UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
PROGRAMA DE PÓS GRADUAÇÃO EM
BIODIVERSIDADE ANIMAL

Francine Cenzi De Ré

EVOLUÇÃO MOLECULAR E PADRÕES MACRO E MICRO EVOLUTIVOS EM *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

Santa Maria, RS, Brasil
2016

Francine Cenzi De Ré

EVOLUÇÃO MOLECULAR E PADRÕES MACRO E MICRO EVOLUTIVOS EM *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

Tese apresentada ao Curso de Doutorado do Programa de Pós Graduação em Biodiversidade Animal, área de concentração em Sistemática e Biologia Evolutiva, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Doutora em Biodiversidade Animal.**

Orientador: Prof. Dr. Elgion Lucio da Silva Loreto
Co-orientadora: Profª. Dra. Lizandra Jaqueline Robe

Santa Maria, RS, Brasil
2016

Francine Cenzi De Ré

EVOLUÇÃO MOLECULAR E PADRÕES MACRO E MICRO EVOLUTIVOS EM *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

Tese apresentada ao Curso de Doutorado do Programa de Pós Graduação em Biodiversidade Animal, área de concentração em Sistemática e Biologia Evolutiva, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Doutora em Biodiversidade Animal.**

Aprovada em 28 de março de 2016:

_____
Elgion Lúcio da Silva Loreto, Dr.
(Presidente/Orientador)

_____
Lizandra Jaqueline Robe, Drª
(Presidente/Co-orientadora)

_____
Vera Lúcia da Silva Valente, Drª (UFRGS)
(Examinadora)

_____
Maríndia Deprá, Drª (UFPEL)
(Examinadora)

_____
Ana Lucia Anversa Segatto (UFSM)
(Examinadora)

_____
Daniel Ângelo Sganzerla Graichen (UFSM – CESNORS)
(Examinador)

Santa Maria, RS
2016

Dedico esta tese aos meus pais Darci e Marinez,
meus irmãos Fabrício e Francisco e ao meu noivo Cássio.

“Freedom of thought is best promoted by the gradual illumination of men’s minds which follows from the advance of science.”

Charles Darwin

AGRADECIMENTOS

Ao Programa de Pós Graduação em Biodiversidade Animal (PPGBA – UFSM).

Ao meu orientador Prof. Dr. Elgion pela acolhida no laboratório ao longo de todos esses anos, da graduação ao doutorado, pela confiança depositada no meu trabalho ao me sugerir um projeto tão lindo e ao mesmo tempo tão desafiador, pela sua dedicação e ensinamentos constantes. Também gostaria de agradecer o apoio nas questões pessoais que enfrentei ao longo do doutorado que foi fundamental para que eu retomasse o trabalho com disposição e felicidade.

A Prof. Dra. Lizandra, minha querida amiga e também orientadora, que como uma mãe conduz o seu filho a dar os primeiros passos, ela me ensinou com muita paciência, a dar os primeiros passos na ciência desde a graduação até os dias de hoje. Obrigada pela confiança, amizade, conselhos, incentivos e toda a sua dedicação. Com certeza, sou grata por tudo que aprendi com você, tanto como profissional como ser humano. Aproveito para estender meus agradecimentos, a toda sua família: Sandro, Isabela e Charles por abrirem as portas de casa e me receberem com tanto carinho sempre que precisei.

A minha banca examinadora Daniel, Vera, Maríndia, Ana Lucia, André e Marlise pela atenção, críticas, sugestões e por lerem meu trabalho em um curto espaço de tempo.

Aos meus queridos amigos Larissa, Paloma, Sinara, Valéria, Pedro, Michelle, Vanessa, Gabriela, Tailini, Mauro, João Pedro, que me acompanharam nessa longa caminhada e sempre me ajudaram em todos os momentos que precisei. Obrigada pelos momentos de alegria e descontração (que não foram poucos), abraços sinceros e por me acolherem nos momentos de angústia. Vocês foram fundamentais!

A Sinara, especialmente, pelo apoio incondicional dia e noite, literalmente.

Aos demais colegas do lab: Stela, Dani, Mari, Marcos, Camila, Zé, Tiago, Raquel, Nader e Gabriel, valeu por tudo! Trabalhar com vocês é a certeza de um excelente ambiente de trabalho. Aproveito para estender meus agradecimentos a todos os ex colegas que passaram pelo laboratório e conviveram comigo ao longo desses anos.

Aos colegas e amigos do LABEEM – FURG, Thaisa, Henrique, Daiana, Bruna e Diego que sempre me acolheram muito bem e estavam sempre dispostos a me auxiliar.

Ao Gabriel por todos os ensinamentos transmitidos durante o desenvolvimento desta tese e por estar sempre disposto a me ajudar com questões técnicas e discutir aspectos do meu trabalho.

A Stela, Liz e ao Pedro pelas coletas de Cestrum.

Ao Ronaldo, por todas as amostras seqüenciadas, pelos momentos de descontração no laboratório e pela grandiosa amizade.

As minhas amigas de Jacutinga: Andréia, Dani, Giovana e Jossana, que apesar da distância sempre torceram por mim, compreendendo as minhas ausências e retribuindo com amizade sincera.

As minhas amigas de infância: Karin, Karine e Monique por todo apoio e carinho de sempre.

Ao meu noivo Cássio, por ser meu professor de Linux e informática em geral, me auxiliar com os programas, edição das figuras, mas principalmente por me acolher nos momentos de aflição, me acalmar e comemorar comigo as minhas vitórias. Seu carinho e amor, presentes em todos os momentos, são essenciais para mim e me tornam uma pessoa muito mais feliz. Obrigada, meu amor! Agradeço também a minha sogra, Maria, pelo carinho e acolhida de sempre.

Ao meu pai Darci, minha mãe Marinez e aos meus manos Fabrício e Francisco pelo amor incondicional, incentivo, carinho e por me confortarem nos momentos de tristeza e preocupação. Obrigada por fazerem o possível para que eu chegasse até aqui e por tornarem a nossa família esse "ninho" de amor e união.

A minha querida amiga Naninha, por todo amor e carinho comigo, pelos lanches, comidinhas de mãe enquanto fiquei longe de casa, mates e cafés.

A Laura, minha afilhada, e Angélica pelo apoio de sempre e por estarem diariamente perto, no meu coração.

A Ana Luisa, minha afilhada, ao Pedro e a Lari por alegrarem até os meus dias mais difíceis.

Enfim, serei eternamente grata, de coração, a todas as pessoas que me ajudaram de perto ou de longe, com um sorriso confortante, palavras de incentivo ou com técnicas e conhecimentos científicos.

RESUMO

EVOLUÇÃO MOLECULAR E PADRÕES MACRO E MICRO EVOLUTIVOS EM *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

AUTORA: Francine Cenzi De Ré
ORIENTADOR: Elgion Lucio da Silva Loreto
CO-ORIENTADORA: Lizandra Jaqueline Robe

O grupo *flavopilosa* foi proposto por Wheeler e colaboradores em 1962. Em termos taxonômicos, parece ser um grupo monofilético, pertencente à radiação *virilis-repleta* do subgênero *Drosophila.* No entanto, o exato posicionamento do grupo dentro dessa radiação ainda é alvo de debate, havendo ampla incongruência entre diferentes marcadores. De acordo com a classificação proposta até o momento, o grupo *flavopilosa* é dividido em dois subgrupos, *nesiota* e *flavopilosa*, que compreendem um total de 16 espécies mais *Drosophila incompta*, organismo modelo desta tese. *D. incompta* é estritamente adaptada à exploração de flores do gênero *Cestrum*, tanto como sítios de oviposição, quanto como substrato para o desenvolvimento larval. Para isso, essa espécie desenvolveu uma série de adaptações fenotípicas exclusivas que facilitam a exploração desse recurso. Além disso, acredita-se que *D. incompta* apresente uma série de adaptações moleculares a sua especialização ecológica, e isto deve estar particularmente refletido no conjunto gênico de receptores olfativos (ROs) e gustativos (RGs) da espécie. Também em virtude de seus padrões ecológicos restritos, a distribuição de *D. incompta* é completamente dependente da distribuição de seus hospedeiros, que parecem ser abundantes na região Neotropical. No Brasil, especificamente, as plantas do gênero *Cestrum* são distribuídas ao longo da Mata Atlântica e do Cerrado, os quais apresentam formações vegetacionais que mudaram consideravelmente nos períodos glaciais e interglaciais do Quaternário. De fato, as oscilações climáticas desse período, marcado pela redução da temperatura e umidade do Hemisfério Sul, parecem ter ocasionado contrações na distribuição da Mata Atlântica e sua substituição por outros tipos de formações vegetais, condizentes ao clima, como o Cerrado e a Caatinga. Dada a complexidade dos cenários macroevolutivos, microevolutivos e moleculares envolvendo *D. incompta*, o objetivo geral desta tese é caracterizar os aspectos filogenéticos, moleculares e filogeográficos associados à especialização ecológica de *D. incompta* às flores do gênero *Cestrum* de Solanaceae. Para isso, caracterizamos o genoma mitocondrial de *D. incompta*, inferimos seu correto posicionamento filogenético a partir de dados de filogenômica, identificando as possíveis fontes de incongruência, caracterizamos o repertório gênico de ROs e RGs da espécie e analisamos seus padrões de diversidade e estruturação ao longo da região Sul do Brasil. Nossos resultados mostram que o genoma mitocondrial de *D. incompta* apresenta perfeita sintenia com as outras espécies do gênero *Drosophila*, sendo constituído de 13 genes codificadores de proteínas, 22 *tRNAs*, 2 *rRNAs* e uma região rica em A-T. Além disso, análises de polimorfismo ao longo deste genoma indicam a presença de níveis pronunciados de diversidade intra-específica. Em relação ao posicionamento filogenético, foi possível demonstrar que enquanto os genomas mitocondriais suportam o agrupamento de *D. incompta* com *D. mojavensis*, os genes nucleares recuperam *D. incompta* e *D. virilis* como espécies irmãs. Como, em geral, estes posicionamentos são mantidos mesmo quando os efeitos da saturação são controlados, acredita-se que a incongruência entre os dois conjuntos de dados deve ser um reflexo de diferenças em suas histórias evolutivas. Em relação ao repertório de ROs e RGs, encontramos 28 e 12 genes pertencentes a cada uma destas famílias no genoma de *D. incompta*, respectivamente. A redução no número de genes com relação a outras espécies de *Drosophila* parece ser adaptativo devido ao padrão de ecologia restrita. Ainda assim, esses genes parecem estar sob efeito de seleção purificadora. Por fim, as análises filogeográficas mostram que as populações de *D. incompta* da região Sul do Brasil passaram por um evento de expansão populacional entre 175 e 100 mil anos atrás seguido de um período de estabilidade que se estende até os dias atuais. Os altos níveis de diversidade e a ausência de um padrão geográfico de estruturação genética parecem refletir a ocorrência de altos níveis de fluxo gênico na espécie, como resposta às oscilações na abundância e disponibilidade dos recursos explorados.

Palavras-chave : *Drosophila incompta*. Filogeografia. Incongruência mito-nuclear.

ABSTRACT

MOLECULAR EVOLUTION AND MACRO AND MICRO EVOLUTIONARY PATTERNS IN *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

AUTHOR: Francine Cenzi De Ré
ADVISOR: Elgion Lucio da Silva Loreto
CO-ADVISOR: Lizandra Jaqueline Robe

The *flavopilosa* group was proposed by Wheeler and colleagues in 1962. In taxonomic terms, it seems to be a monophyletic group, belonging to *virilis-repleta* radiation of the *Drosophila* subgenus. However, the exact position of the group within that radiation is still under discussion and there is wide incongruence between different markers. According to the classification proposed to date, the *flavopilosa* group is divided into two subgroups, *nesiota* and *flavopilosa*, comprising a total of 16 species plus *Drosophila incompta*, model organism of this thesis. *D. incompta* is strictly adapted to the exploitation of the *Cestrum* flowers, both as oviposition sites, and as a substrate for larval development. For this, this species has developed a number of unique phenotypic adaptations that facilitate the exploitation of this resource. Furthermore, it is believed that *D. incompta* presents a series of molecular adaptations regarding their ecological specialization, and this must be reflected particularly in the gene set of olfactory and gustatory receptors. Also because of their limited ecological patterns, the distribution of *D. incompta* is completely dependent on the distribution of their hosts, who seem to be abundant in the Neotropics. In Brazil, specifically, the plants of *Cestrum* genus are distributed over the Biomes Cerrado and Atlantic Forest, two vegetation formations which have changed considerably in the glacial and interglacial periods of the Quaternary. In fact, climate fluctuations that period, marked by the reduction of temperature and humidity in the Southern Hemisphere appear to have caused contractions in the distribution of the Atlantic Forest and its replacement by other, consistent climate, such as the Cerrado and Caatinga. Given the complexity of macro, microevolutionary and molecular scenarios involving *D. incompta*, the general aim of this thesis is to characterize the phylogenetic, phylogeographic and molecular aspects associated with ecological specialization of *D. incompta* to the flowers of the *Cestrum*. For this, we characterize the mitochondrial genome of *D. incompta*, infer its correct phylogenetic position from phylogenomic data, identifying possible sources of incongruity, characterized the gene repertoire of ORs and GRs and analyze their patterns of diversity and structuring throughout Southern Brazil. Our results show that the mitochondrial genome of *D. incompta* shows perfect synteny with the other species of the genus *Drosophila*, consisting of 13 protein-coding genes, 22 *tRNAs*, 2 *rRNAs* and a A-T rich region. Besides, polymorphism analysis over this genome indicate the presence of pronounced levels of intra-specific diversity. Regarding the phylogenetic position, it was demonstrated that while the mitochondrial genome supports the clade formed by *D. incompta* and *D. mojavensis*, the nuclear genes recover *D. incompta* and *D. virilis* as sister species. As, in general, these positions are maintained even when the effects of saturation are controlled, it is believed that the incongruity between the two data sets must be a reflection of differences in their evolutionary histories. Regarding the repertoire of ORs and GRs, we find 28 and 12 genes belonging to each of these families in the genome of *D. incompta*, respectively. The reduction in the number of genes relative to other species of *Drosophila* appears to be adaptative due to restricted ecological pattern. Still, these genes appear to be under effect of purifying selection. Finally, the phylogeographic analysis shows that populations of *D. incompta* of southern Brazil experienced a population expansion event between 175 and 100 thousand years ago followed by a period of stability that extends to the present day levels of diversity and the lack of a geographic pattern of genetic structure seem to reflect the occurrence of high levels of gene flow in the species, in response to changes in the abundance and availability of the resources exploited.

Keywords : *Drosophila incompta.* Phylogeography. Mito-nuclear discordance.

SUMÁRIO

1 INTRODUÇÃO GERAL

1.1 A FAMÍLIA DROSOPHILIDAE

Drosophilidae é considerada uma das famílias mais diversas e amplamente distribuídas da ordem Diptera, apresentando aproximadamente 4.200 espécies distribuídas ao longo de 77 gêneros (BÄCHLI, 2015). Segundo THROCKMORTON (1975), a família Drosophilidae teve origem nas regiões tropicais do Velho Mundo, há cerca de 50 milhões de anos atrás, no período conhecido como Eoceno. Em termos taxonômicos, é uma família monofilética (GRIMALDI, 1990; REMSEN AND O'GRADY, 2002) e subdividida em duas subfamílias, Steganinae e Drosophilinae (TODA, 2007). Representantes de Drosophilidae podem ser encontrados nos mais diversos ambientes, ocorrendo em planícies, pântanos e savanas, a nível do mar ou em montanhas, desde os trópicos até as tundras. Entretanto, os bosques e as florestas são os centros de maior abundância (THROCKMORTON, 1975).

Membros da família Drosophilidae desempenham um papel ecológico importante nos ecossistemas aos quais estão inseridos, uma vez que, nos estágios imaturos, alimentam-se de organismos em processo de fermentação. Neste sentido, THROCKMORTON (1975), sugere que a evolução dessa família deu-se, principalmente, a partir de ancestrais que utilizavam fungos crescidos em folhiço como principal fonte de alimento, e a diversificação de substratos seria o resultado de adaptações naturais, com uma certa ênfase na utilização de frutos. A suscetibilidade à evolução de nicho associada à diversificação dos substratos explorados parece fornecer uma ótima explicação para o sucesso evolutivo da família (ROBE et al. 2010) e sua ampla distribuição (THROCKMORTON, 1975).

Entre os 77 gêneros contemplados na família Drosophilidae é importante destacar o gênero *Drosophila*, que compreende um total de 1.188 espécies descritas (BÄCHLI, 2015). Este é, também, o gênero mais representado no Brasil, onde já foi reportada a presença de, pelo menos, 181 espécies (GOTTSCHALK et al. 2008). Diferentes regiões brasileiras têm sido amostradas com o intuito de acessar a diversidade de drosofilídeos e, principalmente, *Drosophila* (GARCIA et al. 2012; POPPE et al. 2012, 2014, ROQUE et al. 2013). Entretanto, essa diversidade pode estar sendo subestimada, devido ao método de amostragem, que, em sua maioria, consiste na metodologia de coleta tradicional descrita por TIDON E SENE (1988). Nessas coletas, indivíduos frugívoros ou generalistas são favorecidos, enquanto espécies antófilas ou de nicho restrito, tendem a ser excluídas. Ainda assim, estudos

específicos têm sido conduzidos com espécies antófilas, pela coleta e acondicionamento das flores, em laboratório, até a eclosão ROBE et al. (2013).

1.1.1 O grupo *flavopilosa* de *Drosophila*

Através da análise minuciosa das genitálias de machos e fêmeas, o grupo *flavopilosa* foi proposto no início da década de 60, do século passado, por WHEELER et al. (1962). Esses autores sugerem, inicialmente, que o grupo *flavopilosa* pertence ao subgênero *Drosophila* e é formado por *D. flavopilosa* e outras 13 espécies distribuídas ao longo da região Neotropical. Atualmente, sabe-se que esse grupo é formado por dois subgrupos que compreendem um total de 17 espécies: o subgrupo *flavopilosa* que inclui *D. acroria, D. crossoptera, D. lauta, D. cordeiroi, D. cestri, D. flavopilosa* e *D. hollisae;* e *nesiota*, formado por *D. incompta*, *D. melina, D. nesiota*, *D. gentica* e *D. mariahelenae,* e mais cinco espécies não alocadas a nenhum subgrupo (BÄCHLI, 2016).

Em relação ao posicionamento filogenético do grupo *flavopilosa*, THROCKMORTON (1975) sugere que o grupo pertence à radiação *virilis-repleta* de *Drosophila* e pode ser derivado tanto da radiação que levou a origem dos grupos *bromelie* e *peruviana*, como da radiação *repleta*. Mais tarde, GRIMALDI (1990) apresenta o grupo *flavopilosa* proximamente relacionado ao clado formado pelas espécies do gênero *Scaptomyza,* levantando assim a possibilidade deste grupo não fazer parte do gênero *Drosophila*. Entretanto, evidências recentes, baseadas em dados moleculares, enfatizam o grupo *flavopilosa* como um dos representantes do gênero *Drosophila*, incluído na radiação *virilis-repleta*, formando um clado irmão ao grupo *annulimana* (ROBE et al. 2005; ROBE et al. 2010). Entretanto, o correto posicionamento do clado formado pelos grupos *flavopilosa* e *annulimana* dentro da radiação *virilis-repleta* é ainda controverso, como pode ser demonstrado pelos resultados obtidos por ROBE et al. (2010), onde diferentes marcadores nucleares apoiaram diferentes resoluções.

