic–morphological characterization of two cryptic species of the *Anastrepha fraterculus* complex (Diptera: Tephritidae). **Annals of the Entomological Society of America,** v. 98, n. 3, p. 367-381, 2005.

SELIVON, D. et al. New variant forms in the *Anastrepha fraterculus* complex (Diptera, Tephritidae). In: B.N., B., Proceedings of the 6th International Symposium on Fruit Flies of Economic Importance., 2004, Irene, South Africa. Isteg Scientific Pub, 6-10 May 2002. p.253-258.

SERVEDIO, M. R. Geography, assortative mating, and the effects of sexual selection on speciation with gene flow. **Evolutionary Applications,** v. 9, n. 1, p. 91-102, 2016.

SILVA, J. G.; BARR, N. B. Recent advances in molecular systematics of *Anastrepha* Schiffner. Proceedings of the 7th International Symposium on Fruit Flies of Economic Importance, 2008, Salvador, Brazil. 10-15 September 2006. p.13-28.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics,** v. 31, n. 19, p. 3210-3212, 2015.

SMITH-CALDAS, M. R. B. et al. Phylogenetic relationships among species of the *fraterculus* group (*Anastrepha*: Diptera: Tephritidae) inferred from DNA sequences of mitochondrial cytochrome oxidase I. **Neotropical Entomology,** v. 30, n. 4, p. 565-573, 2001.

SMITH, S. A.; DUNN, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. **Bioinformatics,** v. 24, n. 5, p. 715-716, 2008.

SOBRINHO, I. S.; DE BRITO, R. A. Evidence for positive selection in the gene *fruitless* in *Anastrepha* fruit flies. **BMC Evolutionary Biology,** v. 10, n. 1, p. 293, 2010.

______. Positive and purifying selection influence the evolution of *doublesex* in the *Anastrepha fraterculus* species group. **PLoS ONE,** v. 7, n. 3, p. e33446-e33446, 2012.

SOLÍS-LEMUS, C.; ANÉ, C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. **PLoS Genetics,** v. 12, n. 3, p. e1005896, 2016.

SOLÍS-LEMUS, C.; YANG, M.; ANÉ, C. Inconsistency of species tree methods under gene flow. **Systematic Biology,** v. 65, n. 5, p. 843-851, 2016.

SORAGGI, S.; WIUF, C.; ALBRECHTSEN, A. Improved D-statistic for low-coverage data. **bioRxiv**, 2017.

STAMATAKIS, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics,** v. 30, n. 9, p. 1312-1313, 2014.

STONE, A. **The fruit flies of the genus *Anastrepha***. Washington, DC, US: USDA. Misc. Publ., 1942. 439p.

SUTTON, B. D. et al. Nuclear ribosomal internal transcribed spacer 1 (ITS1) variation in the *Anastrepha fraterculus* cryptic species complex (Diptera, Tephritidae) of the Andean region. **ZooKeys,** v. 540, 2015.

THAN, C.; RUTHS, D.; NAKHLEH, L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. **BMC Bioinformatics,** v. 9, n. 1, p. 322, 2008.

THE *HELICONIUS* GENOME CONSORTIUM. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. **Nature,** v. 487, n. 7405, p. 94-98, 2012.

VACHASPATI, P.; WARNOW, T. ASTRID: Accurate Species TRees from Internode Distances. **BMC Genomics,** v. 16, n. 10, p. S3, 2015.

VAN DONGEN, S. **Graph clustering by flow simulation**. 2000. (Ph.D thesis). University of Utrecht, Utrecht, Netherlands.

VANÍČKOVÁ, L. et al. Current knowledge of the species complex *Anastrepha fraterculus* (Diptera, Tephritidae) in Brazil. **ZooKeys,** v. 540, 2015.

VERA, M. T. et al. Mating incompatibility among populations of the South American fruit fly *Anastrepha fraterculus* (Diptera: Tephritidae). **Annals of the Entomological Society of America,** v. 99, n. 2, p. 387-397, 2006.

WANG, J.; XIONG, K.-C.; LIU, Y.-H. *De novo* transcriptome analysis of Chinese citrus fly, *Bactrocera minax* (Diptera: Tephritidae), by high-throughput illumina sequencing. **PLOS ONE,** v. 11, n. 6, p. e0157656, 2016.

YANG, Y.; SMITH, S. A. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. **Molecular Biology and Evolution,** v. 31, n. 11, p. 3081-3092, 2014.

YU, Y.; DEGNAN, J. H.; NAKHLEH, L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. **PLOS Genetics,** v. 8, n. 4, p. e1002660, 2012.

YU, Y.; NAKHLEH, L. A maximum pseudo-likelihood approach for phylogenetic networks. **BMC Genomics,** v. 16, n. 10, p. S10, 2015.

ZDOBNOV, E. M. et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. **Nucleic Acids Research,** v. 45, n. D1, p. D744-D749, 2017.

ZHANG, W. et al. Genome-wide introgression among distantly related *Heliconius* butterfly species. **Genome Biology,** v. 17, n. 1, p. 25, 2016.

ZHANG, W.; KUNTE, K.; KRONFORST, M. R. Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using *de novo* transcriptome assemblies. **Genome Biology and Evolution,** v. 5, n. 6, p. 1233-1245, 2013.

ZHENG, W. et al. RNA sequencing to characterize transcriptional changes of sexual maturation and mating in the female oriental fruit fly *Bactrocera dorsalis*. **BMC Genomics,** v. 17, n. 1, p. 194, 2016.

ZUCCHI, R. A. Espécies de *Anastrepha*, sinonímias, plantas hospedeiras e parasitóides. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000a. p.41-48.

______. Taxonomia. In: MALAVASI, A. e ZUCCHI, R. A. (Ed.). **Moscas-das-frutas de importância econômica no Brasil: Conhecimento básico e aplicado**. Ribeirão Preto, Brazil: Holos, 2000b. p.1-24.

Appendix 1

#Commands used in this study.

#Cleanning the reads

```
java -jar trimmomatic-0.35.jar PE -threads 20 -phred33 rawreads_R1.fastq.gz rawreads_R2_001.fastq.gz filteredreads_PE1.fq.gz filteredreads_SE1.fq.gz filteredreads_PE2.fq.gz filteredreads_SE2.fq.gz HEADCROP:1 ILLUMINACLIP:TruSeq-adapters.fa:2:30:10 LEADING:15 TRAILING:15 SLIDINGWINDOW:5:20 MINLEN:50 2> output.std
```

#Assembly with Stranded library

```
Trinity --seqType fq --max_memory 500G --left filteredreads_PE1.fq.gz --right filteredreads_PE2.fq.gz --CPU 80 --min_contig_length 200 --SS_lib_type RF --output outputdir_trinity
```

#Assembly without Stranded library

```
Trinity --seqType fq --max_memory 500G --left filteredreads_PE1.fq.gz --right filteredreads_PE2.fq.gz --CPU 80 --min_contig_length 200 --output outputdir_trinity
```

#Assessing of quality assembly

```
BUSCO.py -i Trinity.fasta -o output -l arthropoda_odb9 -m tran -c 50 --long -sp fly -e 1e-06 -z
```

#Prediction of CDSs with Stranded library

```
TransDecoder.LongOrfs -t Trinity.fasta -S
```

#Prediction of CDSs without Stranded library

```
TransDecoder.LongOrfs -t Trinity.fasta
```

#Annotation using pfam

```
hmmscan --cpu 10 --domtblout pfam.domtblout Pfam-A.hmm Trinity.fasta.transdecoder_dir/longest_orfs.pep
```

#Annotation using Unitprot

```
blastp -query Trinity.fasta.transdecoder_dir/longest_orfs.pep -db uniref90.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 90 > blastp.outfmt6
```

#Final results of TransDecoder

```
TransDecoder.Predict -t Trinity.fasta --retain_pfam_hits pfam.domtblout --retain_blastp_hits blastp.outfmt6
```