Os representantes do grupo *flavopilosa* apresentam ecologia restrita, uma vez que exploram apenas flores do gênero *Cestrum* (Solanaceae) como sítio de oviposição, desenvolvimento larval e alimentação (BRNCIC, 1966; HOFMANN, 1985). Para otimizar a exploração desse recurso, essas moscas compartilham algumas características morfológicas adaptativas, típicas para o micro-hábitat utilizado: 1) são de pequeno a médio porte; 2) têm coloração amarelo claro, que é críptica para as flores de *Cestrum*; 3) as fêmeas possuem espinhos fortes na região externa do ovipositor (utilizados para escarificar a superfície das flores antes da oviposição); e 4) depositam seus ovos em avançado estágio de

desenvolvimento embrionário (a fim de reduzir o período de pupação, dada a natureza efêmera das flores) (BRNCIC, 1983; LUDWIG et al. 2002). No Brasil, foram registradas seis espécies de *Drosophila* do grupo *flavopilosa* (*D. cestri, D. cordeiroi, D. flavopilosa, D. hollisae, D. incompta* e *D. mariahelenae*), em um total de sete diferentes espécies de *Cestrum* (*C. amictum, C. calycinum, C. corymbosum, C. intermedium, C. parqui, C. schlenchtendalii, C. sendtnerianum* e *Sessea brasiliensis*) (SANTOS E VILELA, 2005).

SEPEL et al. (2000) mostram que o período de floração de *Cestrum* não é sincronizado ao longo do ano. Desse modo, quando esse recurso é escasso, as populações de *Drosophila* podem atravessar períodos de *bottlenecks*, que são seguidos de expansões populacionais, nas épocas em que as flores tornam-se abundantes. No entanto, esses mesmos autores sugerem que, na presença do recurso, a taxa de ocupação das flores de *Cestrum* é diferenciada entre as espécies, o que também pode contribuir para os eventos de redução no tamanho populacional. Certamente, a competição entre as espécies do grupo que ocorrem em simpatria e sintopia, bem como a disponibilidade de recursos pertencentes às diferentes espécies de *Cestrum* são fatores chave na compreensão da dinâmica evolutiva destas espécies. Além disso, devido à íntima associação do grupo *flavopilosa* às flores de *Cestrum*, a distribuição dessas espécies é afetada não apenas pelas condições ambientais limitantes ao grupo, como também pelas variáveis climáticas que afetam suas plantas hospedeiras (ROBE et al. 2013).

O padrão de ecologia restrita, que inclui o comportamento alimentar diferenciado em relação às moscas generalistas, já foi reportado para outras espécies de *Drosophila*, como, por exemplo, para *D. sechellia*, que está intimamente relacionada aos frutos de uma única espécie de planta, *Morinda critifolia*, utilizando-os como fonte de alimentação (JONES, 2005; MCBRIDE, 2007). Além dessa espécie, níveis menores de especialização também são encontrados em espécies do grupo *repleta* de *Drosophila*. No *cluster* de *D. buzzatii*, por exemplo, as espécies utilizam cactos em decomposição como fonte exclusiva de alimentação nos estágios larvais (PEREIRA, VILELA E SENE, 1983; MANFRIN E SENE, 2006) e *D. mojavensis* utiliza cactos tanto como fonte de recurso (MATZKIN, 2014) quanto para o acasalamento (KELLEHER E MARKOW, 2009). A fonte do recurso escolhido para alimentação, bem como os sítios de acasalamento e oviposição, são fundamentais para a sobrevivência e adequação de todos os drosofilídeos. Embora um dos grandes objetivos da genética evolutiva seja entender as mudanças moleculares subjacentes à variação fenotípica dentro e entre espécies, grande parte das variações genéticas associadas ao perfil de ecologia

restrita ainda são desconhecidas. O estudo das espécies grupo *flavopilosa*, oferece uma oportunidade ímpar de correlação entre dados genéticos e ecológicos (HOFMANN, 1985).

1.1.1.1 Drosophila incompta

D. incompta pertence ao grupo *flavopilosa*, subgrupo *nesiota* (BÄCHLI, 2016) e é frequentemente encontrada em simpatria e até mesmo sintopia com *D. flavopilosa, D. cordeiroi* e *D. cestri* na região Sul do Brasil (ROBE et al. 2013). Embora a maior parte das espécies do grupo *flavopilosa* ainda não tenham sido incluídas em nenhum estudo filogenético, essas quatro espécies constituem um agrupamento monofilético, no qual *D. incompta* aparece como ramificação basal (ROBE et al. 2010a).

Do ponto de vista ecológico, *D. incompta* pode ser considerada uma espécie altamente epecializada, cuja distribuição e abundância depende da disponibilidade do recurso ecológico que ela explora, neste caso, flores do gênero *Cestrum*. A partir de uma compilação de dados da literatura e resultados de coletas pontuais, SANTOS E VILELA (2005) mostram que *D. incompta* já foi coletada em pelo menos oito espécies do gênero *Cestrum* (*C amictum, C. calycinum, C. corymbosum, C. intermedium, C. nocturnum, C. parqui, C. Schlechtendalii* e *C. Sendtnerianum*) e também em *Sessia brasiliensis.* Uma abordagem mais detalhada do perfil de oviposição desta espécie mostra que, ao contrário das outras espécies do grupo *flavopilosa* que ovipositam em flores fechadas, *D. incompta* tem preferência por ovipositar em flores abertas (NAPP E BRNCIC, 1978). De fato, coletas recentes realizadas pelo nosso grupo de pesquisa (entre 2010 – 2015) mostraram a predominância significativa de *D. incompta* nas eclosões obtidas a partir de locais em que as flores já estavam abertas. Por outro lado, nos locais em que as flores ainda estavam fechadas no momento da coleta, houve um número extremamente significativo de *D. cestri* (dados não publicados). Estas duas espécies também parecem apresentar certa diferenciação em termos de nicho abiótico, e embora *D. incompta* pareça ser mais tolerante que *D. cestri* em relação a variáveis relacionadas a temperatura e precipitação, as duas se sobrepõem na maior parte de suas distribuições (ROBE et al. 2013).

1.2 GENÔMICA E EVOLUÇÃO MOLECULAR

Há décadas, as espécies do gênero *Drosophila* são conhecidas por auxiliarem na elucidação de diversos aspectos da biologia evolutiva, tais como: genética de populações, evolução ecológica, fenômenos de especiação, filogenia, evolução a nível de genoma e

desenvolvimento (POWELL, 1997). Neste sentido, o grupo *flavopilosa*, consideradas as suas propriedades anteriormente mencionadas, apresenta potencial para ser utilizado como um modelo no estudo da evolução de espécies com ecologia restrita, especiação geográfica *versus* ecológica e padrões e impactos da competição intra/interespecífica (ROBE et al. 2013).

Os avanços recentes das tecnologias de sequenciamento em larga escala facilitam o acesso ao genoma completo de qualquer espécie, inclusive aquelas que, por apresentarem padrões ecológicos peculiares, não podem ser mantidas em laboratório como por exemplo, *D. incompta.* Além disso, tais tecnologias possibilitam estudar os padrões de variação, moldados ou não pela ação da seleção natural, em todos os níveis de complexidade desde os genes até as populações.

Além do modelo de estudos tradicional, *D. melanogaster*, até o momento, há 25 genomas de *Drosophila* disponíveis para download e análise: *D. albomicans, D. buzzatii, D. americana, D. ananassae, D. biarmipes, D. bipectinata, D. busckii, D. elegans, D. erecta, D. eugracilis, D. ficusphila, D. grimshawi, D. kikkawai, D. miranda, D. mojavensis, D. perimilis, D. pseudoobscura, D. rhopaloa, D. sechellia, D. simulans, D. suzukii, D. takahashii, D. virilis, D. willistoni, D. yakuba.* Essas espécies apresentam uma variedade de estratégias comportamentais e ecológicas, variando em termos alimentares desde hábitos generalistas, como os encontrados em *D. ananassae, D. melanogaster* e *D. simulans*, até hábitos especialistas, como os apresentados por *D. sechellia*, que se alimenta de um único tipo de fruto (CLARK et al. 2007).

1.2.1 O genoma mitocondrial de insetos

O genoma mitocondrial de insetos é o mais amplamente estudado dentre todos os invertebrados, com quase 500 espécies sequenciadas até o momento. A representatividade deste conjunto de dados é excelente, uma vez que ele inclui todas as ordens de insetos (CAMERON, 2014). Em geral, o genoma mitocondrial nestas espécies é representado por uma molécula pequena e circular, que apresenta de 15–20 kb de tamanho e compreende 13 genes codificadores de proteínas, envolvidos no processo de fosforilação oxidativa, dois genes para *rRNAs* ribossomais e vinte e dois genes para *tRNAs* (revisão em BOORE, 1999). Além disso, estes genomas ainda apresentam uma região não codificadora, conhecida como região rica em A+T, que está diretamente relacionada ao controle da replicação e da transcrição do mtDNA.

Segundo LANG et al. (1999), a organização deste conjunto de genes parece ser comum ao longo do mtDNA de diferentes espécies animais, sendo conservada dentro dos filos. No entanto, estudos recentes revelam que este padrão de uniformidade pode não ser aplicado a alguns grupos animais, especialmente para filos que não apresentam simetria bilateral, incluindo Cnidaria, Ctenophora, Placozoa, e Porifera (LAVROV, 2014).

O primeiro genoma mitocondrial de insetos a ser caracterizado quanto ao número e arranjo gênico foi o de *Drosophila yakuba*, na década de oitenta (CLARY E WOLSTENHOLME, 1985). A partir de então, houve um crescente número de genomas de insetos disponíveis, principalmente nos últimos treze anos, o que tem resultado em um grande fluxo de dados, sendo os genes mitocondriais, as sequências mais comuns no GenBank (CAMERON, 2014). Até o momento há 14 genomas mitocondriais de *Drosophila* disponíveis: *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi* (MONTOOTH et al. 2009), *D. yakuba* (CLARY E WOLSTENHOLME, 1985), *D. melanogaster* (GARESSE, 1988), *D. simulans*, *D. mauritiana* (BALLARD 2000a, b), *D. pseudoobscura* (TORRES et al. 2009), *D. sechellia* (BALLARD 2000a, b) e *D. littoralis* (ANDRIANOV et al. 2010). Além desses, dados brutos dos genomas de *D. elegans*, *D. biarmipes* e *D. rhopaloa* estão disponíveis para download, embora os seus genomas mitocondriais ainda não tenham sido montados e anotados.

1.2.2 Evolução de receptores gustativos e olfativos

Ao longo da evolução, os animais desenvolveram eficientes sistemas sensoriais que permitem avaliar as características do meio externo e são imprescindíveis para a localização e avaliação de recursos alimentares, localização de substratos favoráveis à reprodução, o reconhecimento de parceiros em potencial e, ainda, a evasão de predadores e outras ameaças. Dentro dos invertebrados, os insetos formam o grupo com maior número de espécies conhecidas e estas apresentam diversos tipos de comportamento, variando de acordo com o estilo de vida e habitat (GRIMALDI E ENGLE, 2005). São essas variações que tornam interessantes os estudos com espécies ecologicamente restritas, cujos genomas devem estar repletos de sinapomorfias adaptativas úteis para a exploração de seus recursos peculiares.

Estudos recentes envolvendo genes associados à pigmentação corporal (GILBERT et al. 2007; MATUTE et al. 2009 e WITTKOPP et al. 2009 e WITTKOPP et al. 2010), à recepção de sinais olfativos e gustativos (MCBRIDE, 2007 e SAMBANDAN et al. 2006) e ao metabolismo e detoxificação necessários para o uso de recursos particulares ou à

especialização a determinado hospedeiro (MATZKIN et al. 2006) foram desenvolvidos com drosofilídeos e revelaram padrões interessantes de evolução molecular, tanto no que diz respeito ao número e localização dos fatores genéticos, quanto no que se refere às taxas de substituições não sinônimas encontradas em genes particulares. A comparação dos padrões obtidos indica que muitas mudanças devem ocorrer como resultado da ação de seleção positiva ou a partir de níveis relaxados de restrição seletiva, que ocorrem em associação às mudanças de nicho. RUBIN et al. (2010), por exemplo, mostraram que existem importantes diferenças entre organismos no que diz respeito ao número de cópias de genes envolvidos em diferentes processos celulares e de desenvolvimento.

Neste contexto, os drosofilídeos fornecem um leque de opções para o estudo da adaptação a diferentes condições ecológicas. Além disso, a presença de circuitos neuronais simples, ciclo de vida rápido e altas taxas de reprodução transforma os mesmos em organismos modelo ideais para estudos associados a evolução de genes relacionados ao olfato e paladar. A vasta disponibilidade de ferramentas de bioinformática e a identificação dos genes que fazem parte do complexo de receptores olfativos (RO) e gustativos (RG) (CLYNE et al. 1999; ROBERTSON et al. 2003; HALLEM E CARLSON, 2004) também facilitam esse processo.

Segundo ROBERTSON et al. (2003), em *D. melanogaster*, há 60 genes presentes em cada família de RO e RG, codificando 62 e 68 proteínas, respectivamente. Entretanto, esse repertório gênico pode variar, consideravelmente, entre diferentes espécies de insetos. *Bombyx mori*, por exemplo, possui 48 ROs (WANNER et al. 2007), enquanto em *Linepithema humile* o número de genes que fazem parte deste complexo chega a 367 (SMITH et al. 2011a). Acredita-se que o número de genes em cada complexo esteja diretamente relacionado à diversidade de estímulos experimentados por cada espécie. Neste sentido, a ecologia restrita de *D. incompta* fornece condições ideais para a avaliação dos mecanismos moleculares e evolutivos subjacentes à especialização dos sistemas olfativos e gustativos, permitindo, pela análise comparativa das condições encontradas em outras espécies, a identificação dos processos envolvidos na especialização ecológica.

1.3 FILOGEOGRAFIA: CONCEITOS, MARCADORES E APLICAÇÕES

O termo “filogeografia” foi proposto por Avise e colaboradores em 1987 (AVISE et al. 1987) com o intuito de integrar biólogos evolucionistas nas áreas de filogenia e genética de populações. Este campo da ciência utiliza informações genéticas para estudar a distribuição

geográfica de linhagens genealógicas, especialmente a nível intraespecífico (AVISE, 2000). Através da amostragem adequada dos indivíduos e de seus genes, torna-se possível testar hipóteses biogeográficas e inferir os processos de origem, distribuição e manutenção da biodiversidade, bem como descrever a evolução do isolamento reprodutivo de unidades populacionais (BEHEREGARAY, 2008).

Até meados da década de 60, a maioria dos estudos evolutivos baseava-se na análise de caracteres morfológicos, fisiológicos ou fenotípicos. Embora esses marcadores tenham contribuído para o desenvolvimento de ferramentas de análises, seu uso era bastante limitado a espécies que poderiam ser mantidas em laboratório (TURCHETTO – ZOLET et al. 2013). Atualmente, os estudos filogeográficos contemplam diferentes marcadores moleculares, que variam de acordo com o objetivo e o modelo biológico a ser estudado. O sequenciamento de DNA possibilitou o acesso a genes que apresentam taxas de evolução distintas e que se localizam em diferentes partes do genoma, possibilitando a avaliação dos níveis taxonômicos em diversas escalas geográficas (AVISE, 2000). Entre os marcadores amplamente utilizados em estudos filogeográficos, pode-se destacar o DNA organelar - cloroplasto nas plantas e mitocondrial nos animais - e também fragmentos de DNA provenientes de regiões nucleares não codificadoras, que por suas altas taxas evolutivas, tornam-se ideais para estudos evolutivos baseados no período do Quaternário. Por outro lado, marcadores como microssatélites e AFLPs são mais apropriados para estudos de eventos recentes (HEWITT, 2000).

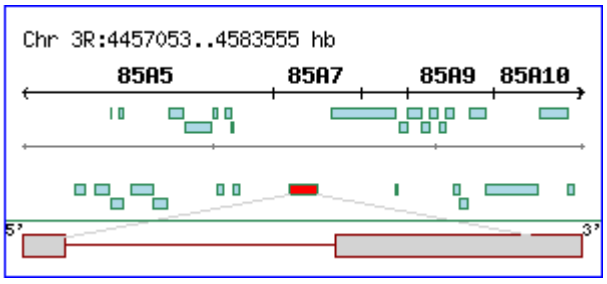
Os marcadores moleculares abordados no artigo IV desta tese são de duas fontes distintas: mitocondrial (*Citocromo Oxidade subunidade I - COI* e *Citocromo Oxidase Subunidade II - COII*) e nuclear *(Hunchback)*. Os genes mitocondriais evoluem mais rapidamente que os nucleares, de modo a acumular um maior número de diferenças, mesmo quando espécies próximas são comparadas (HEBERT et al. 2004b). Além disso, o genoma mitocondrial apresenta uma série de outras vantagens: é haplóide, apresenta herança materna, recombinação ausente ou limitada, é mais resistente à degradação e apresenta regiões altamente conservadas, que possibilitam a construção de *primers* robustos (ROE E SPERLING, 2007). Outro motivo frequentemente apontado refere-se ao fato de que o tamanho efetivo populacional do genoma mitocondrial é cerca de um quarto do genoma nuclear, o que está intimamente relacionado com o seu menor tempo de coalescência (WHITWORTH et al. 2007).

Ainda assim, a inclusão de marcadores nucleares torna-se interessante em face à independência evolutiva apresentada por diferentes loci. Isto é especialmente importante quando consideramos o fato de que a história de um único marcador não reflete, necessariamente, a história das unidades taxonômicas em questão e que muitas vezes a utilização concomitante de diferentes regiões genômicas torna-se fundamental para a compreensão do verdadeiro cenário evolutivo (ROKAS et al. (2003); POLLARD et al. (2006). O DNA nuclear, entretanto, apresenta alguns desafios que devem ser considerados nas análises filogeográficas. Estes incluem recombinação, seleção (não neutralidade), heterozigosidade, polimorfismos de inserção/deleção, baixa divergência e politomia, variação nas taxas e na história dos genes e/ou sítios, dificuldades de amplificação nas reações de PCR e sequenciamento, etc. (ZHANG E HEWITT, 2003).

1.3.1 O gene *Hunchback*

Localizado no cromossomo 3R de *Drosophila melanogaster*, o gene nuclear *Hb* foi primeiramente caracterizado por TAUTZ et al. (1987), e como sua própria tradução para "corcunda" sugere, desempenha um papel fundamental no estabelecimento do gradiente ântero-posterior do zigoto, especialmente no desenvolvimento do tórax nas moscas. Além de suas funções reguladoras no início e durante a segmentação corporal, o *Hb* também é expresso no sistema nervoso em desenvolvimento (*The interactive Fly* – BRODY, 1999). Em *D. melanogaster*, este gene se estende por 6503pb, apresentando dois exons responsáveis por codificar uma proteína de 758 aminoácidos (Figura 1).

Figura 1 – Mapa genômico do cromossomo 3R em *D. melanogaster*, cuja extensão é representada pelos números acima da figura. Retângulos azuis representam os genes adjacentes ao *Hb* e presentes no mesmo cromossomo. O retângulo vermelho indica o gene *Hb*, o qual foi ampliado em dois outros retângulos cinzas para representar os dois exons, conectados entre si pelo íntron, representado pela linha vermelha



Fonte: FlyBase (2016).

1.3.2 Os genes *COI* e *COII*

Devido às características mencionadas anteriormente acerca dos marcadores mitocondriais, os genes *COI* e *COII* são amplamente utilizados em trabalhos que fazem inferências evolutivas intraespecíficas ou entre espécies proximamente relacionadas. Segundo SIMON et al. (1994), o *COI* é o gene mitocondrial mais conservado em termos de evolução em nível de aminoácidos, enquanto que o *COII* apresenta taxas intermediárias de substituição. Além disso, o marcador *COI* é um dos mais populares para estudos genômicos e filogeográficos (AVISE, 1994) e, recentemente, sua popularidade aumentou ainda mais por ser uma ferramenta eficaz na identificação de metazoários, pela técnica do DNA *barcoding* (HEBERT et al. 2003).

O Citocromo oxidase c é o componente da cadeia respiratória que catalisa a redução de oxigênio para a água, e os genes *COI* (subinudade I), *COII* (subunidade II) e *COIII* (subunidade III) formam o núcleo funcional do complexo da enzima (FlyBase site). Em *D. melanogaster*, *COI* e *COII* apresentam, respectivamente, 1536 e 685 pb, localizando-se em posições adjacentes no genoma (Figura 2).

Figura 2 – Representação gráfica da distribuição espacial dos genes ao longo do genoma mitocondrial de *D. melanogaster* com destaque para o genes *COI* e *COII*

Fonte: Mitos Web Server, 2016.

1.3.3 Filogeografia do gênero *Drosophila* no Brasil

A disponibilidade de trabalhos publicados ou que mencionam a área de filogeografia do gênero *Drosophila*, com notável relevância científica, vem aumentando significativamente. Na prática, essa afirmação pode ser evidenciada pela busca das palavras chaves "*Drosophila* phylogeography*"* em um banco de dados de artigos científicos, como por exemplo, o site www.onlinelibrary.wiley.com. Em 10 de março de 2016 esta busca resultou em 2035 artigos científicos e 42 livros que trazem informações relacionados ao tema e contemplam diversos países, inclusive o Brasil.

Entretanto, a despeito do aumento de publicações na área, ainda há uma quantidade limitada de estudos filogeográficos quando apenas as espécies brasileiras do gênero *Drosophila* são consideradas (MANFRIN et al. 2001; DE BRITO et al. 2002a e b, BRISSON

et al. 2005; MANFRIN E SENE, 2006; MORAES E SENE, 2007; MORAES et al. 2009; FRANCO et al. 2010a e b; FRANCO E MANFRIN, 2013; DE RÉ et al. 2014b; GUSTANI et al. 2015). Com exceção de BRISSON et al. (2005), DE RÉ et al (2014a) e GUSTANI et al. (2015), que exploram aspectos filogeográficos de espécies pertencentes a radiação *tripunctata* (*D. polymorpha, D. maculifrons* e *D. ornatifrons,* respectivamente), todos os outros estudos acima mencionados incluem espécies da radiação *virilis-repleta* de *Drosophila* e contemplam fenômenos que afetaram especialmente as regiões Nordeste e Sudeste do Brasil. Neste sentido, a filogeografia de espécies de *Drosophila* na região Sul do Brasil ainda não foi explorada em maiores detalhes, de modo que o impacto das alterações climáticas do passado sobre a fauna de drosofilídeos residentes desta região ainda não é compreendida.

Apesar de alguns trabalhos com enfoques evolutivos terem sido realizados entre as espécies do grupo *flavopilosa*, principalmente no que e refere à estrutura (BRNCIC, 1962) e polimorfismo cromossômico (BRNCIC, 1966, 1967 e 1968), análise de polimorfismos de isozimas (HOFMANN E NAPP, 1984), posicionamento filogenético (ROBE et al. 2005 e 2010; DE RÉ et al. 2014a) e modelagem de nicho ecológico (ROBE et al. 2013), até o momento não há estudos voltados para a filogeografia das espécies do grupo *flavopilosa*. Isto ocorre a despeito dos inúmeros questionamentos relacionados à dinâmica evolutiva destas espécies, que parece envolver sucessivos eventos de *bottleneck* e efeito fundador ou taxas muito altas de migração, como resposta às oscilações na abundância dos recursos explorados.

1.4 OBJETIVOS

1.4.1 Objetivo geral

Caracterizar os aspectos filogenéticos, moleculares e filogeográficos associados à especialização ecológica de *D. incompta* às flores do gênero *Cestrum* de Solanaceae.

1.4.2 Objetivos específicos

- Montar e caracterizar o genoma mitocondrial completo de *D. incompta*;
- Acessar o atual posicionamento filogenético do grupo *flavopilosa* no gênero *Drosophila* a partir de uma abordagem filogenômica;

- Identificar o repertório gênico e analisar os padrões diferenciais de evolução molecular apresentados por genes que podem ter contribuído para a especialização do grupo *flavopilosa* ao seu hospedeiro (como por exemplo, receptores olfativos e receptores gustativos) no genoma de *D. incompta* em comparação a outros genomas de *Drosophila*;
- Avaliar a diversidade intra-específica dentro e entre diferentes populações de *D. incompta*;
- Determinar as forças ecológicas e evolutivas que modelaram a distribuição destas espécies ao longo da região sul do Brasil;
- Avaliar a verossimilhança de cenários envolvendo expansão populacional e/ou *bottlenecks* sucessivos ao longo da evolução do grupo.

2 ARTIGO 1 - CHARACTERIZATION OF THE COMPLETE MITOCHONDRIAL GENOME OF FLOWER-BREEDING *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

Genetica
DOI 10.1007/s10709-014-9799-9

Characterization of the complete mitochondrial genome of flower-breeding *Drosophila incompta* (Diptera, Drosophilidae)

F. C. De Ré · G. L. Wallau · L. J. Robe · E. L. S. Loreto

Received: 12 August 2014 / Accepted: 18 November 2014

Abstract *Drosophila incompta* belongs to the *flavopilosa* group of *Drosophila*, and has a restricted ecology, being adapted to flowers of *Cestrum* as feeding and oviposition sites. We sequenced, assembled, and characterized the complete mitochondrial genome (mtDNA) of *D. incompta.* In addition, we performed phylogenomic and polymorphism analyses to assess evolutionary diversification of this species. Our results suggest that this genome is syntenic with the other published mtDNA of *Drosophila.* This molecule contains 15,641 bp and encompasses two rRNA, 22 tRNA and 13 protein-coding genes. Regarding nucleotide composition, we found a high A−T bias (76.6 %). The recovered phylogenies indicate *D. incompta* in the *virilis–repleta* radiation, as sister to the *virilis* or *repleta* groups. The most interesting result is the high degree of polymorphism found throughout the *D. incompta* mitogenome, revealing pronounced intrapopulational variation. Furthermore, intraspecific nucleotide diversity levels varied between different regions of the genome, thus allowing the use of different mitochondrial molecular markers for analysis of population structure of this species.

Keywords Mitochondrial DNA · Polymorphism · *Flavopilosa* group · Mitogenome

Electronic supplementary material The online version of this article (doi:10.1007/s10709-014-9799-9) contains supplementary material, which is available to authorized users.

F. C. De Ré · G. L. Wallau · L. J. Robe · E. L. S. Loreto (✉)
Programa de Pós Graduação em Biodiversidade Animal, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil
e-mail: elgion@base.ufsm.br

G. L. Wallau
Programa de Pós-Graduação em Ciências Biológicas, Universidade Federal do Pampa, São Gabriel, Rio Grande do Sul, Brazil

L. J. Robe
Programa de Pós-Graduação em Biologia de Ambientes Aquáticos Continentais, Universidade Federal do Rio Grande, Rio Grande, Rio Grande do Sul, Brazil

E. L. S. Loreto
Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil

Introduction

The genus *Drosophila* is an excellent model system to explore several issues related to evolutionary biology. Besides *Drosophila melanogaster*, one of the most traditional eukaryotic model organisms, the genus has more than two-dozen completely sequenced genomes (Clark et al. 2007; Chen et al. 2014). In addition, this genus has 1,178 described species (Bächli 2014), 181 of which have been recorded in Brazil (Gottschalk et al. 2008). Throughout their evolution, these species have acquired special adaptations enabling them to explore different resources, such as fruits, flowers, fungi and even bat guano, in various specialization levels (Clark et al. 2007). This diversification, combined with the solid knowledge of genetics and genomics, makes this genus a major model for micro- and macro-evolutionary studies.

Within the genus *Drosophila*, one of the most interesting groups of species is the *flavopilosa* group, with potential to be used as a model in the study of restricted ecology evolution, geographical versus ecological speciation, and intra/interspecific competition (Robe et al. 2013). The members of this group are ecologically specialized since

Published online: 22 November 2014

they only use flowers in the genus *Cestrum* (Solanaceae) as sites for oviposition, larval development, and feeding (Brncic 1966; Hofmann 1985). In order to exploit these resources, these species share several adaptations, i.e. 1) they are small-to medium-sized, with body lengths up to 2.0 mm (Wheeler et al. 1962) and have a light-yellow body color, similar to *Cestrum* host flowers, and 2) females have strong spines on the outer region of the ovipositor and lay their eggs in advanced stages of embryonic development in order to reduce preadult development times, given the ephemeral nature of the flowers (Brncic 1983; Ludwig et al. 2002).

Sepel et al. (2000) showed that *Cestrum* flowers exhibit non-synchronized flowering. These authors monitored the flowers and their associated *D. flavopilosa* group species in two different areas every week for 1 year, and recorded significant *Drosophila* population bottlenecks when flowers became scarce. When flowers again became abundant, large drosophilid population expansions followed. Although sites with a large number of host plants could provide a continuous source for oviposition and feeding, variable occupancy rates between species were also suggested as potential causes for seasonal fluctuations in population sizes. More recently, Robe et al. (2013) showed that the distribution of the *flavopilosa* group species also seems to be widely affected by alterations in environmental conditions within the distribution limits of their host plants. Despite low levels of environmental variability expected for species with such biotic- and abiotically restricted ecological patterns, especially for those that may be subject to frequent bottlenecks, these species populations appear to be quite polymorphic (Brncic 1962; Napp and Brncic 1978; Hofmann and Napp 1984).

The flavopilosa group is a member of the *Drosophila* subgenus (Wheeler et al. 1962) and comprises 17 described species that have been further subdivided into the *flavopilosa* and *nesiota* subgroups. One member of the former subgroup, *D. incompta*, occurs in Brazil, Argentina, Colombia, Panamá, México, and the Caribbean islands (Bächli 2014). According to molecular evidence (Robe et al. 2005, 2010), the *flavopilosa* group belongs to the *virilis–repleta* radiation of the *Drosophila* subgenus, clustering with its sister group, the *D. annulimana* group. However, no phylogenomic analyses have yet been used in order to better characterize the phylogenetic positioning of this group.

Whole-genome sequencing has allowed study of variation in patterns of DNA and RNA, whether or not related to the action of natural selection, from genes to populations (Chen et al. 2014). Mitochondrial genomes (mtDNA) comprise a set of genes necessary for operation and maintenance of aerobic cellular metabolism in eukaryotic species (Matoba et al. 2006). Additionally, mtDNA controls some fundamental functions for energy production, such as oxidative phosphorylation (OXPHOS), production of most reactive oxygen species (ROS), and regulation of cell apoptosis (Wallace 2005).

In general, the majority of animal mtDNAs are 15–20 kb in length, and are composed of 13 protein-coding genes (involved in the oxidative phosphorylation process), two rRNAs of the mitochondrial ribosome, and 22 tRNAs necessary for translation of the proteins encoded by mtDNA (reviewed in Boore 1999). Moreover, these genomes contain a control region comprised of promoters and control elements for the replication and transcription (Taanman 1999). The organization of this set of genes seems to be common throughout the mtDNA from different animal species, and is conserved within phyla (Lang et al. 1999). Nevertheless, recent studies have revealed that this uniformity may not apply if different animal mtDNAs are compared, especially for non-bilaterally symmetrical animals including Cnidaria, Ctenophora, Placozoa, and Porifera (Lavrov 2014). Mollusc mitogenomes also vary in size and nucleotide composition, and some bivalve species have an altered system of mtDNA inheritance involving two kinds of mtDNA, one transmitted by the mother to all offspring and the second passed on by father exclusively to sons (Zouros et al. 1992; Breton et al. 2007).

In the genus *Drosophila*, there are 14 annotated mtDNAs available in GenBank: *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, *D. grimshawi* (Montooth et al. 2009), *D. yakuba* (Clary and Wolstenholme, 1985), *D. melanogaster* (Garesse 1988), *D. simulans, D. mauritiana* (Ballard 2000a, b), *D. pseudoobscura* (Torres et al. 2009), *D. sechellia* (Ballard 2000a, b), and *D. littoralis* (Andrianov et al. 2010). Further, there are mitogenomic sequences available for at least 15 and 20 *D. simulans* and *D. melanogaster* individuals, respectively (Early and Clark 2013). In the present study, we add to the number of complete *Drosophila* mtDNA by sequencing, assembling and characterizing the mtDNA of *D. incompta.* In addition, we performed phylogenomic and polymorphism analyses in order to understand phylogenetic relationships and the micro-evolutionary patterns associated with the restricted ecology of the *flavopilosa* species group.

Materials and methods

Sampling and processing DNA

Flies of the group *flavopilosa* can't be grown in the laboratory. To obtain such flies, flowers of *Cestrum parquii* (Solanaceae) were collected in a park of Curitiba, Paraná state, south of Brazil (25°25′26.88″S/49°15′57.73″W), and were taken to the laboratory and maintained until

emergence of the adult flies. Following emergence, all individuals were identified by external morphology and male genitalia, according to the pictures provided by Wheeler et al. (1962) and Danko Brncic (unpublished data). These diagnostic morphological characters are consistent with molecular markers (Robe et al. 2013). Total DNA was extracted from a pool of 20 individuals of *D. incompta* using the NucleoSpin Tissue XS kit (Macherey–Nagel–Germany).

Sequencing and mitochondrial genome assembly

The entire genome was sequenced by the Fasteris DNA Sequencing Service with a Solexa-Illumina HiSeq 2000 Next Generation Sequencing (NGS) device. A single-end approach with a read-size of approximately 100 bp was employed. The sequence reads were not quality-trimmed since there was high confidence in scoring every base (a quality plot of all reads used in the mtDNA assembly is presented in Supplementary Figure 1). Once the reads were obtained, the *D. incompta* mtDNA was assembled using the MITObim package (Hahn et al. 2013). The NADH dehydrogenase subunit 5 gene (NADH 5) from *D. virilis* mtDNA (BK006340.1) was used to locate the homologous sequence in the *D. incompta* draft genome, and this sequence was then used as a seed for the mtDNA total assembly. MITObim uses an in silico baiting approach, which was implemented in the MIRAbait module of the MIRA assembler (v3.4.1.1) (Chevreux et al. 1999). The entire set of reads obtained from *D. incompta* individuals were used in this analysis, and MIRAbait extracted the mtDNA reads using kmers of 31 bp.

After the reads were obtained by the in silico baiting approach, they were mapped back to the gapped reference sequence, leading to the extension of the reference *D. incompta* mtDNA. In order to be incorporated as an extension into the mapping assembly, a read required an overlap of at least 30 bp at the edge of the reference, allowing mismatches to sum to 15 % of the total length of a Smith–Waterman alignment overlap (Smith and Waterman 1981). The *D. incompta* total mtDNA sequence was deposited in Genbank under accession number KM275233.

Characterization, annotation and polymorphisms analyses

The characterization and annotation of the assembled *D. incompta* mtDNA was performed in MITOS Web Server (Bernt et al. 2013), using default parameters and UGENE software (Okonechnikov et al. 2012), respectively. Nucleotide composition on a general *D. incompta* mitogenome scale, for the control region, and for different codon positions within each coding gene were measured using Mega 5.0 (Tamura et al. 2011). For *D. incompta* mtDNA, determination of single-nucleotide polymorphisms (SNPs) was carried out with the FreeBayes software (O'Fallon et al. 2013) implemented in the Galaxy platform (Goecks et al. 2010; Blankenberg et al. 2010; Giardine et al. 2005), using default parameters which consider as variants only polymorphisms present in at least 20 % of the reads from a single sample. The relative number of substitutions per gene was calculated by dividing the number of substitutions by the gene length.

In order to compare the observed polymorphisms in *D. incompta* mitochondrial genes with those found in other species, 19 mtDNAs of *D. melanogaster* and 14 of *D. simulans* (geographical origin and accession numbers can be found in Supp. Table 1S) were downloaded from GenBank and aligned using Clustal W, as implemented in Mega 5.0 (Tamura et al. 2011). The inter-populational nucleotide diversity (p_i) was calculated for each gene with DNAsp software (Librado and Rozas 2009), using the DNA polymorphism option. These estimates were obtained as the average of the values for all comparisons per gene and sites with alignment gaps or missing data were not used.

As the presence of nuclear mitochondrial pseudogenes (Numts) can lead to overestimates of mitochondrial polymorphism, their potential contribution to *D. incompta* diversity values was evaluated analyzing the presence of frameshifts, and the number of synonymous (S) and non-synonymous (NS) substitutions in each gene. In addition, mtDNA was extracted from individual *D. incompta* flies using the protocol described by Pissios and Scouras (1992). Fragments of the mitochondrial Cytochrome Oxidase Subunit I (COI) and Cytochrome Oxidase Subunit II (COII) genes were amplified using the following primers pairs, respectively: TYJ1460 (5′-TAC AAT CTA TCG CCT AAA CTT CAG CC-3′) and C1N2329 (5′-ACT GTA AAT ATA TGA TGA GCT CAT ACA-3′) (from Simon et al. 1994); TL2J3037 (5′-ATG GCA GAT TAG TGC AAT GG-3′) and TKN3785 (5′-GTT TAA GAG ACC AGT ACT TG-3′) (Simon et al. 1994). At least two electropherograms from each of 20 individuals were assembled in Gap4 using the Staden Package (Staden 1996) and carefully checked by looking for the presence of multiple peaks for each base.