#Orthology prediction

```
makeblastdb -in all.fa -parse_seqids -dbtype nucl -out all.fa
```

```
blastn -db all.fa -query all.fa -evalue 10 -num_threads 80 -max_target_seqs 1000 -out all.rawblast -outfmt '6 qseqid qlen sseqid slen frames pident nident length mismatch gapopen qstart qend sstart send evalue bitscore'
```

```
python blast_to_mcl.py all.rawblast 0.4
```

```
mcl all.rawblast.hit-frac0.4.minusLogEvalue --abc -te 20 -tf 'gq(5)' -I 1.4 -o hit-frac0.4_I1.4_e5
```

```
python write_fasta_files_from_mcl.py all.fa hit-frac0.4_I1.4_e5 10 mlc_clusters

python fasta_to_tree.py mlc_clusters 20 dna n

python trim_tips.py mlc_clusters .tre 0.2 0.4

python mask_tips_by_taxonID_transcripts.py mlc_clusters mlc_clusters y

python cut_long_internal_branches.py mlc_clusters .mm 0.3 10 homolog_final

python prune_paralogs_MI.py homolog_final .subtree 0.2 0.4 10 MI/tree

#POTION pipeline

perl ~/bin/POTION-1.1.2/bin/potion.pl --conf_file_path config_file (see appendix 2)

#Gene tree inference

raxml -T 30 -f a -x 12345 -# 200 -p 12345 -s input -n output_prefix -m GTRCAT

#Species tree inference

#ASTRAL-II

java -jar astral.4.10.12.jar -i input_trees.tre -o output_species_tree.tre

java -jar astral.4.10.12.jar -q output_species_tree.tre -t 1 -i input_trees.tre -o output_species_tree-t1.tre 2>
output_species_tree-t1.out

#ASTRID

ASTRID-linux -i input_trees.tre -r 1000 -o output_species_tree.tre -c output_trees.cache

#SVDquartets

bowtie2 -p 30 --nofw --very-sensitive -x abisSCSP01_reference_assembly.fas -1 reads_PE1.fq.gz -2
reads_PE2.fq.gz | samtools view  -Sb  - | samtools sort -@ 20 1> file.bam 2> output.err

java -jar -XX:ParallelGCThreads=10 picard.jar MarkDuplicates INPUT=file.bam OUTPUT=output_dedup.bam
M=output_dedup.info TMP_DIR=./tmp ASSUME_SORTED=true REMOVE_DUPLICATES=true
MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000

samtools mpileup -q 20 -d 10000000 -f abisSCSP01_reference_assembly.fas file1.bam file2.bam | java -jar
VarScan.v2.4.2.jar mpileup2snp --min-coverage 6 --min-reads2 1 --min-avg-qual 30 --min-var-freq 0.01 --min-
freq-for-hom 0.75 --p-value 0.05 --strand-filter 1 --output-vcf 1  1> SNPs.vcf

#Paup script

begin paup;

Execute SNPs.nex;

log file=svd_quartets_lineages_bootstrap.log start;

SVDQuartets nquartets=100000 speciesTree=no showSV=yes showScores=yes
qfile=svd_quartets_lineages_bootstrap.qfile  seed=475839 bootstrap nreps=1000 nthreads=10
mrpFile=svd_quartets_lineages_bootstrap.mrp;
```

```
log stop;

savetrees treeWts=yes brLens=yes taxaBlk=yes setStoreCmd=yes supportValues=both
file=svd_quartets_lineages_bootstrap.tre;

saveassum file=svd_quartets_lineages_bootstrap.assum;

end;

quit;
```

Appendix 2

```
# Config file including all parameters used in POTION for dataset including all samples.

###############PROJECT PARAMETERS#######################

mode = site                   # main analysis mode. Currently POTION supports only site-models analysis.

CDS_dir_path = folder_with_CDS_files                   #path to folder containing CDS data

homology_file_path = input_otholog_file             # path to the ORTHOMCL 1.4 main output file

project_dir_path = folder_output    # path to the main directory where results will be created.Parent directory must exist.

max_processors = 32                # number of processors to be used in parallelized steps of POTION

remove_identical = no             # "yes" to remove 100% identical nucleotide groups at the very beginning of

                                  # analysis, "no" otherwise

verbose = 1                   # 1 to print nice log messages telling you what is going on. 0 otherwise

############SEQUENCE/GROUP PARAMETERS################

groups_to_process = all           # Defines which lines of the cluster file (ortholog groups) will be processed.

                         # Use "all" to process every group, "-" to set groups between two given lines

                         # (including the said lines).

                         # Use "!" to not process a specific line, can be used with "-" to specify a

                         # set to not be processed. Useful if groups are taking too long to finish.

                                # Use "," or ";" to set distinct sets

                                # Examples: 1;4-10;12  will process groups 1, 4 to 10 and group 12

                                #           all;!3     will process all groups, except group 3

                                #           all;!3-5   will process all groups, except groups 3 to 5

behavior_about_bad_clusters = 1        # what should POTION do if it finds a cluster with a sequence removed

                         # due to any filter? Possible options are:

                         # 0 - does not filter any sequence (not recommended)

                         # 1 - removal of any flagged sequence

                         # 2 - removal of any group with flagged sequences

homology_filter = 1         # this variable controls for what POTION will do if a group with paralogous

                         # genes is found. Possible options are:

                         # 0 - analyze all sequences within group

                         # 1 - remove all paralogous within group, analyzing only single-copy genes

                         # 2 - remove groups with paralogous genes
                         # 3 - remove single-copy genes, analyzing all paralogous within group together
                         # 4 - remove single-copy genes and split remaining paralogous into individual
                         # species, evaluating each subgroup individually
validation_criteria = 3          # quality criteria to remove sequences. Possible values are:
                         # 1 - checks for valid start codons
                         # 2 - checks for valid stop codons
                         # 3 - checks for sequence size multiple of 3
                         # 4 - checks for nucleotides outside ATCG
                         # 'all' applies every verification
additional_start_codons = ()       # these codons, plus the ones specified in codon table, will be the valid start
                                    # codons for validation purposes
additional_stop_codons = ()        # same as start codons
codon_table = 1                  # codon table id (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi)
absolute_min_sequence_size = 150     # minimum sequence length cutoff for sequence/group further evaluation
absolute_max_sequence_size = 40000    # maximum sequence length cutoff for sequence/group further evaluation
relative_min_sequence_size = 0.70     # sequences smaller than mean|meadian times this value will be filtered
relative_max_sequence_size = 1.3     # sequences greater than mean|meadian times this value will be filtered
sequence_size_average_metric = median   # which average metric will be calculated to determine the
                         # minimum/maximum relative lengths ranges for sequence removal
                         # Possible values are "mean" and "median"
min_group_identity = 60          # mean minimum group identity cutoff in pairwise sequence alignments
max_group_identity = 99.9          # mean maximum group identity cutoff in pairwise sequence alignments
group_identity_comparison = aa      # the kind of sequence that will be used when computing mean group identity
                                     # possible values are "nt" or "aa"
min_sequence_identity = 60        # minimum (mean/median) sequence identity cutoff in pairwise sequence alignments
max_sequence_identity = 99.99        # maximum (mean/median) sequence identity cutoff in pairwise sequence alignemnts
sequence_identity_average_metric = median # would you like to use mean or median to measure sequence identity?
                        # possible values are "mean" and "median"


sequence_identity_comparison = aa    # the kind of sequence that will be used when computing sequence identity
                        # possible values are "nt" and "aa"
```

```
min_gene_number_per_cluster = 10        # minimum # genes in group after all filtering steps

max_gene_number_per_cluster = 25        # maximum # genes in group after all filtering steps

min_specie_number_per_cluster = 10      # minimum # species in group after all filtering steps

max_specie_number_per_cluster = 25      # maximum # species in group after all filtering steps

reference_genome_file =                 # genome reference name, leave blank for none (same name used in fasta file)


###########THIRD-PARTY SOFTWARE CONFIGURATION###############


multiple_alignment = mafft        # program used for multiple sequence alignment. Possible values are

                                       # muscle, mafft and prank

bootstrap = 100                   # number of bootstraps in phylogenetic analysis

phylogenetic_tree_speed = fast         # fast or slow analysis? Used in phylip dnaml or proml only

phylogenetic_tree = phyml_nt           # program used for phylogenetic tree reconstruction. Possible values are

                                      # proml dnaml, phyml_aa and phyml_nt

recombination_qvalue = 0.05             # q-value for recombination detection. Must occur for all the specified tests

rec_minimum_confirmations = 2            # minimum number of significant recombination tests positives

rec_mandatory_tests = phi NSS maxchi2    # any combination of the three test names, separated by spaces, or N.A. to use

                          # any test

remove_gaps = strictplus                 # numeric values between 0 and 1 will remove columns with that percentage of

                          # gaps. Values of "strict" or "strictplus" will use respectively these

                          # filters to remove unreliable regions (described in trimal article)

PAML_models = m12 m78   # codeml models to be generated. "m12" and/or "m78" values acceptable.

pvalue = 0.05                     # p-values for positive selection detection

qvalue = 0.05                     # q-values for positive selection detection
```