Phylogenetic analyses

Phylogenetic analyses were conducted using two different matrices. The first comprised newly obtained *D. incompta* mtDNA plus the 14 complete mitochondrial sequences of *Drosophila* available in Genbank: *D. melanogaster* (NC_001709.1), *D. simulans* (JQ_691661.1), *D. sechellia* (NC_005780.1), *D. yakuba* (KF_824901.1), *D. erecta*

(BK_006335.1), *D. ananassae* (BK_006336.1), *D. pseudoobscura* (NC_018348.1), *D. persimilis* (BK_006337.1), *D. willistoni* (BK_006338.1), *D. mojavensis* (BK_006339.1), *D. virilis* (BK_006340.1), *D. grimshawii* (BK_006341.1), *D. mauritiana* (NC_005779.1), and *D. littoralis* (NC_011596.1). In this case, a partial mtDNA sequence from *Musca domestica* (EU_154477.1) was used as an out-group. The second data matrix encompassed a concatenation of three mitochondrial genes (COI, COII and ND2), which were joined in order to improve the positioning of *D. incompta* within the *virilis–repleta* radiation. In this case, besides *D. incompta*, 14 other species included in the *virilis–repleta* radiation were considered according to Robe et al. (2010) (Table 2S).

The alignments were made using ClustalW, as implemented in Mega 5.0 (Tamura et al. 2011). For the first analysis, the whole genomes were aligned and indels were not observed in protein-coding genes, although some RNA genes (tRNAY, tRNAK, tRNAL1, rRNAS and rRNAL) contained indels involving one or two nucleotides. For the second matrix, alignments were individually conducted for each gene, after which concatenation was carried out in SeaView (Galtier et al. 1996). In both cases, Bayesian analyses (BA) were performed in MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001), using the GTR+I+G nucleotide substitution model, as selected by the Akaike information criterion (AIC) implemented in the JModelTest (Darriba et al. 2012). The MCMCs were run for 1,000,000 generations, sampling every 100 generations. Next, 25 % of the chains were discarded and those remaining were used to estimate a majority-rule consensus tree and the posterior probability of each clade.

Results

mtDNA genome reconstitution

MITObim took 111 iterations to reach a stationary state where no more reads were added. The complete mtDNA of *Drosophila incompta* is a circular molecule of 15,641 bp (Fig. 1a). A total of 436,275 reads were assembled reaching an average total coverage of 2,806.38 X with a maximum coverage of 5,162.

mtDNA characterization and annotation

Genome annotation led to the characterization of the gene content of the *D. incompta* mtDNA, which encompasses 13 protein-coding genes [NADH dehydrogenase subunit 1–6 and 4L (ND1–6, ND4L), Cytochrome oxidase subunits I, II, III (COI, COII, COIII), ATP synthase subunit 6 and 8 (ATP6, ATP8), Cytochrome b (Cytb)], 22 tRNA genes [tRNAI (anticodon gat), tRNAQ (ttg), tRNAM (cat), tRNAW (tca), tRNAC (gca), tRNAY (gta), tRNAL2 (taa), tRNAK (ctt), tRNAD (gtc), tRNAG (tcc), tRNAA (tgc), tRNAR (tcg), tRNAN (gtt), tRNAS1 (gct), tRNAE (ttc), tRNAF (gaa), tRNAH (gtg), tRNAT (tgt), tRNAP (tgg), tRNAS2 (tga), tRNAL1 (tag), and tRNAV (tac)], and two rRNA genes (rRNAL, rRNAS) (Table 1, Fig. 1a). The mtDNA of *D. incompta* is AT rich, totaling 76.68 % A−T content. The plus strand comprised nine protein coding genes and 14 RNA genes, whereas the minus strand encompassed four protein coding genes and 10 RNA genes (Table 1; Fig. 1a). Besides total gene content, these strands were also differentiated as regards the A+T bias of the whole gene regions, and the A−T content varied from 74 to 84.7 % for the plus and minus strand, respectively.

The control region (in insects, also called the AT-rich region) of this genome was found to be located between rRNAS and the tRNAI, encompassing 733 bp, with a A−T content of nearly 92 % (Table 3S). For protein-coding genes (PCGs), A−T bias was always higher for third codon positions, and often lower for second codon positions, except for COI, COII, COIII and CytB (Table 2). The most and the least biased genes were ND6 and COI, respectively. Furthermore, there was a pattern of nucleotide overlap for five pairs of genes: ND4L–ND4 (6 nt), ATP8–ATP6 (9 nt), tRNAW–tRNAC (7 nt), tRNAG–ND3 (2 nt) and tRNAL1–rRNAL (39 nt).

Phylogenomic and polymorphism analyses

The Bayesian phylogeny reconstructed based on the 15 complete *Drosophila* mitogenomes (Fig. 2a) indicated *D. incompta* as a member of *Drosophila* subgenus, nested inside the *virilis–repleta* radiation. In this case, this species appeared as closely related to a clade formed by two species of the *virilis* group, *D. virilis* and *D. littoralis*, with a posterior probability (PP) of 1.00. However, when the sampling of the *virilis–repleta* radiation species was expanded for only three of the mtDNAs PCGs (COI, COII and ND2), *D. incompta* appeared as sister to the *Drosophila repleta* species group (PP = 0.97) (Fig. 2b).

We also observed abundant polymorphisms throughout the *D. incompta* mtDNA genome, and pronounced intrapopulational variation. The gene that showed the highest number of polymorphic sites was tRNAH, with a relative frequency of 4.6 %, followed by PCGs ATP8 and ND5 (Fig. 3). In contrast, tRNAI, tRNAQ, tRNAG, tRNAA, tRNAF and tRNAP showed an absence of polymorphic sites. Cytb and COIII revealed the lowest number of polymorphic sites among the protein-coding genes. When the polymorphic sites found within the PCGs were

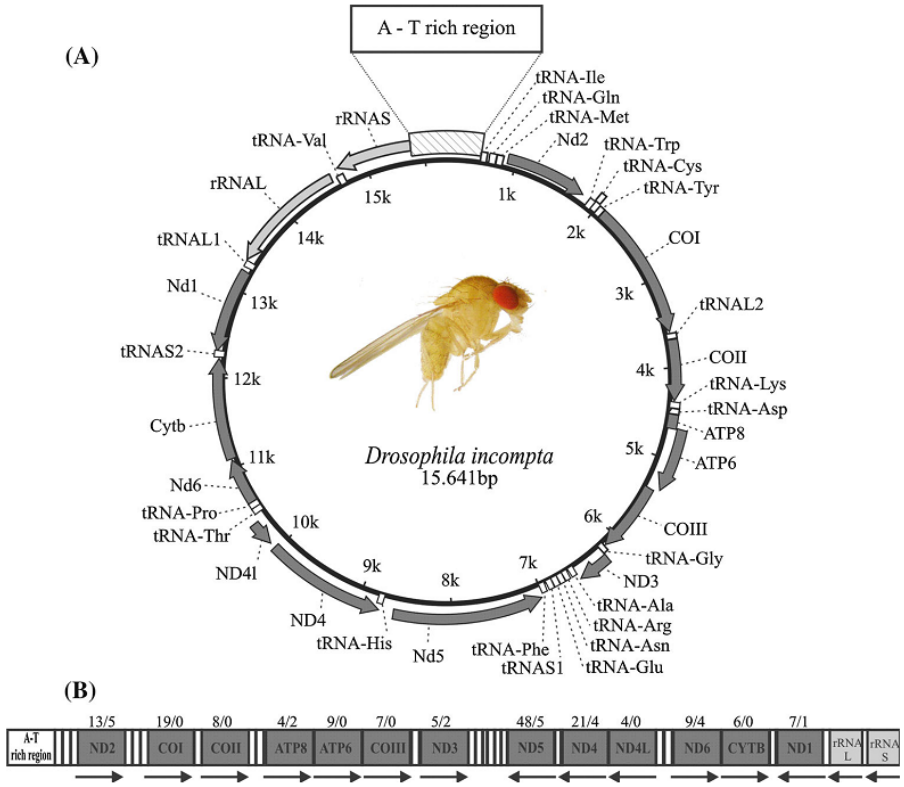


Fig. 1 Summary of *D. incompta* mitochondrial genome content and organization (**a**). ND1–6 and 4L refer to NADH dehydrogenase subunits 1–6 and 4L, COI–III refers to cytochrome c oxidase subunits 1–3, ATP6 and ATP8 refer to ATPase subunits 6 and 8, and Cyt b refers to cytochrome b. Protein-coding genes and rRNA genes are represented by dark and light gray, respectively, whereas tRNA genes are represented by white squares and the control region is crosshatched. (**b**) Number of synonymous/non-synonymous polymorphic substitutions along each mitochondrial protein coding gene. *Arrows* indicate gene direction

characterized for the presence of indels and synonymous/non-synonymous substitutions, only synonymous substitutions were found for COI, COII, COIII, ND4L, ATP6 and CYTB genes (Fig. 1b). The other seven PCGs showed one or few non-synonymous mutations (maximum of five), but the NS/S ratio was always lower than 0.5. No PCGs presented indels, and when electropherograms obtained for single individuals for partial COI and COII fragments were carefully examined, no case of double or multiple peaks was detected. Comparing levels of polymorphism observed in mitochondrial genes of *D. incompta*, revealed similar estimates with those of populations of *D. melanogaster* and *D. simulans*. As can be seen in Fig. 3, ATP8, ND5 and ND6 showed similar diversity values among *D. incompta* and *D. simulans*, although the first was examined at the intra-populational level whereas 12 different populations were included for the second. The tRNAN and tRNAH genes had higher polymorphism levels for *D. incompta*, while the other mitochondrial genes had higher polymorphism values in *D. simulans*. Overall inter-populational mitochondrial diversity observed for *D. melanogaster* was significantly reduced when compared with the inter-populational and intra-populational polymorphisms of *D. simulans* and *D. incompta*, respectively.

Discussion

D. incompta mtDNA is 15,641 bp in length, which is similar to other *Drosophila* mtDNA, from 14,874 bp in *D. grimshawii* (Montooth et al. 2009) to 19,517 bp in *D. melanogaster* (Garesse 1988), and also in other insects e.g., *Apis mellifera* (16,343 bp) (Crozier and Crozier 1993) and *Cervaphis quercus* (15,272 bp) (Wang et al. 2014), crustaceans, *Artemia franciscana* (15,822 bp) (Valverde et al. 1994), and echinoderms, *Paracentrotus lividus* (15,697 bp) (Cantatore et al. 1989). This variation in size in different *Drosophila* genomes is mainly due to insertions and deletions in the control region. Genome content and organization are conserved across the 15 *Drosophila* mtDNA genomes, as also seen by Montooth et al. (2009).

The AT-rich region consists of 733 bp in *D. incompta*, which is smaller than that reported for other Drosophilidae species, with a maximum of 1,149 bp for *D. obscura* (Saito et al. 2005) and a minimum of 842 bp for *Zaprionus indianus* (da Silva et al. 2009) (Table 3S). The A+T bias of this region of the *D. incompta* mitogenome (92 %) is intermediate between the extremes found for other Drosophilidae species, ranging from 88.8 % in *D. albomicans* (Saito et al. 2005) to 95.5 % in *D. mauritiana* (Tsujino et al. 2002) (Table 3S).

Table 1 Mitochondrial genome content of *D. incompta*

Gene	Position			Strand
	Start	Stop	Length (bp)	
tRNAI (gat)	607	670	64	+
tRNAQ (ttg)	701	769	69	–
tRNAM (cat)	780	848	69	+
ND2	870	1,775	906	+
tRNAW (tca)	1,878	1,943	66	+
tRNAC (gca)	1,936	1,997	62	–
tRNAY (gta)	2,012	2,077	66	–
COI	2,082	3,590	1,509	+
tRNAL2 (taa)	3,614	3,679	66	+
COII	3,684	4,355	672	+
tRNAK (ctt)	4,372	4,441	70	+
tRNAD (gtc)	4,442	4,507	66	+
ATP8	4,508	4,666	159	+
ATP6	4,657	5,328	672	+
COIII	5,362	6,144	783	+
tRNAG (tcc)	6,164	6,227	64	+
ND3	6,225	6,575	351	+
tRNAA (tgc)	6,585	6,649	65	+
tRNAR (tcg)	6,665	6,727	63	+
tRNAN (gtt)	6,728	6,793	66	+
tRNAS1 (gct)	6,794	6,861	68	+
tRNAE (ttc)	6,862	6,929	68	+
tRNAF (gaa)	6,948	7,013	66	–
ND5	7,036	8,694	1,659	–
tRNAH (gtg)	8,749	8,813	65	–
ND4	8,818	10,155	1338	–
ND4 1	10,149	10,412	264	–
tRNAT (tgt)	10,445	10,509	65	+
tRNAP (tgg)	10,510	10,574	65	–
ND6	10,577	11,089	513	+
Cytb	11,101	12,225	1,125	+
tRNAS2 (tga)	12,236	12,302	67	+
ND1	12,334	13,266	933	–
tRNAL1 (tag)	13,268	13,332	65	–
rRNAL	13,293	14,659	1,367	–
tRNAV (tac)	14,658	14,729	72	–
rRNAS	14,729	15,513	785	–

Table 2 Nucleotide composition at each codon position in the thirteen protein-coding genes (PCGs) encompassing *D. incompta* mitogenome

PCGs/strand	First codon position		Second codon position		Third codon position	
	A–T (%)	C–G (%)	A–T (%)	C–G (%)	A–T (%)	C–G (%)
ATP6/+	67.8	32.1	62	37.9	92.6	7.6
ATP8/+	82.8	17	68.4	32	96.3	3.8
Cytb/+	61.6	38.7	65.6	34.7	91.3	8.3
COI/+	56.6	42.9	59.7	40.4	90.3	9.4
COII/+	59.9	39.8	68.3	32.2	87.6	12.1
COIII/+	59.1	41	59.6	40.6	90.5	9.2
ND1/–	71.1	28.9	67.6	32.4	89.8	10.2
ND2/+	77.8	22.5	68.9	31.5	92	7.7
ND3/+	75.6	24.8	69.4	30.8	93.7	6
ND4/–	74.4	25.6	67.6	32.4	95.1	4.9
ND4L/–	76.5	23.5	76.5	23.5	96.8	3.2
ND5/–	72.5	27.5	67.8	32.2	94.5	5.5
ND6/+	81.1	18.7	74.9	25.2	96	4.1

Although Clary and Wolstenholme (1985) and Garesse (1988) did not report any overlap pattern for the ND4L–ND4 gene pair in Drosophila, *D. incompta* mtDNA shows an overlap of 6 nt between this pair of genes. A similar pattern was also reported for other insects (Wang et al. 2014). The pair of genes ATP6–ATP8, in turn, showed a 9-nt overlap in the mtDNA of *D. incompta*, which is similar to the 7-nt overlap found in *D. yakuba* (Clary and Wolstenholme, 1985). The tRNAW–tRNAC gene pair also showed a 7-nt overlap both in *D. yakuba* (Clary and Wolstenholme, 1985) and *D. incompta*, although Wang et al. (2014) showed that the mtDNA of *Cervaphis querus* has a 20-nt overlap for this gene pair, while *Apis mellifera* has a 19-nt overlap (Crozier and Crozier 1993).

In general, the phylogenetic relationships recovered through the Bayesian analysis of complete mtDNA is in accordance with the phylogeny presented by Clark et al. (2007). However, here we have included additional species, including *D. mauritiana*, *D. littoralis,* and *D. incompta*. *Drosophila mauritiana* appeared as a sister to the *D. simulans* and *D. sechellia* clade (Fig. 2a), concordant with Hatadani et al. (2009), but discordant with the phylogeny presented by Young and Coleman (2004) in which *D. mauritiana* clustered with *D. sechellia*, and with Da Lage et al. (2007) and Kastanis et al. (2003) in which the new species clustered with *D. simulans*.

Concerning the phylogenetic positioning of *D. incompta*, our mitogenomic analysis adds to current knowledge by supporting the *flavopilosa* group as a member of the *Drosophila* subgenus, clustered inside the *virilis–repleta* radiation (Robe et al. 2010). These results are in contrast with those presented by Grimaldi (1990), who clustered these species with the genus *Scaptomyza*. Within the *virilis–repleta* radiation, mitogenomic data supported *D. incompta* as closely related to the clade formed by *D. virilis* and *D. littoralis*, two species from the *virilis* group (Fig. 2a), although COI + COII + ND2 concatenated data presented the target species as sister to the *repleta* group (Fig. 2b). Previous molecular evidence based on nuclear

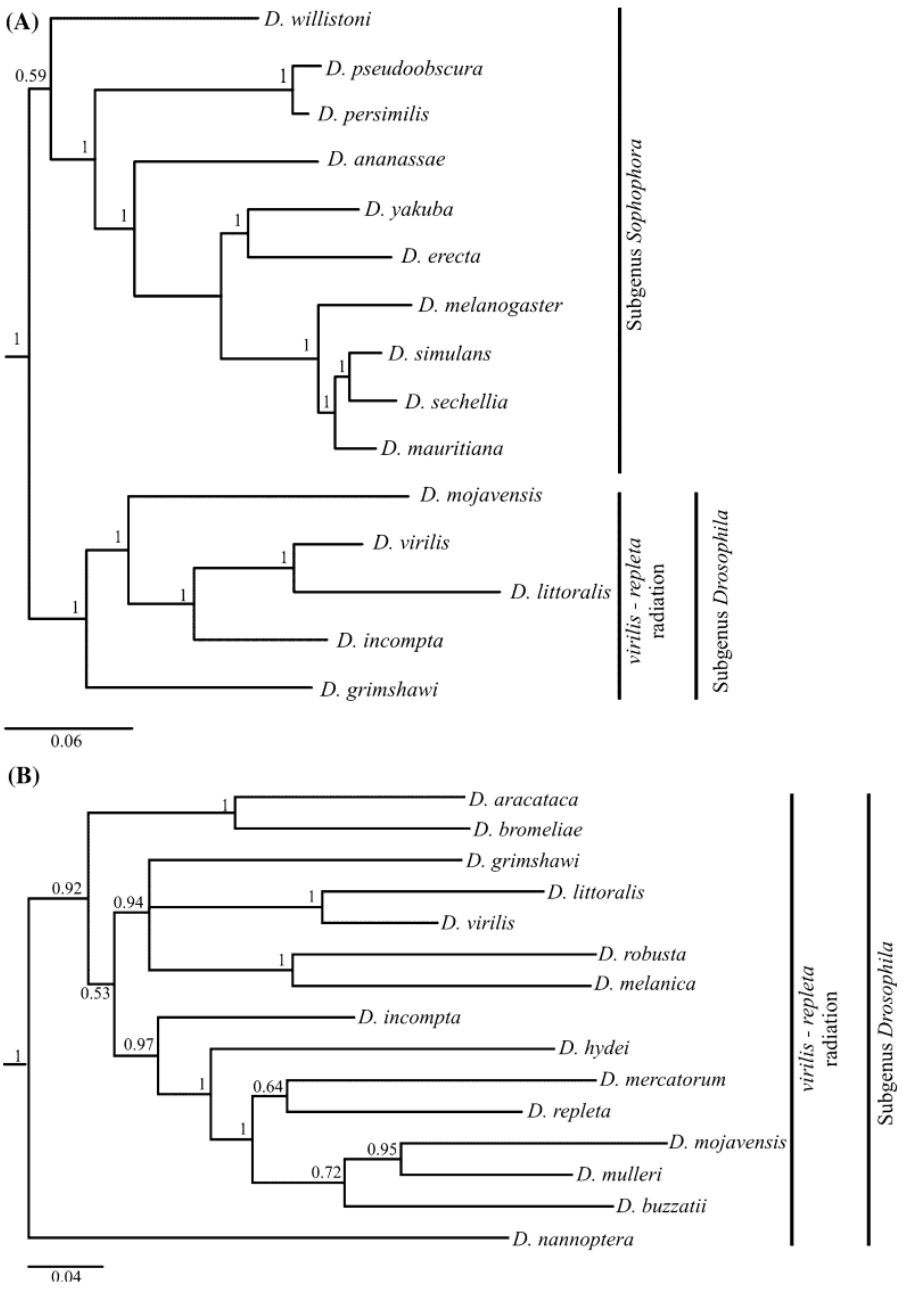


Fig. 2 Majority-rule consensus phylogenetic tree obtained using a Bayesian analysis with the GTR + I + G model, as proposed by the AIC test for the 15 complete *Drosophila* subgenus mitochondrial genomes (**a**) and for the *virilis–repleta* radiation concatenate COI + COII + ND2 genes (**b**). The outgroups (*Musca domestica* and *D. immigrans*, respectively) were omitted from the tree. The posterior probability of each clade is indicated before its respective internal branch

genes supported *D. incompta* as sister to the *annulimana* group (Robe et al. 2005, 2010), which was only included in the partial concatenated analysis. So, the phylogenetic positioning of the *D. flavopilosa* species group is still controversial, and uniform species sampling added to simultaneous inclusion of multiple nuclear and mitochondrial markers are required to better understand this complex evolutionary pattern.