Appendix 3

# Config file including all parameters used in POTION for dataset including *fraterculus* group and *A. bistrigata* samples.

```
##############PROJECT PARAMETERS######################

mode = site                        # main analysis mode. Currently POTION supports only site-models analysis.

CDS_dir_path = folder_with_CDS_files             #path to folder containing CDS data

homology_file_path = input_otholog_file          # path to the ORTHOMCL 1.4 main output file

project_dir_path = folder_output     # path to the main directory where results will be created.Parent directory must exist.

max_processors = 32                  # number of processors to be used in parallelized steps of POTION

remove_identical = no                # "yes" to remove 100% identical nucleotide groups at the very beginning of

                                      # analysis, "no" otherwise

verbose = 1                        # 1 to print nice log messages telling you what is going on. 0 otherwise


###########SEQUENCE/GROUP PARAMETERS###############

groups_to_process = all               # Defines which lines of the cluster file (ortholog groups) will be processed.

                              # Use "all" to process every group, "-" to set groups between two given lines

                              # (including the said lines).

                              # Use "!" to not process a specific line, can be used with "-" to specify a

                              # set to not be processed. Useful if groups are taking too long to finish.

                                   # Use "," or ";" to set distinct sets

                                   # Examples: 1;4-10;12  will process groups 1, 4 to 10 and group 12

                                   #       all;!3    will process all groups, except group 3

                                   #       all;!3-5   will process all groups, except groups 3 to 5

behavior_about_bad_clusters = 1        # what should POTION do if it finds a cluster with a sequence removed

                              # due to any filter? Possible options are:

                              # 0 - does not filter any sequence (not recommended)

                              # 1 - removal of any flagged sequence

                              # 2 - removal of any group with flagged sequences

homology_filter = 1            # this variable controls for what POTION will do if a group with paralogous

                              # genes is found. Possible options are:

                              # 0 - analyze all sequences within group
```

```
                                   # 1 - remove all paralogous within group, analyzing only single-copy genes
                                   # 2 - remove groups with paralogous genes
                                   # 3 - remove single-copy genes, analyzing all paralogous within group together
                                   # 4 - remove single-copy genes and split remaining paralogous into individual
                                   # species, evaluating each subgroup individually
validation_criteria = 3             # quality criteria to remove sequences. Possible values are:
                                   # 1 - checks for valid start codons
                                   # 2 - checks for valid stop codons
                                   # 3 - checks for sequence size multiple of 3
                                   # 4 - checks for nucleotides outside ATCG
                                   # 'all' applies every verification
additional_start_codons = ()        # these codons, plus the ones specified in codon table, will be the valid start
                                   # codons for validation purposes
additional_stop_codons = ()         # same as start codons
codon_table = 1                     # codon table id (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi)
absolute_min_sequence_size = 150    # minimum sequence length cutoff for sequence/group further evaluation
absolute_max_sequence_size = 40000  # maximum sequence length cutoff for sequence/group further evaluation
relative_min_sequence_size = 0.70   # sequences smaller than mean|meadian times this value will be filtered
relative_max_sequence_size = 1.3    # sequences greater than mean|meadian times this value will be filtered
sequence_size_average_metric = median   # which average metric will be calculated to determine the
                                   # minimum/maximum relative lengths ranges for sequence removal
                                   # Possible values are "mean" and "median"
min_group_identity = 80             # mean minimum group identity cutoff in pairwise sequence alignments
max_group_identity = 99.9           # mean maximum group identity cutoff in pairwise sequence alignments
group_identity_comparison = aa      # the kind of sequence that will be used when computing mean group identity
                                  # possible values are "nt" or "aa"
min_sequence_identity = 80          # minimum (mean/median) sequence identity cutoff in pairwise sequence alignments
max_sequence_identity = 99.99       # maximum (mean/median) sequence identity cutoff in pairwise sequence alignemnts
sequence_identity_average_metric = median # would you like to use mean or median to measure sequence identity?
                                  # possible values are "mean" and "median"
sequence_identity_comparison = aa   # the kind of sequence that will be used when computing sequence identity
                                  # possible values are "nt" and "aa"
```

```
min_gene_number_per_cluster = 10        # minimum # genes in group after all filtering steps

max_gene_number_per_cluster = 17        # maximum # genes in group after all filtering steps

min_specie_number_per_cluster = 10      # minimum # species in group after all filtering steps

max_specie_number_per_cluster = 17      # maximum # species in group after all filtering steps

reference_genome_file =                 # genome reference name, leave blank for none (same name used in fasta file)

############THIRD-PARTY SOFTWARE CONFIGURATION###############

multiple_alignment = mafft              # program used for multiple sequence alignment. Possible values are

                            # muscle, mafft and prank

bootstrap = 100                         # number of bootstraps in phylogenetic analysis

phylogenetic_tree_speed = fast          # fast or slow analysis? Used in phylip dnaml or proml only

phylogenetic_tree = phyml_nt            # program used for phylogenetic tree reconstruction. Possible values are

                            # proml dnaml, phyml_aa and phyml_nt

recombination_qvalue = 0.05             # q-value for recombination detection. Must occur for all the specified tests

rec_minimum_confirmations = 2           # minimum number of significant recombination tests positives

rec_mandatory_tests = phi NSS maxchi2   # any combination of the three test names, separated by spaces, or N.A. to use

                            # any test

remove_gaps = strictplus                # numeric values between 0 and 1 will remove columns with that percentage of

                            # gaps. Values of "strict" or "strictplus" will use respectively these

                            # filters to remove unreliable regions (described in trimal article)

PAML_models = m12 m78                   # codeml models to be generated. "m12" and/or "m78" values acceptable.

pvalue = 0.05                           # p-values for positive selection detection

qvalue = 0.05                           # q-values for positive selection detection
```

Appendix 4

Table S1. Sampling information of individuals from *Anastrepha* lineages.

Sample	Species	Species group	City (State/Province[b])	Country	Coordinates	Host
A. sororcula SP	*A. sororcula*	*fraterculus*	Ilha Bela (SP)	Brazil	23°47'19.52"S 45°21'42.02"W	Guava
A. obliqua RJ	*A. obliqua*	*fraterculus*	Conceição do Jacareí (RJ)	Brazil	23° 1'54.32"S 44° 9'54.14"W	Starfruit
A. obliqua GO	*A. obliqua*	*fraterculus*	Goiania (GO)	Brazil	16°41′58″S 49°16′35″W	Jacote
A. obliqua PR1	*A. obliqua*	*fraterculus*	Capanema (PR)	Brazil	25°39'45.54"S 53°48'28.74"W	*Eugenia uvalha*
A. obliqua PR2	*A. obliqua*	*fraterculus*	Marialva (PR)	Brazil	23°30'56.68"S 51°49'34.11"W	Jocote
A. distincta SP	*A. distincta*	*fraterculus*	São Carlos (SP)	Brazil	21°57'33"S 47°53'54"W	Inga
A. turpiniae MG	*A. turpiniae*	*fraterculus*	Três Marias (MG)	Brazil	18°12'13.28"S 45°14'23.34"W	Guava
A. turpiniae SP	*A. turpiniae*	*fraterculus*	Araraquara (SP)	Brazil	21°48'55.81"S 48°12'5.34"W	Guava
A. fraterculus SP1	*A. fraterculus*	*fraterculus*	São Carlos (SP)	Brazil	22° 1'49.84"S 47°54'27.90"W	Guava
A. fraterculus ES	*A. fraterculus*	*fraterculus*	Muniz Freire (MG)	Brazil	20°27'52.13"S 41°24'54.88"W	*Plinia cauliflora*
A. fraterculus RJ	*A. fraterculus*	*fraterculus*	Conceição do Jacareí (RJ)	Brazil	23° 1'54.32"S 44° 9'54.14"W	Tropical-almond
A. fraterculus RS	*A. fraterculus*	*fraterculus*	Dois Irmãos (RS)	Brazil	29°57'7"S 51°11'33"W	*Cattley guava*
A. fraterculus TUC	*A. fraterculus*	*fraterculus*	Tucumán (TUC)	Argentina	- -	-
A. fraterculus SC	*A. fraterculus*	*fraterculus*	Itapema (SC)	Brazil	27°05'36"S 48°37'08"W	Barbados cherry
A. fraterculus SP2	*A. fraterculus*	*fraterculus*	Porto Ferreira (SP)	Brazil	21°50'59"S 47°29'42"W	Rangpur
A. fraterculus BA	*A. fraterculus*	*fraterculus*	Ubaitaba (BA)	Brazil	14°18'37.66"S 39°19'18.08"W	Guava
A. bistrigata	*A. bistrigata*	*striata*	São Carlos (SP)	Brazil	22° 1'49.84"S 47°54'27.90"W	Guava
A. pseudoparallela	*A. pseudoparallela*	*pseudoparallela*	Porto Ferreira (SP)	Brazil	21°50'59"S 47°29'42"W	Passion fruit
A. grandis	*A. grandis*	*grandis*	Porto Ferreira (SP)	Brazil	21°50'59"S 47°29'42"W	Pumpkin
A. serpentina	*A. serpentina*	*serpentina*	Araraquara (SP)	Brazil	21°48'55.81"S 48°12'5.34"W	*Pouteria campechiana*

[a] Samples consisted of a pool of individuals.

[b] SP: São Paulo; RJ: Rio de Janeiro; GO: Goiás; PR: Paraná; MG: Minas Gerais; RS: Rio Grande do Sul; BA: Bahia; TUC: Tucumán.

Table S2. Sequenced and filtered pair-end reads of female reproductive transcriptomes from *Anastrepha* specimens.

Sample	Sequenced PE reads*	Read length (bp)	Filtered PE reads*	Removed reads Only forward	Only Reverse	Both PE*
A. sororcula SP	21,453,462	100	20,495,047	596,373	175,997	186,045
A. obliqua RJ	17,688,068	125	16,519,954	749,341	221,955	196,818
A. obliqua GO	20,247,452	100	18,846,636	881,217	289,655	229,944
A. obliqua PR1	19,068,596	125	17,772,333	794,035	275,119	227,109
A. obliqua PR2	20,627,384	125	19,241,590	874,179	287,259	224,356
A. distincta SP	20,987,635	100	19,859,708	725,559	176,999	225,369
A. turpiniae MG	25,540,362	100	24,403,199	717,680	187,141	232,342
A. turpiniae SP	22,995,090	100	21,844,276	743,105	193,184	214,525
A. fraterculus SP1	28,241,337	100	25,649,251	1,127,424	965,178	499,484
A. fraterculus ES	21,640,288	100	20,645,615	629,686	153,174	211,813
A. fraterculus RJ	18,488,145	125	17,253,617	829,672	202,836	202,020
A. fraterculus RS	22,252,650	100	21,155,504	680,320	182,726	234,100
A. fraterculus TUC	25,881,894	100	24,990,366	488,553	295,549	107,426
A. fraterculus SC	22,679,589	100	21,607,546	694,369	166,134	211,540
A. fraterculus SP2	20,842,744	100	19,903,058	591,330	155,453	192,903
A. fraterculus BA	22,131,648	100	21,052,866	671,460	193,394	213,928
A. bistrigata	19,595,706	125	18,195,636	922,285	246,368	231,417
A. pseudoparallela	19,861,690	125	18,586,823	791,238	275,688	207,941
A. grandis	21,185,565	100	20,384,121	463,096	168,300	170,048
A. serpentina	20,420,542	125	19,111,638	799,692	294,824	214,388

*PE: Pair-end

Table S3. Assembly statistics of female reproductive transcriptomes from *Anastrepha* specimens.

Sample	Assembled bases	Contigs	Unigenes[1]	Contigs > 1,000bp[2]	Contigs > 10,000bp[3]	Longest contig (bp)	Median	Average	N50
A. sororcula SP	56,459,957	57,724	11,067	15,592	93	27,787	390	978.10	2,161
A. obliqua RJ	62,051,237	56,399	11,140	17,820	88	22,386	477	1,100.22	2,285
A. obliqua GO	66,121,389	67,290	12,134	18,860	46	21,941	436	982.63	2,016
A. obliqua PR1	69,128,257	70,205	11,732	19,361	75	17,083	422	984.66	2,099
A. obliqua PR2	74,594,148	82,003	13,140	20,632	76	25,155	403	909.65	1,876
A. distincta SP	55,929,239	58,911	12,022	15,836	84	26,158	403	949.39	2,015
A. turpiniae MG	67,559,555	74,205	12,242	18,494	96	27,958	388	910.44	1,936
A. turpiniae SP	67,847,361	67,085	13,698	18,408	154	29,262	411	1,011.36	2,225
A. fraterculus SP1	62,530,301	68,777	11,895	17,217	120	27,168	405	909.17	1,856
A. fraterculus ES	54,372,681	60,798	12,443	15,332	64	22,073	400	894.32	1,838
A. fraterculus RJ	69,374,899	75,587	12,293	19,365	56	19,106	412	917.82	1,867
A. fraterculus RS	57,826,073	64,115	12,597	16,101	106	28,919	392	901.91	1,885
A. fraterculus TUC	61,598,433	64,816	11,901	17,353	50	23,477	430	950.36	1,938
A. fraterculus SC	64,001,819	67,350	12,184	17,475	122	27,848	396	950.29	2,066
A. fraterculus SP2	64,643,582	68,137	12,552	17,878	109	27,851	403	948.73	2,016
A. fraterculus BA	63,241,603	70,181	14,341	17,091	107	27,469	393	901.12	1,885
A. bistrigata	76,754,085	85,312	13,646	21,583	71	19,490	411	899.69	1,795
A. pseudoparallela	63,085,992	73,908	13,890	18,470	46	15,403	419	853.57	1,620
A. grandis	67,898,836	62,134	11,367	18,193	229	27,563	408	1,092.78	2,524
A. serpentina	65,686,623	82,656	14,634	18,648	45	16,300	397	794.70	1,466

[1] Number of filtered unigenes: Trinity components with non-redundant CDSs and filtered by expression.
[2] Number of contigs longer than 1,000bp.
[3] Number of contigs longer than 10,000bp.

Table S4. Quality assessment summary of raw assemblies and filtered unigenes of female reproductive transcriptomes from *Anastrepha* specimens. Each dataset was compared against 1,066 single-copy orthologs of Arthropoda from OrthoDB.