As the flies of the *flavopilosa* group do not grow in the laboratory, we sampled individuals from a single natural population. Judging by the number of individuals sampled for the mtDNA sequencing (20), the polymorphism we detected (up to 4.6 %) in some genes can be considered high for mtDNA. As our data should not be directly compared to those of other *Drosophila* mtDNA because they were assembled from inbred isofemale strains

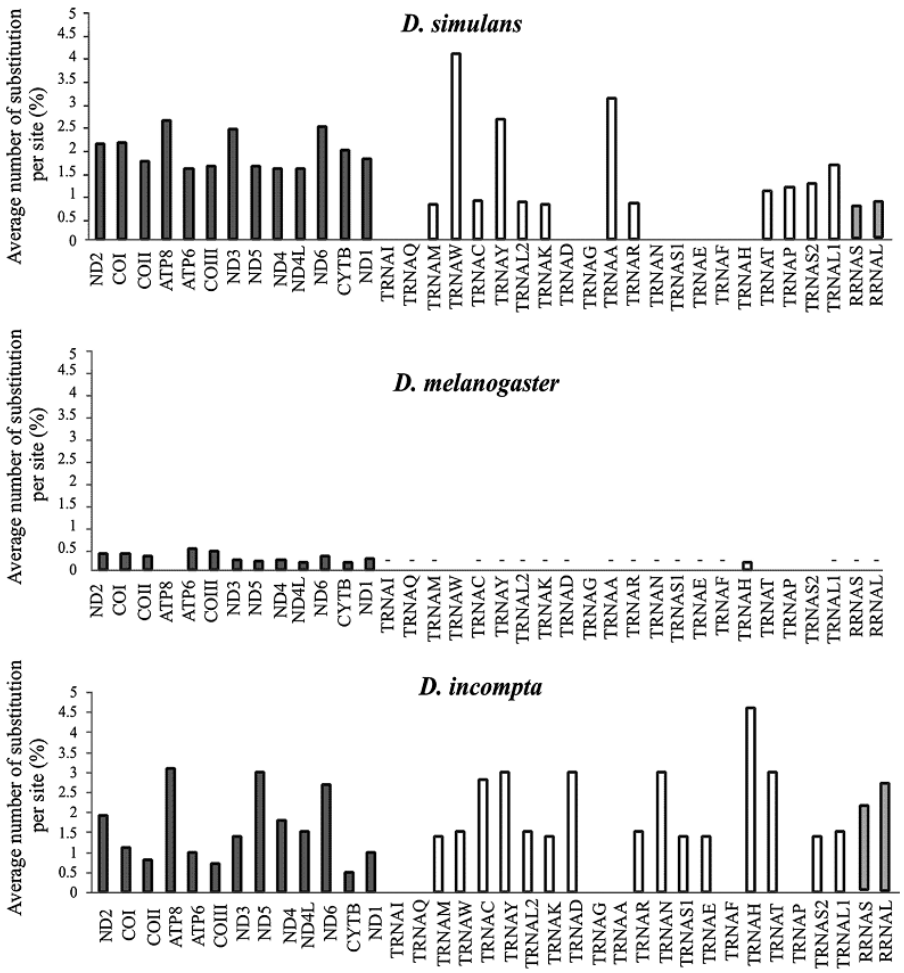


Fig. 3 Intraspecific nucleotide diversity throughout the 37 genes that encompass the mitochondrial genome of *D. incompta*, *D. simulans* and *D. melanogaster*. For *D. melanogaster* sequences, some populations showed absence or incomplete sequences for tRNAs. Due to this lack of information, these genes were excluded from the analysis and were represented by "*asterisk*"

(minimizing the number of polymorphic sites), we have performed analogous analyses with the publicly available mitogenomes of members of different populations of *D. simulans* and *D. melanogaster*. It is remarkable that the intrapopulational polymorphism levels observed in *D. incompta* were almost similar to those observed worldwide in *D. simulans* and much higher than those observed in *D. melanogaster*. The low levels of mtDNA nucleotide diversity observed in *D. melanogaster* may be related to the recent invasion by *Wolbachia* endosymbiont, dated to 8,000 years ago (Richardson et al. 2012).

Nuclear mitochondrial pseudogenes (Numts) appear to evolve neutrally and with specific nuclear mutation bias. For these reasons, they can potentially inflate the polymorphism estimates for mitochondrial genes (Bensasson et al. 2001; Rogers and Griffiths-Jones 2012). We cannot reject the hypothesis that the presence of Numts in *D. incompta* genome has affected our results. However, several evidences suggest that observed polymorphism levels were not related to the presence of Numts: (1) absence of frameshifts and in-frame nonsense mutations in the protein-coding genes; (2) polymorphic mutations were mainly synonymous (87.4 %); (3) at least for COI and COII genes, samples from several individuals, sequenced using Sanger methods, showed single peaks along the entire sequence (if Numts were present double peaks would be expected in some bases). Furthermore, it is important to emphasize that *D. incompta* polymorphism levels were high throughout the entire mtDNA. As Numts generally encompass a small region of the mitogenome (Richly and Leister 2004), congruence between different markers (Song et al. 2008; Linares et al. 2009) can also be taken as a good indicator that the high polymorphism levels detected here were not just an artifact due to the co-sequencing of nonfunctional mtDNA copies. Thus, we propose the high polymorphism levels encountered in the mtDNA of *D. incompta* may be a consequence of putative high migration rates associated with the cyclical

population expansions and retractions seasonally faced by this species (Sepel et al. 2000).

For other *Drosophila* species, more nucleotide diversity data are available mainly for the mitochondrial genes COI and COII. In general, nucleotide diversity for these genes ranges from zero to 1.0 % (Table 4S). The only nucleotide diversity values similar to those observed in *D. incompta* (1.12 % for COI) were observed by Franco and Manfrin (2013) for *D. gouveai* (1.07 %) and *D. seriema* (1.11 %), although in these cases different populations were sampled in eastern Brazil and not only from a single population as is the case here for *D. incompta*. High levels of polymorphisms were already reported for the *flavopilosa* group species through the use of different genetic markers, including allozymes (Hofmann and Napp 1984). We argue here that these studies support the hypothesis that *D. incompta* populations persist through cyclic and recurrent bottlenecks, with the possible aid of new migrants for population re-establishment. Although particular adaptations involved in the evolution of the *D. incompta* need to be assessed in the nuclear genome, the description and the characterization of the mtDNA presented here has revealed micro- and macro-evolutionary patterns related to the origin, establishment, and persistence of this species. Moreover, the description of *D. incompta* mtDNA polymorphisms may guide the choice of gene markers for phylogeographic studies, which may additionally clarify the processes associated with the evolution of this interesting species.

Acknowledgments This study was supported by research grants and fellowships from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), PRONEX-FAPERGS (10/0028-7) and FAPERGS (11/0938-0).

References

Andrianov B, Goryacheva I, Mugue N, Sorokina S, Gorelova T, Mitrofanov V (2010) Comparative analysis of the mitochondrial genomes in *Drosophila virilis* species group (Diptera: Drosophilidae). Trends Evol Biol 2:e4

Bächli G (2014) TaxoDros: the database on taxonomy of Drosophilidae, v. 1.03, Database 2010/2012. http://taxodros.unizh.ch/. Last accessed on 13 June 2014

Ballard JWO (2000a) Comparative genomics of mitochondrial DNA in *Drosophila simulans*. J Mol Evol 51:64–75

Ballard JWO (2000b) Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. J Mol Evol 51:48–63

Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol 16:314–321

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Stadler PF (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol 69:313–319

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol Chapter 19:Unit 19.10.1-21

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27(8):1767–1780

Brehm A, Harris DJ, Hernández M, Cabrera VM, Larruga JM, Pinto FM, González AM (2001) Structure and evolution of the mitochondrial DNA complete control region in the *Drosophila subobscura* subgroup. Insect Mol Biol 10:573–578

Breton S, Beaupré HD, Stewart DT, Hoeh WR, Blier PU (2007) The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? Trends Genet 23:465–474

Brncic D (1962) Chromosomal structure of populations of *Drosophila flavopilosa* studied in larvae collected in their natural breeding sites. Chromosoma 13:183–195. doi:10.1007/BF00326570

Brncic D (1966) Ecological and cytogenetic studies of *Drosophila flavopilosa*, a *Neotropical* species living in *Cestrum* flowers. Evolution 20:16–29

Brncic D (1983) The *flavopilosa* group of species as an example of flower-breeding species. In: Ashburner M, Carson HL, Thompson JN (eds) The genetics and biology of *Drosophila*, vol 6d. Academic Press, New York, pp 360–377

Cantatore P, Roberti M, Rainaldi G, Gadaleta MN, Saconne C (1989) The complete nucleotide sequence, gene order and genetic code of the mitochondrial genome of *Paracentrotus lividus*. J Biol Chem 264:10965–10975

Chen ZX, Sturgill D, Qu J et al (2014) Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. Genome Res 24:1209–1223

Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information. In: German conference on bioinformatics, pp. 45–56

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman T, Kellis M, Gelbart W et al (2007) Evolution of genes and genomes on the *Drosophila phylogeny*. Nature 450:203–218

Clary DO, Wolstenholme DR (1985) The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. J Mol Evol 22:252–271

Clary DO, Wolstenholme DR (1987) Drosophila mitochondrial DNA: conserved sequences in the A + T-rich region and supporting evidence for a secondary structure model of the small ribosomal RNA. J Mol Evol 25:116–125

Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. Genetics 133:97–117

da Silva NM, de Souza Dias A, da Silva Valente VL, Valiati VH (2009) Characterization of mitochondrial control region, two intergenic spacers and tRNAs of *Zaprionus indianus* (Diptera: Drosophilidae). Genetica 137:325–332

DaLage JL, Dergoat GJ, Maczkowiak F, Silvain JF, Cariou ML, Lachaise D (2007) A phylogeny of Drosophilidae using the *Amyrel* gene: questioning the *Drosophila melanogaster* species group boundaries. J Zool Syst Evol Res 45:47–63

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772

De Brito RA, Manfrin MH, Sene FM (2002) Mitochondrial DNA phylogeography of Brazilian populations of *Drosophila buzzatii*. Genet Mol Biol 25:161–171

Early AM, Clark AG (2013) Monophyly of *Wolbachia pipientis* genomes within *Drosophila melanogaster*: geographic

structuring, titre variation and host effects across five populations. Mol Ecol 22:5765–5778

Franco FF, Manfrin MH (2013) Recent demographic history of cactophilic *Drosophila* species can be related to Quaternary palaeoclimatic changes in South America. J Biogeogr 40:142–154. doi:10.1111/j.1365-2699.2012.02777.x

Galtier N, Gouy M, Gautier C (1996) SeaView and Phylo win two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Garesse R (1988) *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. Genetics 118:649–663

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451–1455

Goecks J, Nekrutenko A, Taylor J, The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86. doi:10.1186/gb-2010-11-8-r86

Gottschalk MS, Hofmann PRP, Valente VLS (2008) Diptera, Drosophilidae: historical occurrence in Brazil. Check List 4:85–518

Grimaldi DA (1990) A phylogenetic, revised classification of the genera in the Drosophilidae (Diptera). Bulletin of the American Museum of Natural History 197:1–139

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res 41:E129. doi:10.1093/nar/gkt371

Hatadani LM, McInerney JO, Medeiros HFD, Junqueira ACM, Azeredo Espin AMD, Klaczko LB (2009) Molecular phylogeny of the *Drosophila tripunctata* and closely related species groups (Diptera: Drosophilidae). Mol Phylogenet Evol 51:595–600

Hofmann PRP (1985) Variabilidade genética em espécies de nível ecológico restrito. Ciência e Cultura 37:579–581

Hofmann PRP, Napp M (1984) Genetic-environmental relationships in *Drosophila incompta*, a species of restricted ecology. Braz J Genet 7:21–39

Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Hurtado LA, Erez T, Castrezana S, Markow TA (2004) Contrasting population genetic patterns and evolutionary histories among sympatric Sonoran Desert cactophilic *Drosophila*. Mol Ecol 13:1365–1375

Kastanis P, Eliopoulos E, Goulielmos GN, Tsakas S, Loukas M (2003) Macroevolutionary relationships of species of *Drosophila melanogaster* group based on mtDNA sequences. Mol Phylogenet Evol 28:518–528

Lang BF, Gray MW, Burger G (1999) Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet 33:351–397

Lavrov DV (2014) Mitochondrial genomes in invertebrate animals. In Molecular life sciences. Springer, New York, pp 1–8

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452

Linares MC, Soto-Calderon ID, Lees DC, Anthony NM (2009) High mitochondrial diversity in geographically widespread butterflies of Madagascar: a test of the DNA barcoding approach. Mol Phylogenet Evol 50:485–495

Ludwig A, Vidal NM, Loreto ELS, Sepel LMN (2002) *Drosophila incompta* development without flowers. Drosoph Inf Serv 85:40–41

Matoba S, Kang JG, Patino WD et al (2006) p53 regulates mitochondrial respiration. Science 312:1650–1653

Mirol PM, Routtu J, Hoikkala A, Butlin RK (2008) Signals of demographic expansion in *Drosophila virilis*. BMC Evol Biol 8:59

Monnerot M, Solignac M, Wolstenholme DR (1990) Discrepancy in divergence of the mitochondrial and nuclear genomes of *Drosophila teissieri* and *Drosophila yakuba*. J Mol Evol 30:500–508

Montooth KL, Abt DN, Hofmann JW, Rand DM (2009) Comparative genomics of Drosophila mtDNA: novel features of conservation and change across functional domains and lineages. J Mol Evol 69(1):94–114

Moraes EM, Yotoko KSC, Manfrin MH, Solferini VN, Sene FM (2009) Phylogeography of the cactophilic species *Drosophila gouveai*: demographic events and divergence timing in dry vegetation enclaves in eastern Brazil. J Biogeogr 36:2136–2147. doi:10.1111/j.1365-2699.2009.02145.x

Napp M, Brncic D (1978) Eletrophoretic variability in two closely related Brazilian species of the flavopilosa species group of *Drosophila*. Braz J Genet 1:1–10

O'Fallon BD, Wooderchak-Donahue W, Crockett DK (2013) A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. Bioinformatics 29:1361–1366. doi:10.1093/bioinformatics/btt172

Okonechnikov K, Olga G, Mikhail F (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166–1167

Pissios P, Scouras ZG (1992) Rapid isolation of mitochondrial DNA: mitochondrial DNA from *Drosophila serrata*. Experientia 48:67l–673

Reed LK, Nyboer M, Markow TA (2007) Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. Mol Ecol 16:1007–1022. doi:10.1111/j.1365-294X.2006.02941.x

Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, Bergman CM (2012) Population genomics of the Wolbachia endosymbiont in *Drosophila melanogaster*. PLoS Genet 8:e1003129

Richly E, Leister D (2004) NUMTs in sequenced eukaryotic genomes. Mol Biol Evol 21:1081–1084

Robe LJ, Valente VLS, Budnik M, Loreto ELS (2005) Molecular phylogeny of the subgenus *Drosophila* (Diptera, Drosophilidae) with an emphasis on *Neotropical* species and groups: a nuclear versus mitochondrial gene approach. Mol Phylogenet Evol 36:623–640. PMID:15970444. doi:10.1016/j.ympev

Robe LJ, Loreto ELS, Valente VLS (2010) Radiation of the *Drosophila* subgenus (Drosophilidae, Diptera) in the *Neotropics*. J Zool Syst Evol Res 48:310–21. doi:10.1111/j.1439-0469.2009.00563.x

Robe LJ, De Ré FC, Ludwig A, Loreto ELS (2013) The *Drosophila flavopilosa* species group (Diptera, Drosophilidae): an Array of exciting questions. Fly 7:59–69

Rogers H, Griffiths-Jones S (2012) Mitochondrial pseudogenes in the nuclear genomes of *Drosophila*. PLoS One:e32593

Saito S, Tamura K, Aotsuka T (2005) Replication origin of mitochondrial DNA in insects. Genetics 171:1695–1705

Sepel LMN, Golombieski RM, Napp M, Loreto ELS (2000) Seasonal fluctuations of *D. cestri* and *D. incompta*, two species of the flavopilosa group. Drosoph Inf Serv 83:122–126

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P (1994) Evolution, weighting and phylogenetic utility of mitochondrial genes sequences and a compilation of conserved polymerase chain reaction primers. Ann Entomol Soc Am 87:651–701

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. doi:10.1016/0022-2836(81)90087-5

Song H, Buhay JE, Whiting MF, Crandall KA (2008) Many species in one: DNA barcoding over estimates the number of species when

nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Aci USA 105:13486–13491

Staden R (1996) The Staden sequence analysis package. Mol Biotechnol 5:233–241

Taanman JW (1999) The mitochondrial genome: structure, transcription, translation and replication. Biochim Biophys Acta 1410:103–123

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739. doi:10.1093/molbev/msr121

Torres TT, Dolezal M, Schlötterer C, Ottenwälder B (2009) Expression profiling of *Drosophila* mitochondrial genes via deep mRNA sequencing. Nucleic Acids Res 37:7509–7518. doi:10.1093/nar/gkp856

Tsujino F, Kosemura A, Inohira K, Hara T, Otsuka YF, Obara MK, Matsuura ET (2002) Evolution of the A+T-rich region of mitochondrial DNA in the melanogaster species subgroup of *Drosophila*. J Mol Evol 55:573–583

Valverde JR, Batuecas B, Moratilla C, Marco R, Garesse R (1994) The complete mitochondrial DNA sequence of the crustacean *Artemia franciscana*. J Mol Evol 39:400–408

Wallace Douglas C (2005) A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annu Rev Genet 39:359

Wang Y, Huang XL, Qiao GX (2014) The complete mitochondrial genome of *Cervaphis quercus* (Insecta: Hemiptera: Aphididae: Greenideinae). Insect Sci 21:278–290. doi:10.1111/1744-7917.12112

Wheeler MR, Takada H, Brncic D (1962) The *flavopilosa* species group of *Drosophila*. Studies in genetic II. Univ Texas Publ 6 205:396–412

Young I, Coleman AW (2004) The advantages of the ITS2 region of the nuclear rDNA cistron for analysis of phylogenetic relationships of insects: a *Drosophila* example. Mol Phylogenet Evol 30:236–242

Zouros E, Freeman KR, Ball AO, Pogson GH (1992) Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel *Mytilus*. Nature 359:412–414

3 ARTIGO 2 - INFERRING THE PHYLOGENETIC POSITION OF THE *Drosophila flavopilosa* GROUP: INCONGRUENCE BETWEEN AND WITHIN MITOCHONDRIAL AND NUCLEAR PHYLOGENOMIC DATASETS

Inferring the phylogenetic position of the *Drosophila flavopilosa* group: incongruence between and within mitochondrial and nuclear phylogenomic datasets

De Ré, F.C.,[1] Robe, L.J.,[1,2] Wallau, G.L.,[1,3] Loreto, E.L.S.[1,4*]

1. Programa de Pós Graduação em Biodiversidade Animal, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil.
2. Programa de Pós-Graduação em Biologia de Ambientes Aquáticos Continentais Universidade Federal do Rio Grande, Rio Grande, Rio Grande do Sul, Brazil.
3. Departamento de Entomologia, Centro de Pesquisas Aggeu Magalhães - FIOCRUZ-CPqAM, Recife, PE, Brazil.
4. Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil.