Sample	Data type	Complete BUSCOs[a]	Complete single-copy BUSCOs[b]	Complete duplicated BUSCOs[c]	Fragmented BUSCOs[d]	Missing BUSCOs[e]
A. sororcula SP	Assembly	1035	743	292	7	24
	Unigenes	1025	1017	8	5	36
A.obliqua RJ	Assembly	1029	651	378	14	23
	Unigenes	1020	1010	10	11	35
A.obliqua GO	Assembly	1031	633	398	9	26
	Unigenes	1016	1007	9	9	41
A.obliqua PR1	Assembly	1026	678	348	18	22
	Unigenes	1011	1001	10	16	39
A.obliqua PR2	Assembly	1034	627	407	9	23
	Unigenes	1016	1007	9	8	42
A.distincta SP	Assembly	1033	737	296	12	21
	Unigenes	1024	1017	7	10	32
A.turpiniae MG	Assembly	1036	712	324	13	17
	Unigenes	1026	1014	12	10	30
A.turpiniae SP	Assembly	1039	719	320	5	22
	Unigenes	1023	1013	10	3	40
*A.fraterculus S*P1	Assembly	1032	656	376	10	24
	Unigenes	1019	1013	6	10	37
A.fraterculus ES	Assembly	1034	771	263	13	19
	Unigenes	1022	1013	9	12	32
A.fraterculus RJ	Assembly	1032	652	380	12	22
	Unigenes	1024	1018	6	6	36
A.fraterculus RS	Assembly	1032	766	266	16	18
	Unigenes	1022	1019	3	14	30
A.fraterculus TUC	Assembly	1040	723	317	14	12
	Unigenes	1024	1016	8	11	31
A.fraterculus SC	Assembly	1037	757	280	12	17
	Unigenes	1025	1018	7	11	30
*A.fraterculus S*P2	Assembly	1040	744	296	7	19
	Unigenes	1032	1027	5	6	28
A.fraterculus BA	Assembly	1037	769	268	13	16
	Unigenes	1027	1018	9	8	31
A.bistrigata	Assembly	1029	652	377	16	21
	Unigenes	1008	1003	5	17	41
A.pseudoparallela	Assembly	1023	665	358	24	19
	Unigenes	1001	990	11	24	41
A.grandis	Assembly	1037	704	333	8	21
	Unigenes	1025	1011	14	6	35
A.serpentina	Assembly	1023	610	413	21	22
	Unigenes	1009	1001	8	20	37

[a] Complete BUSCOs: Number of Benchmarking Universal Single-Copy Orthologs (BUSCOs) found completed in a transcriptome.

[b] Complete single-copy BUSCOs: Number of BUSCOs found completed and in only one copy (without redundancy) in a transcriptome.
[c] Complete duplicated BUSCOs: Number of BUSCOs found completed and in more than one copy (with redundancy) in a transcriptome.
[d] Fragmented BUSCOs: Number of BUSCOs found incomplete in a transcriptome.
[e] Missing BUSCOs: Number of BUSCOs did not find in a transcriptome.

Table S5. Annotation of unigenes from female reproductive transcriptomes of *Anastrepha* lineages.

Sample	**CDSs**	***D. melanogaster* annotation**	**nr annotation***	**Interproscan**	**Blast2GO**
A. sororcula SP	11286	9550	10891	10510	8280
A.obliqua RJ	11381	9446	10990	10614	8270
A.obliqua GO	12415	9999	11768	11461	8887
A.obliqua PR1	12018	9653	11489	11016	8591
A.obliqua PR2	13463	10421	12716	12302	9491
A.distincta SP	12239	10128	11810	11257	8731
A.turpiniae MG	12513	10204	11987	11448	8891
A.turpiniae SP	12170	9936	11665	11180	8712
*A.fraterculus S*P1	12136	9911	11677	11211	8692
A.fraterculus ES	12672	10418	12249	11637	9004
A.fraterculus RJ	12561	10222	12030	11565	8939
A.fraterculus RS	12845	10521	12385	11802	9116
A.fraterculus TUC	14728	11105	13636	12863	9852
A.fraterculus SC	12449	10153	11964	11421	8879
*A.fraterculus S*P2	12831	10398	12367	11802	9062
A.fraterculus BA	14630	11216	13423	13211	10033
A.bistrigata	13927	11083	13375	12653	9722
A.pseudoparallela	14144	11511	13637	12870	9844
A.grandis	11597	9886	11156	10707	8475
A.serpentina	14929	11621	14172	13581	10257

*nr: Non-redundant protein database of the GenBank (nr) including only proteins of arthropods.

Table S6. ABBA-BABA tests among *Anastrepha* lineages H1 and H2 as *fraterculus* group lineages.

			A. serpentina[a]		
H1	**H2**	**H3**	***D*-statistic ±SE**	**z**	**q-value**
A. obliqua	*A. sororcula*	*A. grandis*	0.022053 ±0.00005	3.13	0.016567
A. obliqua	*A. sororcula*	*A. pseudoparallela*	-0.002286 ±0.000053	-0.31	1.000000
A. obliqua	*A. sororcula*	*A. bistrigata*	0.016399 ±0.000049	2.34	0.101813
A. obliqua	*A. fraterculus* C2	*A. grandis*	-0.003877 ±0.000036	-0.65	1.000000
A. obliqua	*A. fraterculus* C2	*A. pseudoparallela*	-0.025575 ±0.00004	-4.03	0.001890
A. obliqua	*A. fraterculus* C2	*A. bistrigata*	-0.009151 ±0.000036	-1.52	0.403222
A. obliqua	*A. distincta*	*A. grandis*	0.013893 ±0.000046	2.04	0.182625
A. obliqua	*A. distincta*	*A. pseudoparallela*	0.00247 ±0.000049	0.35	1.000000
A. obliqua	*A. distincta*	*A. bistrigata*	0.006018 ±0.000047	0.87	0.905868
A. obliqua	*A. turpiniae*	*A. grandis*	0.006507 ±0.00004	1.03	0.769580
A. obliqua	*A. turpiniae*	*A. pseudoparallela*	-0.009571 ±0.000043	-1.46	0.431427
A. obliqua	*A. turpiniae*	*A. bistrigata*	0.00024 ±0.000042	0.04	1.000000
A. obliqua	*A. fraterculus* C1	*A. grandis*	-0.001293 ±0.000034	-0.22	1.000000
A. obliqua	*A. fraterculus* C1	*A. pseudoparallela*	-0.018342 ±0.000037	-3.03	0.018793
A. obliqua	*A. fraterculus* C1	*A. bistrigata*	-0.004598 ±0.000035	-0.77	0.997507
A. sororcula	*A. fraterculus* C2	*A. grandis*	-0.021741 ±0.000042	-3.37	0.008916
A. sororcula	*A. fraterculus* C2	*A. pseudoparallela*	-0.026262 ±0.000045	-3.92	0.001890
A. sororcula	*A. fraterculus* C2	*A. bistrigata*	-0.022941 ±0.000044	-3.48	0.008551
A. sororcula	*A. distincta*	*A. grandis*	-0.002851 ±0.000053	-0.39	1.000000
A. sororcula	*A. distincta*	*A. pseudoparallela*	0.004545 ±0.000059	0.59	1.000000
A. sororcula	*A. distincta*	*A. bistrigata*	-0.00854 ±0.000058	-1.12	0.731640
A. sororcula	*A. turpiniae*	*A. grandis*	-0.013734 ±0.000048	-1.97	0.194268
A. sororcula	*A. turpiniae*	*A. pseudoparallela*	-0.009207 ±0.000052	-1.28	0.581544
A. sororcula	*A. turpiniae*	*A. bistrigata*	-0.016295 ±0.00005	-2.30	0.105390
A. sororcula	*A. fraterculus* C1	*A. grandis*	-0.025018 ±0.00004	-3.95	0.001890
A. sororcula	*A. fraterculus* C1	*A. pseudoparallela*	-0.017839 ±0.000045	-2.67	0.047646
A. sororcula	*A. fraterculus* C1	*A. bistrigata*	-0.01978 ±0.000041	-3.09	0.017060