*Corresponding author: Elgion L. S. Loreto (elgion@base.ufsm.br)

Keywords: Bayesian Concordance Analysis; *Drosophila incompta*; gene tree versus species tree; saturation; phylogenetic signal.

Abstract:

Inconsistencies and incongruences in phylogenetics may result from inaccurate reconstructions, or from real differences between gene and species trees. The *flavopilosa* group of species is assumed to be part of the *virilis-repleta* radiation of the *Drosophila* subgenus but its correct positioning within this clade is still controversial. In this sense, concordance between and within mitochondrial and nuclear loci were inferred, and this held a more confident description of the phylogenetic position of *D. incompta*. This approach also allowed addressing the possible causes of the extensive incongruence among phylogenies in regard to this relationship. For this, we did Bayesian analysis and Bayesian concordance analysis (BCA) using different datasets: two mitogenomic datasets, evaluated in the broader (18 Drosophilidae species + *Musca domestica*, used as outgroup) and restricted forms (four species of the sister radiations *virilis-repleta*/Hawaiian Drosophilidae + *D. simulans*, used as outgroup), encompassing 13 or 15 genes, respectively; and two involving 25 nuclear loci, again considered in the broader (18 Drosophilidae species + *Musca domestica*, used as outgroup) and restricted forms (four species of the sister radiations *virilis-repleta*/Hawaiian Drosophilidae + *D. simulans*, used as outgroup). Our results suggests that there are inconsistencies between datasets derived from mitochondrial and nuclear genomes in *Drosophila*, once the two groups of markers present contrasting signals in regard to the phylogenetic positioning of the *D. flavopilosa* group. While mitochondrial genomes support the clade formed by *D. incompta* and *D. mojavensis*, the nuclear genome, supports *D. incompta and D. virilis* as a sisters species. Thus, our results highlight the importance of integrative approaches for phylogenetic reconstruction.

Keywords: Long branch attraction, Bayesian concordance analysis, phylogenomic approaches, genes trees, species trees.

1. Introduction

With the advent of DNA and protein sequencing techniques in the mid-70s, the study of evolutionary events and relationships among different species has become more accessible, improving considerably the power of phylogenetic inferences. However, these first molecular phylogenetic analyses were based on limited datasets, and even on single genes, preventing the proper resolution of some nodes (Delsuc et al. 2005). According to Rokas et al. (2003) and Pollard et al. (2006), there might be widespread incongruence between phylogenies recovered from individual genes and, therefore, reliance on single or on a small number of genes has a significant probability of supporting incorrect relationships for the studied taxa.

Recently, large-scale sequencing has enabled the use of larger data sets, which have been crucial to elucidating the phylogenetic relationships among controversial or poorly-resolved clades, especially for deep nodes (Dunn et al. 2008). This phylogenomic approach solves several constraints associated to smaller dataset, but also presents potential problems related to the choice of the appropriate method of phylogenetic reconstruction, resulting in a lot of artifacts. There is also an ample debate about the best approach for clarifying the phylogenetic relationship of a species: sampling many genes for few species (Weigert et al. 2014), sampling many species for few loci (Jetz et al. 2012) or even combining these two strategies (Zheng and Wiens, 2016).

Inconsistencies and incongruences in phylogenetics may result from inaccurate reconstructions, or from real differences between gene and species trees. According to Delsuc et al. (2005), the three main kinds of bias that are not efficiently handled by most current phylogenetic reconstruction methods, leading to incorrect estimation of gene trees are compositional bias, long-branch attraction and heterotachy, but the presence of saturation may certainly be added to this list. Moreover, most methods do not consider that different parts of the genome may have been subject to different evolutionary histories, due to the occurrence

of: 1) incomplete lineage sorting, 2) introgressive hybridization; 3) paralogous gene sampling and 4) horizontal gene transfer (Pamilo and Nei, 1988; Maddison, 1997 and Zachos, 2009).

Although these analytical and biological factors may affect both, mitochondrial and nuclear markers, the strength of the effect is not homogeneously distributed. In this sense, whereas introgression seems to prevail in the mitochondrial genome (Keck and Near, 2010), especially among closely related animal species (Chan and Levin, 2005), incomplete lineage sorting is more common for nuclear genes (Wong et al. 2007), given their larger effective population sizes. Moreover, the effect of incomplete lineage sorting or deep coalescence (Maddison, 1997) is more likely when the effective population size of the ancestral population was large and when the time between speciation events was short (Pamilo and Nei, 1988). In regard to analytical bias, mitochondrial genes are frequently saturated due to the higher substitution rates (Brown et al. 1979) and to the highly biased nucleotide composition (Clary and Wolstenholme, 1985). At contrast, recombination affects mainly the nuclear genome, and can lead to inconsistencies on phylogenetic reconstructions because every site at every locus may have its own evolutionary history. According to Schierup and Hein (2000), besides implying a network evolutionary pattern, if ignored, recombination can lead to a large overestimation of terminal and total branch lengths, and even to the loss of the molecular clock.

Several of these aspects have been previously reported for the genus *Drosophila*, whose ancient diversification (Robe et al. 2010a) and peculiar evolutionary pattern involving multiple and successive radiations (Throckmorton, 1975) may hinder robust phylogenetic assignments. In this sense, inconsistencies of biological origin were detected, for example, in phylogenetic studies of the *Drosophila melanogaster* (Pollard et al. 2006), the *Drosophila willistoni* (Robe et al. 2010b) and the *Drosophila cardini* (De Ré et al. 2010) species groups, highlighting the genus as a putatively common target of incomplete lineage sorting and

introgression. Nevertheless, saturation is also commonly invoked to explain incongruent or poorly supported relationships within the genus (Robe et al. 2005 and 2010b; Silva-Bernardi et al. 2006) and this analytical bias seems to be especially challenging in regard to deep nodes.

Another yet misunderstood case of phylogenetic incongruence within *Drosophila* refers to the positioning of species of the *Drosophila flavopilosa* group. This group encompasses 17 described species (Bächli, 2016) with restricted ecology, which use flowers of *Cestrum* (Solanaceae) as sites for oviposition, larval development and feeding (Brncic 1966; Hofmann 1985). This group of species is assumed to be part of the *virilis-repleta* radiation of the *Drosophila* subgenus (Throckmorton, 1975; Robe et al. 2010a; De Ré et al. 2014b), but its positioning within this clade is still controversial. In this sense, whereas some genes support the *flavopilosa* group as closely related to the *repleta* group, to the exclusion of *D. virilis* (Robe et al. 2005; and *amd* in Robe et al. 2010a), other genes suggest the *flavopilosa* group as closer to *D. virilis* than to the *repleta* group (*ddc* in Robe et al. 2010b). Although this last positioning was recently supported by a whole-mitogenomic phylogenetic analysis (De Ré et al. 2014b), partial mitochondrial data (genes COI, COII and ND2) suggested the alternative resolution. In face of the recent sequencing by our group of the whole genome of *D. incompta*, a member of the *flavopilosa* group, we assess here the actual positioning of the *flavopilosa* group within *Drosophila* phylogeny through a Bayesian Concordance Analysis performed in a phylogenomic perspective. In this sense, concordance between and within mitochondrial and nuclear loci were inferred, and this held a more confident description of the phylogenetic position of *D. incompta*. This approach also allowed addressing the possible causes of the extensive incongruence among phylogenies in regard to this relationship.

2. Material and Methods

2.1. Data sets

The analyses were independently performed on four different datasets: two mitogenomic datasets, evaluated in the broader (18 Drosophilidae species + *Musca domestica*, used as outgroup) and restricted forms (four species of the sister radiations *virilis-repleta*/Hawaiian Drosophilidae + *D. simulans*, used as outgroup), encompassing 13 or 15 genes, respectively (Table 1); and two involving 25 nuclear loci (Table 2), again considered in the broader (18 Drosophilidae species + *Musca domestica*, used as outgroup) and restricted forms (four species of the sister radiations *virilis-repleta*/Hawaiian Drosophilidae + *D. simulans*, used as outgroup). For the mitochondrial data, with the exception of *D. elegans, D. rhopaloa* and *D. biarmipes*, mitogenome sequences were downloaded from GenBank (accession numbers presented on Table 1). For the nuclear genes, with the exception of *D. incompta*, sequences were downloaded using the GBrowse tool of Flybase (Attrill et al. 2015), based on the search for sequences orthologous to *D. melanogaster* genes, whose Ids are given on Table 2. These genes were chosen randomly, have variable functions and are single copy genes, avoiding the inclusion of paralogous sequences or pseudogenes, which do not reflect the evolutionary history of the species in question. Moreover, this set of genes shows a satisfactory assembly, considering reads available for *D. incompta*.

For *D. elegans*, *D. rhopaloa* and *D. biarmipes*, SRA's from the whole genome were downloaded from GenBank (Experiment Numbers SRX094534, SRX095455 and SRX095626, respectively) and the mitogenomes were assembled using the MITObim package (Hahn et al. 2013). Partial mitochondrial nucleotide sequences of 1704 bp (AF164596), 1515 bp (S76764.1) and 1629 bp (AY958403), were used to locate homologous sequence in the draft genomes of *D. elegans*, *D. rhopaloa* and *D. biarmipes*, respectively, and these sequences were then used as seed for the mtDNA total assemblies. MITObim uses an in silico baiting approach, which was implemented in the MIRAbait module of the MIRA assembler (v3.4.1.1) (Chevreux et al. 1999). The characterization and annotation of these

three assembled genomes was performed in MITOS Web Server (Bernt et al. 2013) and UGENE software (Okonechnikov et al. 2012), respectively (data not shown).

For *D. incompta*, total DNA was extracted for a pool of 20 individuals and subjected to a single-end genome sequencing approach, as performed by Fasteris DNA Sequencing Service with a Solexa-llumina HiSeq 2000 Next Generation Sequencing Device. The reads so obtained were further used to assemble each of the 25 employed nuclear genes in MITObim, using individual orthologous genes from *D. virilis* (Table 2) to locate homologous sequences in the draft genome of *D. incompta*. The mitogenome of this species was already assembled, annotated and characterized (De Ré et al. 2014b).

In each of these cases, orthologous sequences were aligned through the Clustal W algorithm, as implemented in Mega 5.0 (Tamura et al. 2011) and then submitted to Gblocks Server (Castresana, 2000) in order to delete poorly aligned of divergent and possibly saturated regions.

2.3. Phylogenetic analyses

Species trees were reconstructed for each of the four previously described combined datasets through Bayesian Concordance Analysis (Ané et al. 2007), in order to measure the number of genes providing phylogenetic signal to each particular clade or the level of concordance across genes. This analysis had the first-stage MCMCMC (Metropolis Coupled MCMC) carried out for the individual genes in MrBayes 3.2 (Ronquist et al. 2012), following the evolutionary models suggested by the AIC Test (Akaike, 1974), as implemented by MrModelTest 2.3 (Nylander, 2004) in conjunction with PAUP 4.0a146 (Swofford, 2003) (Table 3). In each of these cases, the Markov Chain Monte Carlo (MCMC) was run for 5,000,000 generations, sampling trees every 1,000 generations. The runs were stopped only after the convergence value fell below 0.01, and a burn-in of 25% of the generations was applied before summarizing trees and parameters. The number of distinct topologies

recovered by each gene along this search was quantified in order to get a measure of phylogenetic uncertainty.

The second-stage BCA analysis was implemented in BUCKy 1.2 (Ané et al. 2007) with alpha set to 1. For this second-stage, two runs with four chains were performed, each of which considering a total of 5,000,000 pos-burnin generations, sampling every 1,000. The primary concordance tree jointly with concordance factors was filtered from the *.concordance file and visualized in FigTree 3.1 program (Rambaut, 2009). BCA was also used to measure the posterior probability that two loci share the same topology, reflecting the level of concordance between loci.

Additionally, alternative species trees were constructed by a concatenated approach, for each of the four data sets, in MrBayes 3.2, according to the strategy described above. In this case, concatenated matrices were constructed for the different datasets in FASconCAT v. 1.0 (Kück and Meusemann, 2010).

2.3.1. Detection of saturation and compositional heterogeneity

Given that saturation and compositional heterogeneity are frequently pointed as important factors leading to biased tree reconstructions, their potential effects on the individual phylogenetic inferences were further assessed. For this, each locus was individually tested for substitutional saturation in TreSpEx program (Struck, 2014), using the broader datasets and measuring the slope of the linear regression between patristic and uncorrected distances (the higher the slope the less saturated is the dataset). The general results were represented in density plots generated in the R package. Additionally, the putative effect of compositional heterogeneity among taxa was evaluated for each individual gene through the Relative Composition Frequency Variability (RCFV) (Zhong et al. 2011) values, as measured from the broader matrixes in BaCoCa v. 1.1 (Kück and Struck, 2014) (the higher the RCFV, the more the composition of an individual sequence differs from the overall trend in the

dataset). In order to summarize these results, density plots and heatmaps with hierarchical clustering were generated with R.

3. Results

Bayesian Concordance Analysis (BCA) and concatenated analysis were performed on the four developed datasets: 1) the mitogenomic broad dataset (19 species x 13 genes); 2) the nuclear broad dataset (19 species x 25 genes); 3) the mitogenomic restricted dataset (5 species x 15 genes); and 4) the nuclear restricted dataset (5 species x 25 genes). The slope of the saturation linear regression (Struck, 2014) ranged from 0.00031 (for *ND5*) to 0.00507 (for *ATP8*) and 0.00018 (for *VPS35*) to 0.00115 (for *PRL1RB*), for the mitochondrial and nuclear data, respectively, whereas R^2 ranged from 0.7016 (for *ND4*) to 0.9359 (for *ATP8*) and 0.5094 (for *ADAR*) to 0.9377 (for *CAP60A*), respectively (Table 3). These values show that even after GBlocks pruning, saturation remained high for most of the evaluated loci. In fact, only three mitochondrial (*ATP8*, *ND3* and *ND4L*) and six nuclear loci (*CG14435*, *MP20*, *NAP1*, *PRL1RB*, *RAB35* and *VPS20*), presented slope values above 0.00118815 and 0.0005633, respectively, and these were further used in new BCA and concatenated reconstructions using the respective broad datasets (matrices of 19 x 3 and 19 x 6, respectively). RFCV values also warned the effect of skewed nucleotide composition, and ranged from 0.0003 (in *D. ananassae* and *D. erecta*) to 0.0026 (in *D. littoralis*), or from 0.0006 (in *D. persimilis*) to 0.0069 (in *D. willistoni*), considering the mitogenomic or nuclear datasets, respectively (Supplementary Material – Table S1 and S2). Mean RFCV values presented by each loci ranged from 0.000895 (for *COI*) to 0.001932 (for *ATP8*) and from 0.001816 (for *PRL1RB*) to 0.006621 (for *CG4585*), for mitochondrial and nuclear genes, respectively. The results concerning phylogenetic inferences are presented below.

3.1. Mitogenomic broad data set

Concerning the phylogenetic relationships, the primary concordance tree (PCT) recovered through Bayesian Concordance Analysis (BCA) for the total set of 13 mitochondrial protein coding genes, clustered *D. incompta* and *D. mojavensis*, with a concordance factor (CF) of 0.58; in other words, 58% of the mitochondrial loci present phylogenetic information supporting these as sister species (Fig. 1a). Interestingly, when the same dataset was employed in a concatenation strategy, this clade presented a posterior probability (PP) of 1.00, severely overestimating confidence for this relationship (Fig. 1b). This same pattern was also evidenced for other relationships within the tree, and although CF values for basal branching's were always of the order of 0.20 to 0.50, only one of the clades recovered in the total evidence tree (TET) presented PP values bellow 0.90. Moreover, the topology of the PCT and TET also differed in regard to the positioning of *D. grimshawi* and *D. willistoni* (Fig. 1). In this last case, both the basal branching of *D. willistoni* in the BCA and its clustering with species of the subgenus *Drosophila* in the TET are totally unexpected (O'Grady e Kidwell, 2002; Throckmorton, 1975), and are probably an artifact related to the biased nucleotide composition presented by this species (Rodríguez-Trelles and Tarrío, 2000). In this sense, the phylogenetic signal of the mitochondrial genes was probably swamped by the several homoplasies shared between *D. willistoni* and *D. virilis*, which present similar nucleotide composition patterns, as indicated by the heatmaps with hierarchical clustering (Fig. 2a). Besides, the positioning of this species is also probably affected by the phenomenon of Long Branch Attraction (LBA).

Saturation was another factor that could be hindering robust phylogenetic assessments. In fact, ten of the 13 protein coding genes exhibited significant signs of saturation, with slope values lower than or equal to 0.00118815, and these were: *ATP6, COI, COII, COIII, CYTB, ND1*, *ND2*, *ND4*, *ND5* and *ND6*. Nevertheless, when only the remaining three genes (*ATP8*, *ND3* and *ND4L*) presenting slope values higher than 0.00118815 were used in the BCA, the

same clade clustering *D. incompta* with *D. mojavensis* was recovered, but now with even lower CF (0.227) (Fig 1c). This strategy also resulted in the loss of the phylogenetic signal for the monophyly and the internal groupings of the *melanogaster* group and subgroup (see Lewis et al. (2005) for more information) and altered the branching site of *D. grimshawi* in regard to the total mitogenomic broad data set BCA analysis. In this last case, neither mitogenomic phylogeny recovered the expected branching pattern, since this species is part of a group considered to be the sister clade of the *virilis-repleta* radiation (Remsen and O'Grady, 2002; Robe et al. 2010a; Tatarenkov et al. 2001).