A. fraterculus C2	*A. distincta*	*A. grandis*	0.016256 ±0.000044	2.45	0.079050
A. fraterculus C2	*A. distincta*	*A. pseudoparallela*	0.032028 ±0.000049	4.58	0.000420
A. fraterculus C2	*A. distincta*	*A. bistrigata*	0.017951 ±0.000046	2.65	0.047646
A. fraterculus C2	*A. turpiniae*	*A. grandis*	0.011325 ±0.000037	1.85	0.223699
A. fraterculus C2	*A. turpiniae*	*A. pseudoparallela*	0.02262 ±0.000044	3.42	0.008750
A. fraterculus C2	*A. turpiniae*	*A. bistrigata*	0.011849 ±0.000042	1.83	0.224404
A. fraterculus C2	*A. fraterculus* C1	*A. grandis*	0.002487 ±0.000031	0.45	1.000000
A. fraterculus C2	*A. fraterculus* C1	*A. pseudoparallela*	0.012919 ±0.000034	2.20	0.129029
A. fraterculus C2	*A. fraterculus* C1	*A. bistrigata*	0.005638 ±0.000032	1.00	0.788090
A. distincta	*A. turpiniae*	*A. grandis*	-0.007884 ±0.000053	-1.08	0.732992
A. distincta	*A. turpiniae*	*A. pseudoparallela*	-0.01544 ±0.00006	-1.99	0.194229
A. distincta	*A. turpiniae*	*A. bistrigata*	-0.008339 ±0.000058	-1.09	0.732992
A. distincta	*A. fraterculus* C1	*A. grandis*	-0.017385 ±0.00004	-2.75	0.041559
A. distincta	*A. fraterculus* C1	*A. pseudoparallela*	-0.021886 ±0.000046	-3.24	0.012758
A. distincta	*A. fraterculus* C1	*A. bistrigata*	-0.01215 ±0.000042	-1.88	0.222677
A. turpiniae	*A. fraterculus* C1	*A. grandis*	-0.01101 ±0.000035	-1.87	0.222677
A. turpiniae	*A. fraterculus* C1	*A. pseudoparallela*	-0.010237 ±0.00004	-1.63	0.334714
A. turpiniae	*A. fraterculus* C1	*A. bistrigata*	-0.005164 ±0.000036	-0.86	0.905868

[a] *A. serpentina* was used as outgroup and mapping reference.

[b] Average *D*-statistic and standard error (SE) were calculated by m-delete blocked Jackknife approach.

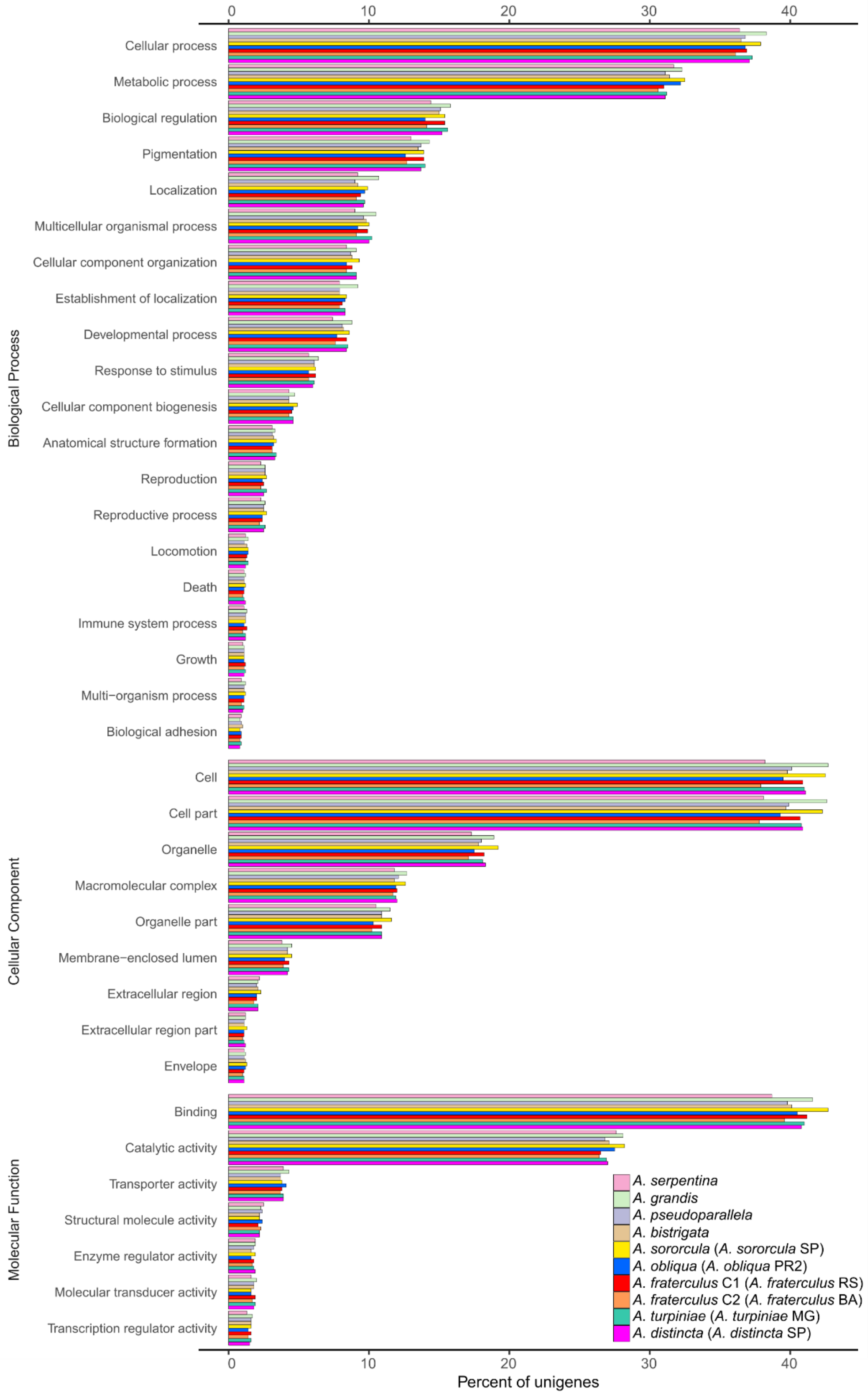
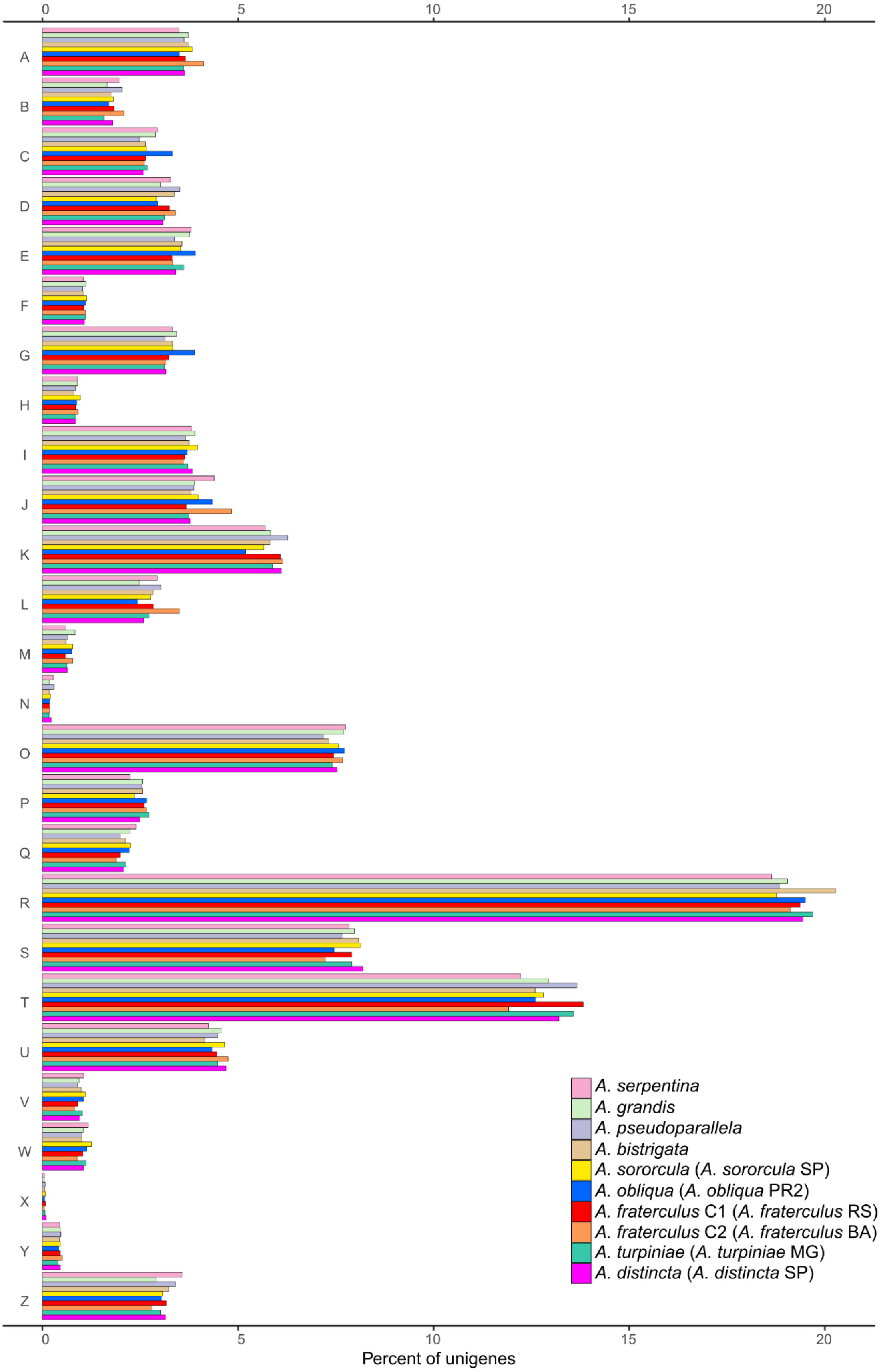


Figure S1. Functional annotation using gene ontology classification of unigenes from female reproductive transcriptomes of 10 *Anastrepha* lineages. The x-axis indicates percentage of unigenes. The y-axis indicates level 2 GO terms with percentages higher than 1%.



0
5
10
15
20
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
Percent of unigenes
A. serpentina
A. grandis
A. pseudoparallela
A. bistrigata
A. sororcula (A. sororcula SP)
A. obliqua (A. obliqua PR2)
A. fraterculus C1 (A. fraterculus RS)
A. fraterculus C2 (A. fraterculus BA)
A. turpiniae (A. turpiniae MG)
A. distincta (A. distincta SP)

Figure S2. Functional annotation using KOG classification of unigenes from female reproductive transcriptomes of 10 *Anastrepha* lineages. The x-axis indicates the percentage of unigenes. The y-axis indicates the following functional category of KOG groups: A, RNA processing and modification; B, Chromatin structure and dynamics; C, Energy production and conversion; D, Cell cycle control, cell division, chromosome partitioning; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; G, Carbohydrate transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; J, Translation, ribosomal structure and biogenesis; K, Transcription; L. Replication, recombination and repair; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; O, Posttranslational modification, protein turnover, chaperones; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown; T, Signal transduction mechanisms; U, Intracellular trafficking, secretion, and vesicular transport; V, Defense mechanisms; W, Extracellular structures; Y, Nuclear structure; and Z, Cytoskeleton.

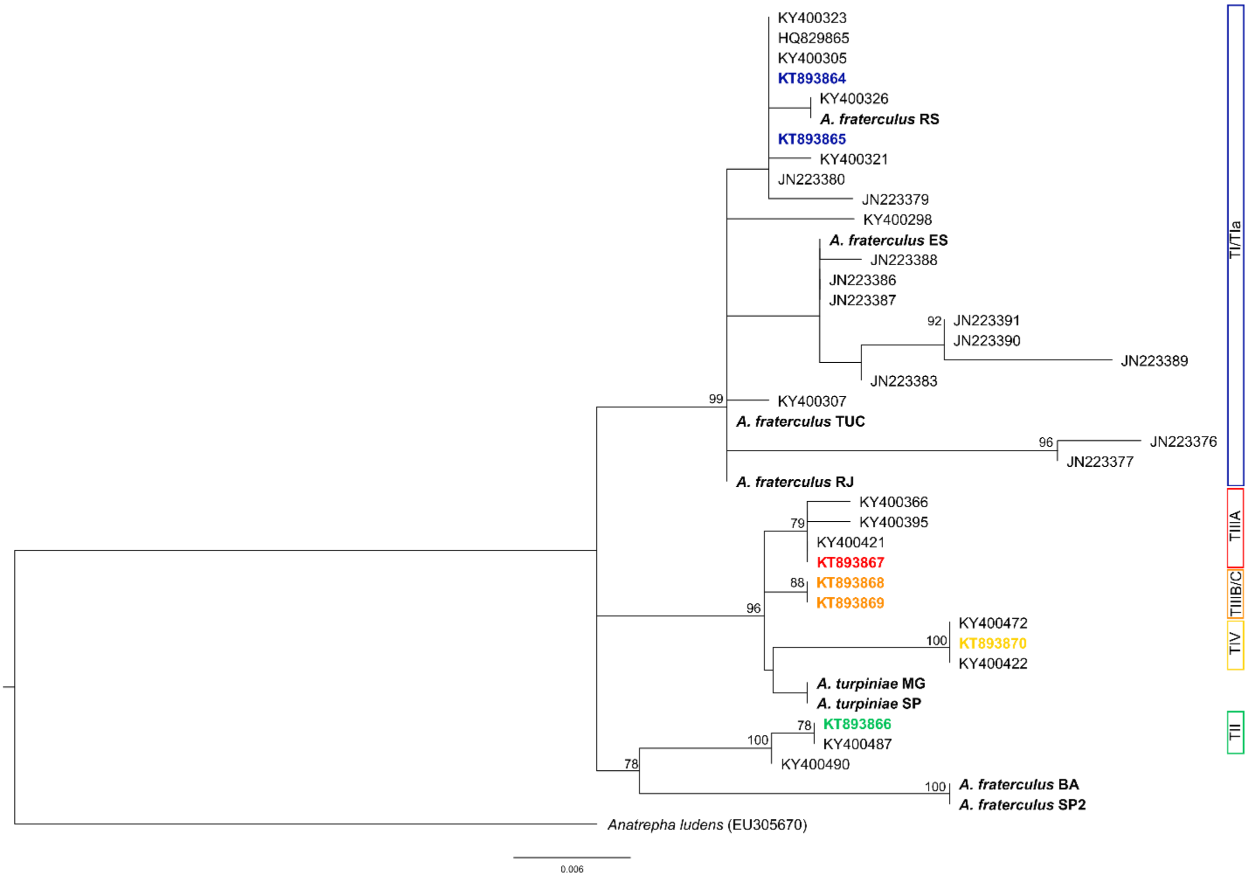


Figure S3. Maximum likelihood phylogeny among *A.fraterculus* complex and *A. turpiniae* specimens based on 517pb from nuclear ribosomal internal transcribed spacer 1 (ITS1). Bootstrap support higher than 75% are shown above the nodes. Samples from this study are indicated in bold. Colors indicate *A. fraterculus* complex lineages identified by Sutton et al. (2015).