3.2. Nuclear broad dataset

Contrary to the mitogenomic dataset, which presented *D. incompta* as sister to *D. mojavensis*, the nuclear broad dataset clustered *D. incompta* with *D. virilis*, both, in the PCT (Fig. 3a) and in the TET (Fig. 3b). The CF presented by this clade in the PCT (0.976) reveals that 98% of the sampled genes shared this evolutionary history, which is also highly supported in the TET (PP = 1.00). The CF's presented by other clades of the nuclear PCT were also, in general, higher than those of the mitochondrial PCT, and except for the positioning of *D. biarmipes*, the PCT is totally congruent with the TET. Nevertheless, unconventional relationships were again recovered for *D. willistoni*, presented as an early offshoot of *Drosophila* both, in the PCT (CF = 0.682) and in the TET (PP = 1.00).

Although nuclear loci were in general more concordant between each other than mitochondrial loci (see Table 4 a and b), as for the mitochondrial genome, most nuclear genes showed high levels of saturation, and only six genes presented slope values higher than 0.0005633 (*CG14435*, *MP20*, *NAP1*, *PRL1RB*, *Rab35* and *VPS20*). In this case, the exclusion of genes with high saturation levels did not affect the positioning of *D. incompta*, whose clustering with *D. virilis* was supported by 99% of the sampled genes (CF = 0.994). Nevertheless, the withdrawn of the saturated nuclear genes seems to have caused the recovery

of primary phylogenetic signal concerning the positioning of *D. willistoni* inside the *Sophophora* subgenus (CF = 0.807) (Fig. 3c) (O'Grady e Kidwell, 2002; Clark et al. 2007 - *Drosophila* 12 Genomes Consortium). So, although for the nuclear genes *D. willistoni* does not seem to departure so much from the general nucleotide composition patterns (Fig. 2b), saturation and associated LBA seems to be the best explanation for the unconventional positioning of this species.

3.3. Mitogenomic and nuclear restricted datasets

In order to clear our results, while evaluating the effect of taxon sampling, mitogenomic and nuclear inferences were repeated for a reduced set of species, containing only four OTU`s plus the outgroup. The results of the individual BAs performed in the first-stage of the BCA are summarized in Table 5, where it can be seen that, in general, whereas only seven of the 13 protein encoding mitochondrial loci support the clustering of *D. incompta* with *D. mojavensis* with moderate to high confidence (PP ranging from 0.59 to 0.99), 23 of the 25 nuclear loci support the grouping of *D. incompta* with *D. virilis* with general high confidence (PP ranging from 0.83 to 1.00). These contrasting results can also be seen in the PCT constructed in the second-stage of the BCA, where the mitochondrial clustering of *D. incompta* is discordant (CF = 0.70) (Fig. 1d), whereas nuclear clustering is highly concordant among loci (CF = 0.98) (Fig. 3d).

4. Discussion

In recent years, we were faced by a significant increase in the availability of molecular data for phylogenetic inferences, which advanced from studies involving a single gene to studies comprising hundreds of concatenated loci. Despite this breakthrough, the elucidation of the evolutionary relationships between species involves many challenges that create uncertainties and ambiguities regarding the real species tree. In this sense, there are still controversies regarding the use of larger data sets, as in phylogenomics: whereas some

authors consider that large datasets can lead to congruence in phylogenetic analyzes (Gee, 2003), others argue that the use of larger datasets can enhance the inconsistencies due to the presence of artifacts such as saturation, compositional heterogeneity and increased in substitution rates (Jeffroy et al. 2006). Indeed, our results point to the presence of inconsistencies between datasets derived from mitochondrial and nuclear genomes in *Drosophila*. Generally, the two groups of markers present contrasting signals: (1) in regard to the phylogenetic positioning of the *D. flavopilosa* group; (2) in regard to the level of concordance between loci.

Considering the mitogenomic strategy performed here, the PCT constructed with the 13 mitochondrial protein coding genes shows *D. incompta* as sister to *D. mojavensis* (CF = 0.58), composing a clade closely related to *D. grimshawi* (CF = 0.282) (Fig. 1a). Interestingly, the clustering of *D. incompta* and *D. mojavensis* was also recovered when only the three less saturated mitochondrial genes were employed in the BCA (CF = 0.227) (Fig. 1c) and also when only four species considered part of the *Drosophila* subgenus (Robe et al. 2010a) were used in the same analysis (CF = 0.703) (Fig. 1d). The low CF values illustrate that only a small proportion of the mitogenome supports this relationship, which is at first unexpected given that mtDNA has a single gene tree history. Nevertheless, low CFs in mitogenomic analysis were also reported by Weisrock (2012), which stated this could be related to: (1) the lack of phylogenetic signal in some genes, which might be too conserved or saturated; (2) error in tree reconstruction. Because in our case, the same topology is recovered when artefactual properties of the dataset, as saturation and taxon sampling, are addressed, and as seven of the 13 protein coding genes + *rRNA12S* support *D. incompta* and *D. mojavensis* as sister-species when analyzed in isolation (Table 5), we argue here that the small CFs is a result of phylogenetic signal in a small number of sites. In this sense, although the mitogenomic gene tree might in fact present such a resolution, most of the sites may have lost

phylogenetics signal of may have not mutated at all, so that the mitochondrial genome contains a mixture of saturated and conserved position. It is also important to mention that even with this straightforward inconsistency, Bayesian analyses performed with the combined mitochondrial genes resulted in complete resolution for this clade (PP = 1.00) (Fig. 1b). In fact, concatenation strategies masked prior inconsistencies presented in the mitochondrial genome, inflating support for almost all recovered relationships. Such a result was also previously reported by several other authors and illustrates the need of employing more robust methods of phylogenetic assessment in phylogenomic perspectives.

Among the set of 25 nuclear genes here employed, concordance was higher than that presented by mitochondrial loci (Table 5). In general, these markers supported *D. incompta* as sister to *D. virilis*, to the exclusion of *D. mojavensis*, and CF values for this relationships were of the order of 0.97, 0.99 or 0.98 in the BCA performed with the broad dataset with the 25 genes and with the six less saturated genes, or with the restricted dataset with all genes, respectively (Fig. 3a, c and d). In fact, only *CG42265* and *PRL1RB* supported alternative resolutions for *D. incompta*, but with moderate to low support values (Table 5). This straightforward concordance occurs despite the possibility of independent evolution between nuclear markers, and even despite the greater mean saturation and compositional bias presented by the nuclear data in comparison to the mitogenome (slope of 0.000477 and 0.001267, R2 of 0.80 and 0.85, RCFV of 0.003709 and 0.001218, respectively).

The inconsistencies between the nuclear and mitochondrial genomes (mito-nuclear discordance) have already been highlighted in several animal species, including other species of the genus *Drosophila* such as *Drosophila pseudoobscura* and *D. persimilis* (Powel, 1983)*; D. santomea* and *D. yakuba* (Bachtrog et al. 2006); and *D. simulans* and *D. mauritiana* (Aubert and Solignac 1990). Moreover, a recent review (Toews and Brelsford, 2012) compiled 126 recent cases in animal systems in which there is strong discordance between the

two datasets. Although our results suggest an important effect of saturation in the huge and small inconsistences recovered within mitochondrial and nuclear datasets, respectively, we argue here that the incongruence between these two sets of data is probably derived from the independent evolutionary histories presented by these markers. In this case, the fact that almost all analyzed nuclear genes support the same topological resolution in regard to the positioning of *D. incompta* allows hypothesizing this as the true species history of this species and its group. The different evolutionary history presented by the mitochondrial loci may, in fact, derive from an ancient hybridization between ancestors of the *flavopilosa* and *repleta* groups, with introgression of the mitochondrial genome. Within *Drosophila*, the interference of introgression in the phylogenetic signal presented by mitochondrial genes has been also previously suggested for the *D. cardini* (De Ré et al. 2010) and the *D. willistoni* species groups (Robe et al. 2010b). Moreover, mitochondrial introgression as a result of hybridization events have also been pointed as the source of conflict between mitochondrial and nuclear loci within the genus *Delma* (Brennan, Bauer and Jackman, 2016).

Despite the possibility of mitochondrial introgression, due to saturation, individual mitochondrial genes also seem to be widely uninformative in *Drosophila* when the phylogenetic relationships are meant to resolve the oldest diversification order (Robe et al. 2005). This is illustrated here by the unconventional relationships recovered by the three less saturated mitochondrial loci in regard to the monophyly of the *melanogaster* group and the *melanogaster* subgroups, which are already widely supported (Schawaroch, 2012 and Lewis et al. 2005, respectively). Saturation may also have affected the positioning of *D. willistoni* in both, the mitochondrial and the nuclear PCTs and TETs. Nevertheless, despite the recovery of the conventional grouping of this species as sister to the *saltans* species group (O'grady and Kidwell, 2002) by the less saturated nuclear partitions (Fig. 3c), the mitochondrial less saturated loci cluster this species with the *virilis-repleta/*Hawaiian drosophilid radiation (Fig.

1c). This is probably an outcome of the convergent nucleotide composition patterns presented by *D. willistoni* and *D. virilis* for the set of protein coding mitochondrial genes (Fig. 2a). In fact, it was previously reported that failure to account for nucleotide base composition variation among sequences can lead to incorrectly reconstructed tree topologies - sequences of similar base compositions may become erroneously clustered - (Gaultier and Gouy, 1995) and to branch lengths that reflect changes in nucleotide composition rather than changes in substitution rate (Tourasse and Li 1999). Moreover, the analysis of eight nuclear genes unambiguously corroborated that the common ancestor of *Sophophora* had an elevated GC content (Rodriguez-Trelles, Tarrion and Ayala, 2000), a pattern that seems to have been at least partially maintained in the *saltans* and *willistoni* species groups (Rodriguez-Trelles, Tarrion and Ayala, 2000). Another artifact that can justify the positioning of this species along the nuclear and mitochondrial PCTs and TETs is Long Branch Attraction (LBA) that is a well-known phylogenetic artifact that causes sequences that are on long branches (which can occur because the lineages have accelerated evolutionary rates or because they are on isolated evolutionary branches) to incorrectly appear as closely related (reviewed in Bergsten, 2005).

In general, the results detected by this study highlight the importance of integrative approaches for phylogenetic reconstruction. Only through the use of a set of mitochondrial and nuclear genes in multiple BCA approaches it became possible to approach the real subjacent species tree, and to identify the factors affecting the resulting topologies.

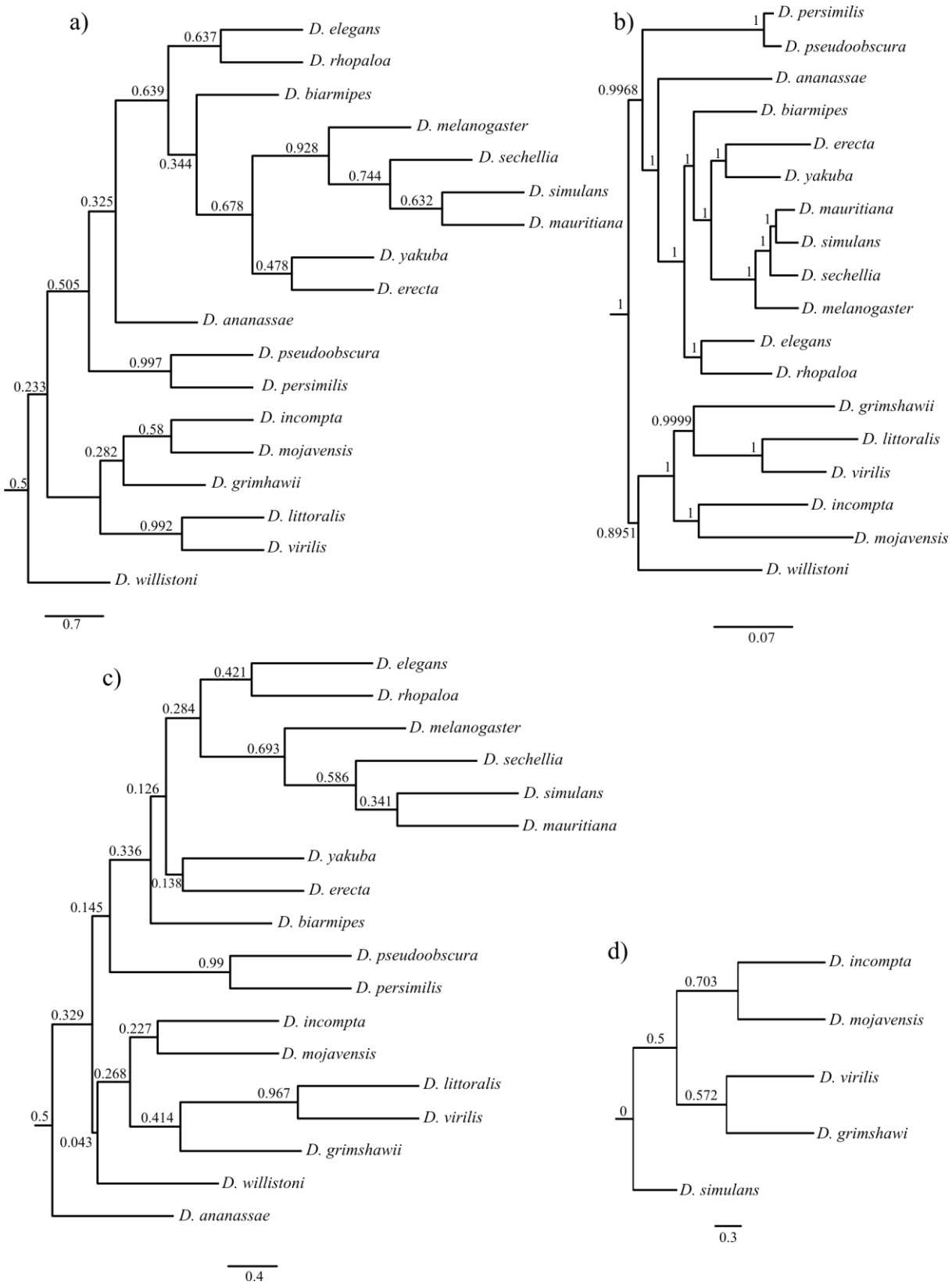


Fig 1: Results of the mitogenomic phylogenetic analyses. a) Primary Concordance Tree (PCT) recovered through Bayesian Concordance Analysis (BCA) for the broad dataset (19 species x 13 protein coding genes). Clade CFs given above branches. b) Total Evidence Tree (TET) recovered by Bayesian Analysis for the concatenated broad dataset (19 species x 13 protein coding genes). PP of each clade given above branches. c) PCT recovered through BCA with the set of three mitochondrial less saturated genes (19 species x 3 protein coding genes). Clade CFs given above branches. d) PCT recovered through BCA for the mitochondrial restricted dataset (5 species x 13 protein coding genes + 2 rRNA genes). The outgroup of 1a, 1b and 1c were omitted from the tree.

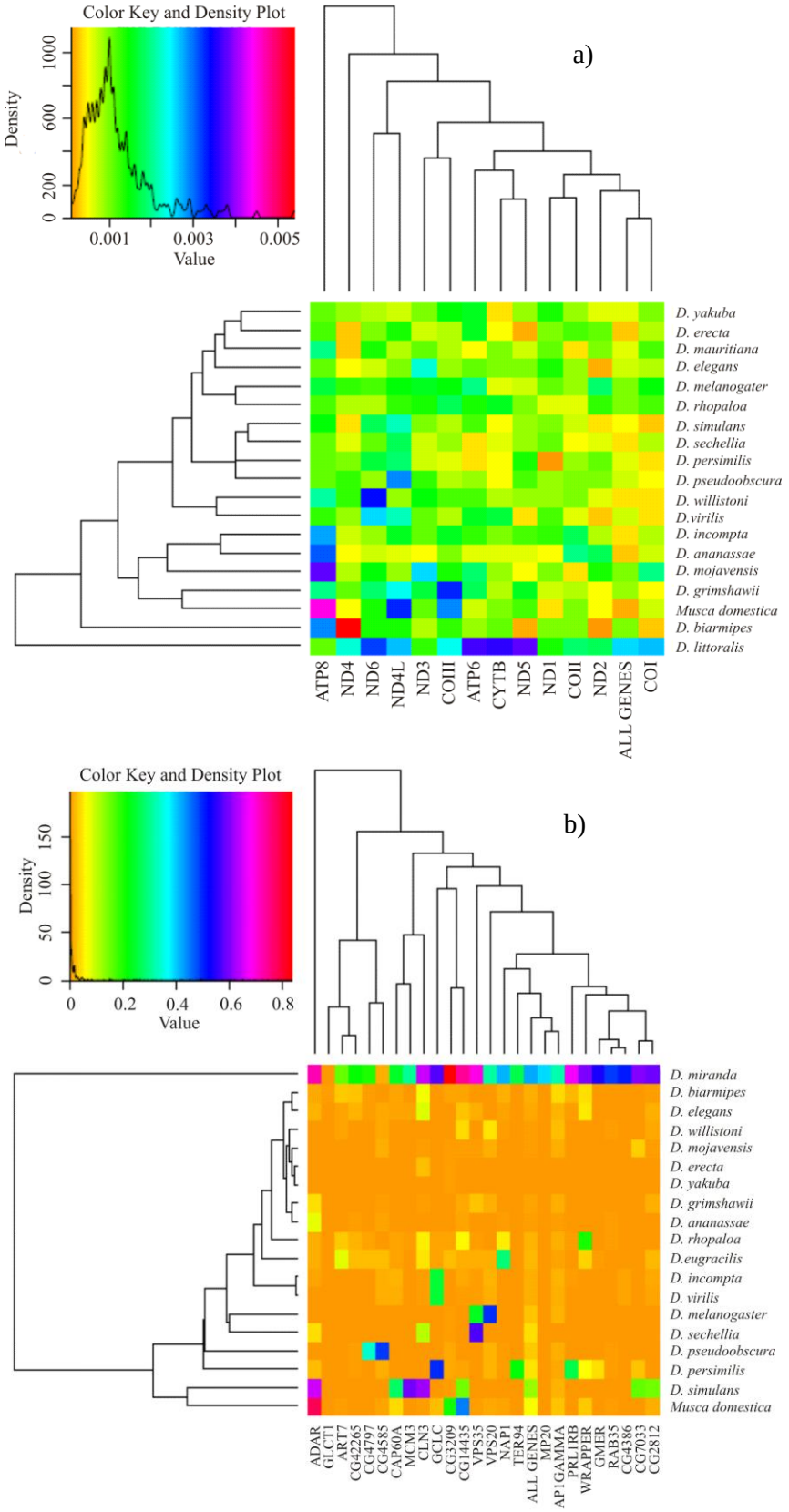


Fig. 2: Taxon to gene heat map and results of hierarchical clustering based on the RCFV values presented by a) Each of the 13 protein coding mitochondrial genes in each of the 19 studied species. b) Each of the 25 nuclear genes and each of the 19 studied species. Y-axis shows the taxa whereas x-axis represents the partitions. Different colors reflect different RCFV values, with color key and density plots presented above the tree.