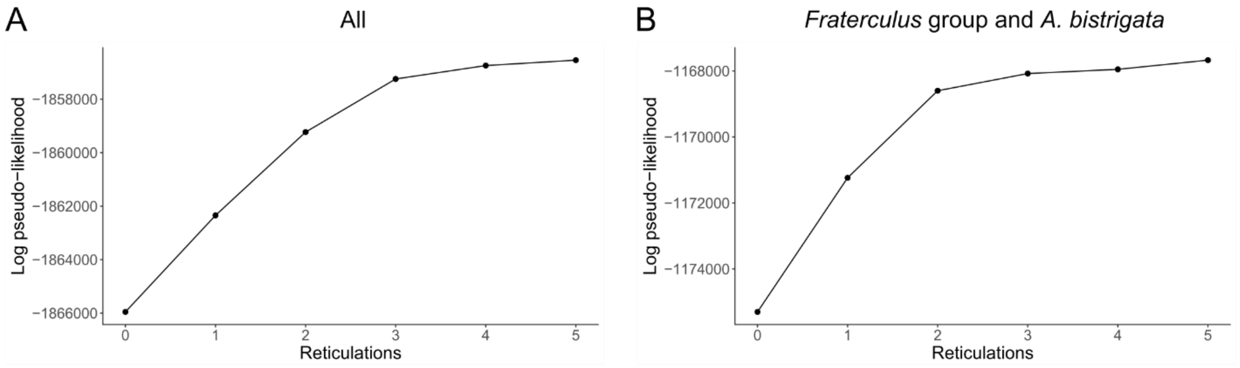


Figure S4. Selection of the optimum species network using all samples (A) and a subset of samples from the *fraterculus* group and *A. bistrigata* (B) indicating that stationary phase of log pseudo-likelihood starts at three reticulations in both datasets.

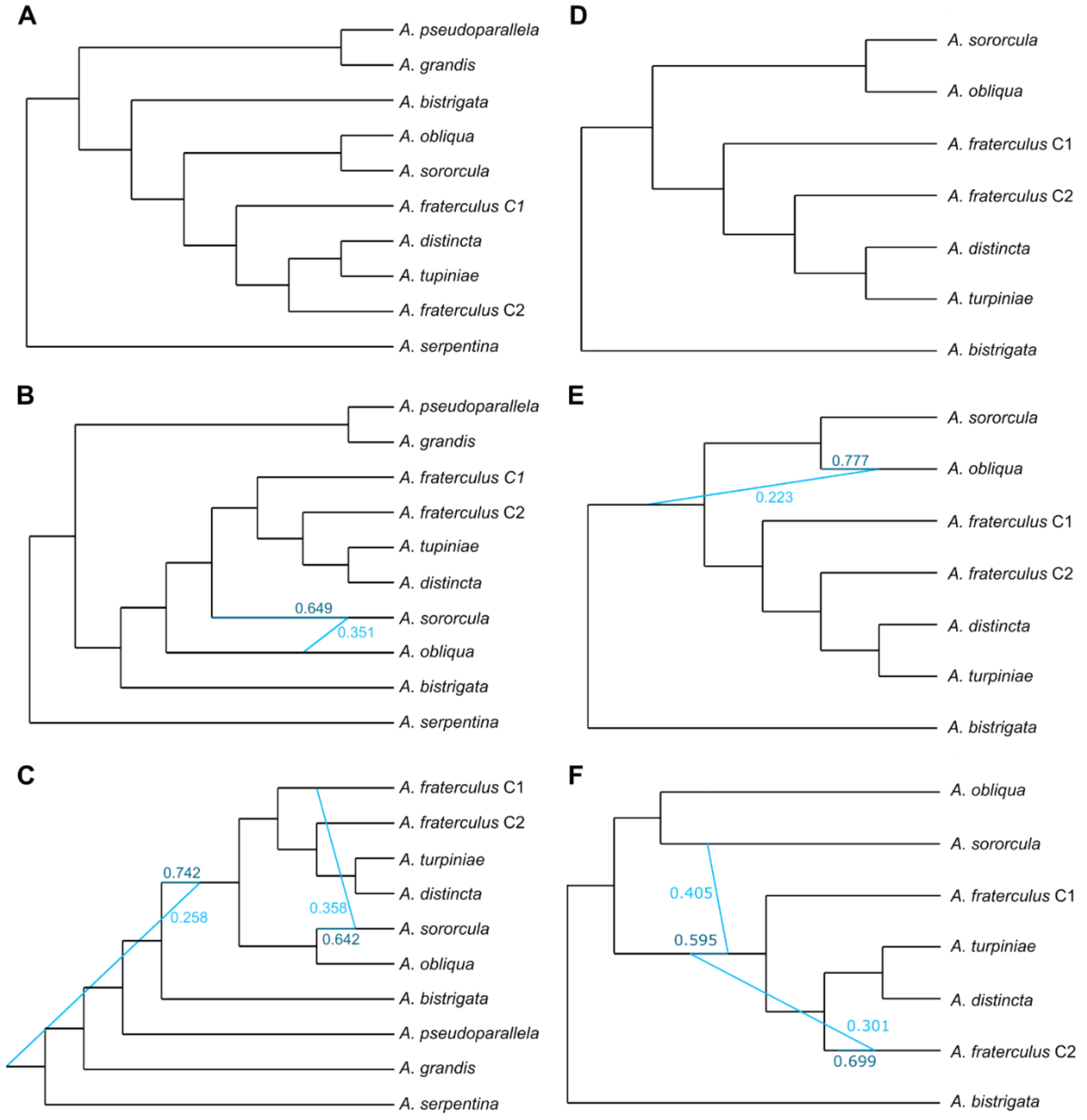


Figure S5. Pseudo-maximum likelihood species networks from different *Anastrepha* lineages. Networks were inferred using 3,220 nuclear genes from *Anastrepha* lineages with 0 (A), 1 (B) and 2 (C) reticulations and using 3,045 nuclear genes from *fraterculus* group lineages and *A. bistrigata* (*striata* species group) with 0 (D), 1 (E) and 2 (F) reticulations. Inheritance probabilities (γ) are shown in sky-blue.

References

SUTTON, B. D. et al. Nuclear ribosomal internal transcribed spacer 1 (ITS1) variation in the *Anastrepha fraterculus* cryptic species complex (Diptera, Tephritidae) of the Andean region. **ZooKeys,** v. 540, 2015.

Final considerations

Several studies that have investigated the evolutionary histories of species in the genus *Anastrepha* have indicated a rapid diversification and little support to most lineages. The use of large genetic datasets may contribute to solve this complex biological scenario, which is why we used RNA-seq data from reproductive and head tissues to investigate both patterns of molecular evolution and levels of expression of *A. fraterculus* and *A. obliqua*, two closely related species of the *fraterculus* group. From this analysis, we found that a substantial portion of sex-biased genes was expressed in reproductive tissue. In contrast, few genes from cephalic tissues showed differential expression due to sex. As reported for other species, male-biased expressed genes showed faster evolutionary rates when compared to female-biased and unbiased genes due to positive selection and relaxed constraints on these genes. Moreover, some of the male-biased and positively selected genes are involved with courtship behavior and fertility, suggesting that these genes may be involved in the differentiation process of these species, and perhaps other species of this group. We also analyzed reproductive transcriptomes from 10 key lineages of *Anastrepha* focusing on the *fraterculus* group to infer the evolutionary history of these taxa using a phylogenomic framework. This data enabled us to establish a robust species tree for these lineages, which revealed that *A. fraterculus* (*sensu latu*) is not a monophyletic group, which agrees with the hypothesis of cryptic diversity in this taxon. Furthermore, we empirically confirmed that the high levels of gene tree discordance are due to incomplete lineage sorting and ancestral hybridization. Although our efforts to understand the mechanisms behind the diversification of this group, there are relevant questions to be addressed, some of which might be answered using large-scale genetic and morphological data from samples collected throughout the geographic distribution of the species and including more species.