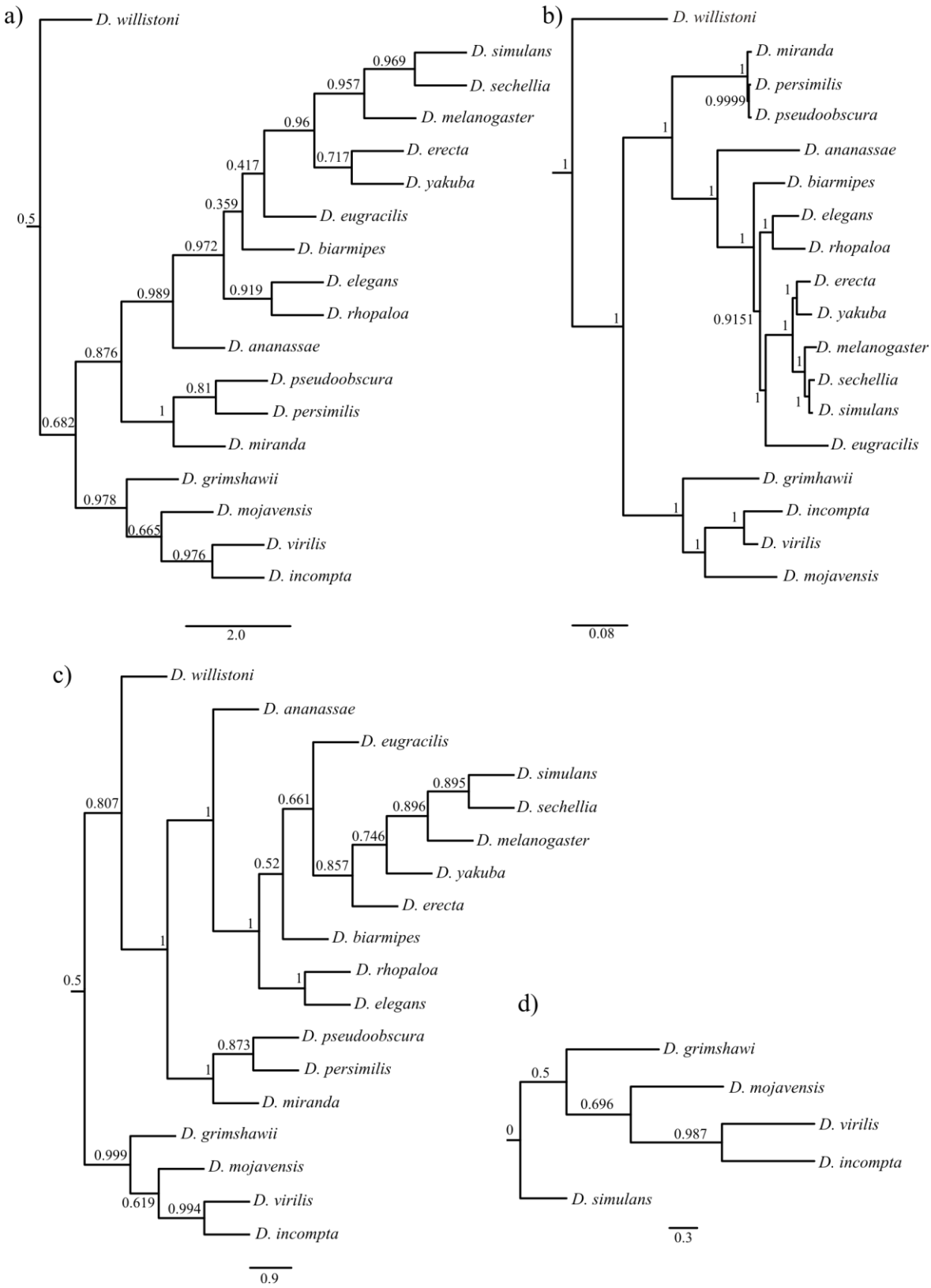


Fig 3: Results of the nuclear phylogenetic analyses. a) Primary Concordance Tree (PCT) recovered through Bayesian Concordance Analysis (BCA) for the broad dataset (19 species x 25 protein coding genes). Clade CFs given above branches. b) Total Evidence Tree (TET) recovered by Bayesian Analysis for the concatenated broad dataset (19 species x 25 protein coding genes). PP of each clade given above branches. c) PCT recovered through BCA with the set of six less saturated nuclear genes (19 species x 6 protein coding genes). Clade CFs given above branches. d) PCT recovered through BCA for the nuclear restricted dataset (5 species x 25 protein coding). The outgroup of 3a, 3b and 3c were omitted from the tree.

Table 1: List of species included in the mitogenomic analyses, encompassing the broader (all 18 Drosophilidae species + *Musca domestica*, used as outgroup) and restricted datasets (four species of the sister radiations *virilis-repleta*/Hawaiian Drosophilidae, marked with an "*" + *D. simulans*, used as outgroup)

Genus	Subgenus	Radiation	Species Group	Species Subgroup	Species	Accession numbers from complete mitochondrial	Reference
Idiomyia		Hawaiian-Drosophilidae	*picture-wing*	*grimshawi*	*D. grimshawii**	BK_006341	Montooth et al. (2009)
Drosophila	*Drosophila*	*virilis-repleta*	*virilis*	*virilis*	*D. virilis**	BK_006340	Montooth et al. (2009)
				-	*D. littoralis*	NC_011596	Andrianov et al. (2010)
			repleta	*mulleri*	*D. mojavensis**	BK_006339	Montooth et al. (2009)
			flavopilosa	*nesiota*	*D. incompta**	KM_275233	De Ré et al. (2014)
	Sophophora	*Sophophora*	*melanogaster*	*melanogaster*	*D. melanogaster*	NC_001709	Garesse (1988)
					D. simulans	JQ_691661	Ballard (2000a, b),
					D. sechellia	NC_005780	Ballard (2000a, b),
					D. mauritiana	NC_005779	Ballard (2000a, b)
					D. yakuba	KF_824901	Clary and Wolstenholme, (1985)
					D. erecta	BK_006335	Montooth et al. (2009)
				ananassae	*D. ananassae*	BK_006336	Montooth et al. (2009)
				suzukii	*D. biarmipes*	-	This thesis
				rhopaloa	*D. rhopaloa*	-	This thesis
				elegans	*D. elegans*	-	This thesis
			obscura	*pseudoobscura*	*D. pseudoobscura*	NC_018348	(Torres et al. (2009)
				pseudoobscura	*D. persimilis*	BK_006337	Montooth et al. (2009)
			willistoni	*willistoni*	*D. willistoni*	BK_006338	Montooth et al. (2009)
Musca					*Musca domestica*	KT444442	Li et al. (2014)

Note: Classification followed Bächli (2016) (http://www.taxodros.uzh.ch/search/prt_rawfile.php?prt=SPECIES-LIST_GR_SR_SC), accessed on February, 22, 2016), except for the radiation subdivision, which followed Remsen & O`Grady (2002) and Robe et al. (2010a). In the broader dataset, genes evaluated were: 1: ATP6, 2: ATP8, 3: COI, 4: COII, 5: COIII, 6: CYTB, 7: ND1, 8: ND2, 9: ND3, 10: ND4, 11: ND4L, 12: ND5 and 13: ND6. Sequences of rRNA12s and rRNA16s were added to these in the restricted mitogenomic dataset.

Table 2: List of single copy orthologous genes included in the analyses of nuclear broader and restricted datasets. Download was performed on Flybase GBrowse tool based on a search for sequences orthologous to *D. melanogaster* genes, entered with the respective IDs. Conversely, for *D. incompta*, genes were assembled from the draft genome using orthologous sequences from *D. virilis*.

Nuclear genes	Full name	Chromosome location	ID for *D. melanogaster*
ADAR	*Adenosine deaminase acting on RNA*	X	FBgn0026086
AP1GAMMA	*AP-1γ*	X	FBgn0030089
ART7	*Arginine methyltransferase 7*	2R	FBgn0034817
CAP60A-RA	*Ca-P60A-RA*	2R	FBgn0263006
CG2812	*CG2812*	2R	FBgn0034931
CG3209	*CG3209*	2R	FBgn0034971
CG4386	*CG4386*	2R	FBgn0034661
CG4585	*CG4585*	2R	FBgn0025335
CG4797	*CG4797*	2R	FBgn0034909
CG7033	*CG7033*	X	FBgn0030086
CG14435	*CG14435*	X	FBgn0029911
CG42265	*CG42265*	X	FBgn0259150
CLN3	*CLN3*	3L	FBgn0036756
GLCT1	*GLCT1*	2R	FBgn0067102
GCLC	*Glutamate-cysteine ligase catalytic subunit*	X	FBgn0040319
GMER	*GDP-4-keto-6-deoxy-D-mannose 3,5-epimerase/4-reductase*	2R	FBgn0267823
MCM3	*Minichromosome maintenance 3*	X	FBgn0024332
MP20	*muscle protein 20*	2R	FBgn0002789
NAP1	*Nucleosome assembly protein 1*	2R	FBgn0015268
PRL1RB	*PRL-1-RB*	2L	FBgn0024734
RAB35	*Rab35*	X	FBgn0031090
TER94-RA	*TER94-RA*	2R	FBgn0261014
VPS20	*Vacuolar protein sorting 20*	2R	FBgn0034744
VPS35	*Vacuolar protein sorting 35*	2R	FBgn0034708
WRAPPER	*Wrapper*	2R	FBgn0025878

Note: The species included in the restricted nuclear dataset are those marked with an "*" on Table 1. The species included in the broader nuclear dataset were the same as those used in the mitogenomic broader dataset (Table 1), to the exclusion of *D. littoralis* and the inclusion of *D. eugracilis*, which is part of the *melanogaster* group, subgroup *eugracilis* (Bächli, 2016).

Table 3: List of mitochondrial and nuclear loci used in the analyses, with their respective sizes in base pairs (pre and pos GBlocks), the total number of distinct topologies recovered in MrBayes 3.2, the selected evolutionary model, the Slope and R^2 values from the saturation test and the mean Relative Composition Frequency Variability (RCFV)

Location	Gene	Total number of characters (pre-GBlocks)	Number of characters retained (pos-GBlocks)	Number of distinct trees	Model	Slope	R^2	Mean RCFV
Mitochondrial	*ATP6*	669	666	3835	GTR+I+G	0.000910623	0.87114819	0.001305
	*ATP8**	162	159	7457	HKY+G	0.005073284	0.9359177	0.001932
	COI	1509	1509	2355	GTR+I+G	0.000338107	0.80912395	0.000895
	COII	672	672	4434	GTR+I+G	0.000889835	0.86566351	0.000974
	COIII	793	709	3762	GTR+G	0.001188146	0.90976333	0.001384
	CYTB	1125	1125	1381	GTR+I+G	0.000392916	0.79880939	0.000984
	ND1	921	921	5402	GTR+I+G	0.000611719	0.82877967	0.000953
	ND2	924	887	1949	HKY+I+G	0.000714475	0.86820752	0.000895
	*ND3**	348	347	4478	GTR+I+G	0.001734750	0.90206912	0.001237
	ND4	1335	1335	2089	GTR+I+G	0.000372058	0.70163222	0.0011
	*ND4L**	259	259	7500	HKY+I+G	0.002892492	0.91846559	0.001658
	ND5	1696	1688	1167	GTR+I+G	0.000306326	0.801298553	0.001016
	ND6	515	507	3359	GTR+I+G	0.001046760	0.848447904	0.001495
Nuclear	*ADAR*	1795	1762	3359	HKY+G	0.000227742	0.509405693	0.001884
	AP1GAMMA	3011	2889	39	SYM+I+G	0.000209574	0.917591158	0.003684
	ART7	2104	1968	83	GTR+I+G	0.000286838	0.912954996	0.003053
	CAP60A	3007	2860	84	GTR+I+G	0.000209934	0.937707547	0.004505
	CG2812	1027	999	171	GTR+G	0.000500112	0.789817488	0.006179
	CG3209	1624	1589	145	GTR+I+G	0.000330783	0.804704363	0.004379
	CG4386	1127	990	164	GTR+I+G	0.000520804	0.829012154	0.003274
	CG4585	1194	1111	338	GTR+I+G	0.000486178	0.824809777	0.006621
	CG4797	1505	1322	147	GTR+I+G	0.000360839	0.818378115	0.003284
	CG7033	1603	1599	282	GTR+I+G	0.00034395	0.802907526	0.004242
	*CG14435**	969	751	524	GTR+G	0.000659851	0.742648548	0.004163
	CG42265	1629	1586	122	GTR+I+G	0.000319853	0.852243464	0.003063
	CLN3	1263	1219	151	GTR+I+G	0.000332721	0.609186674	0.002447

GCLC	1488	1480	50	GTR+I+G	0.000311294	0.698054292	0.003553
GLCT1	1176	1176	385	GTR+I+G	0.000511344	0.921232505	0.003763
GMER	964	962	107	GTR+I+G	0.000563396	0.828378038	0.002963
MCM3	1974	1967	40	GTR+I+G	0.000258186	0.750819534	0.003642
*MP20**	555	552	1012	GTR+G	0.001093187	0.872631568	0.002684
*NAP1**	812	795	354	GTR+I+G	0.000571932	0.828907564	0.005611
*PRL1RB**	547	516	4162	GTR+I+G	0.001150923	0.833166186	0.001816
*RAB35**	610	592	207	GTR+G	0.001052842	0.897389866	0.005295
TER94	2388	2376	49	SYM+I+G	0.000222383	0.877547364	0.003079
*VPS20**	678	630	1007	GTR+I+G	0.00078017	0.794997443	0.004184
VPS35	2416	2405	54	GTR+I+G	0.0001822	0.618003596	0.002405
WRAPPER	1332	1287	27	GTR+I+G	0.000430485	0.837287528	0.002958

Notes: Mitochondrial and nuclear loci presenting slope values above 0.00118815 and 0.0005633, respectively, are marked by an "*", and these were used in further analyses with the aim of reducing the effect of saturation.

Table 4a: Posterior probability (PP) values that two genes share the same topology considering all mitochondrial genes

	ATP6	*ATP8*	*COI*	*COII*	*COIII*	*CYTB*	*ND1*	*ND2*	*ND3*	*ND4*	*ND4L*	*ND5*	*ND6*
ATP6	1	0	0	0	0	0	0	0.3435	0.0213	0	0	0	0
ATP8	0	1	0	0	0	0	0	0	0	0	0	0	0
COI	0	0	1	0	0	0	0	0	0	0	0	0	0
COII	0	0	0	1	0	0.909	0	0	0	0	0	0.7209	0
COIII	0	0	0	0	1	0	0	0	0	0	0	0	0
CYTB	0	0	0	0.909	0	1	0	0	0	0.0113	0	0.753	0
ND1	0	0	0	0	0	0	1	0	0	0	0	0	0
ND2	0.3435	0	0	0	0	0	0	1	0.635	0.6564	0	0	0
ND3	0.0213	0	0	0	0	0	0	0.635	1	0.9603	0	0.2466	0
ND4	0	0	0	0	0	0.0113	0	0.6564	0.9603	1	0	0.2574	0
ND4L	0	0	0	0	0	0	0	0	0	0	1	0	0
ND5	0	0	0	0.7209	0	0.753	0	0	0.2466	0.2574	0	1	0
ND6	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 4b: Posterior probability (PP) values that two genes share the same topology considering all nuclear genes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0094	0	0	0	0	0	0
2	0	1	0	0	0	0	0.0088	0	0	0.0114	0.0088	0	0	0.0088	0.1067	0.9836	0	0	0.2904	0	0	0.9911	0	0.9912	0.9548
3	0	0	1	0	0.9944	0.025	0	0.9742	0	0	0	0	0.0195	0	0	0	0	0	0.002	0.021	0	0	0	0	0
4	0	0	0	1	0	0.046	0	0	0.0098	0.6559	0	0	0	0	0	0.0004	0	0.6674	0.5123	0	0	0.0001	0	0	0
5	0	0	0.9944	0	1	0.0251	0	0.9729	0.0012	0	0	0	0.0193	0	0	0	0	0	0.0017	0.0207	0	0	0	0	0
6	0	0	0.025	0.046	0.0251	1	0	0	0.6306	0.2855	0	0	0.6461	0	0	0	0	0.2846	0.0949	0.0246	0	0	0	0	0.0094
7	0	0.0088	0	0	0	0	1	0	0	0	0.9856	0	0	0.9999	0.8999	0.0039	0	0	0	0	0	0	0	0	0
8	0	0	0.9742	0	0.9729	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0.0098	0.0012	0.6306	0	0	1	0	0	0	0.9667	0	0	0.0002	0	0	0.0628	0.3438	0	0.0001	0	0	0
10	0	0.0114	0	0.6559	0	0.2855	0	0	0	1	0.0075	0	0	0	0.0002	0.0101	0	0.9869	0.4436	0	0	0.0114	0	0.0114	0.0481
11	0	0.0088	0	0	0	0	0.9856	0	0	0.0075	1	0	0	0.9855	0.8886	0.0047	0	0.0075	0	0.0069	0	0	0.0069	0	0.0069
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0.8355	0	0
13	0	0	0.0195	0	0.0193	0.6461	0	0	0.9667	0	0	0	1	0	0	0	0	0	0.065	0.3763	0	0	0	0	0
14	0	0.0088	0	0	0	0	0.9999	0	0	0	0.9855	0	0	1	0.8998	0.0039	0.0001	0	0	0	0	0	0	0	0
15	0	0.1067	0	0	0	0	0.8999	0	0	0.0002	0.8886	0	0	0.8998	1	0.1035	0	0	0.0363	0	0	0.1001	0	0.1001	0.0951
16	0	0.9836	0	0.0004	0	0	0.0039	0	0.0002	0.0101	0.0047	0	0	0.0039	0.1035	1	0	0	0.2924	0.0033	0	0.9925	0.0033	0.9924	0.9559
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0001	0	0	1	0	0	0	0	0	0	0	0
18	0	0	0	0.6674	0	0.2846	0	0	0	0.9869	0.0075	0	0	0	0	0	0	1	0.4513	0	0	0	0	0	0.0367
19	0.0094	0.2904	0.002	0.5123	0.0017	0.0949	0	0	0.0628	0.4436	0	0	0.065	0	0.0363	0.2924	0	0.4513	1	0.1198	0	0.293	0.1148	0.293	0.2751
20	0	0	0.021	0	0.0207	0.0246	0	0	0.3438	0	0.0069	0	0.3763	0	0	0.0033	0	0	0.1198	1	0	0	0.1645	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
22	0	0.9911	0	0.0001	0	0	0	0	0.0001	0.0114	0	0	0	0	0.1001	0.9925	0	0	0.293	0	0	1	0	0.9999	0.9633
23	0	0	0	0	0	0	0	0	0	0	0.0069	0.8355	0	0	0	0.0033	0	0	0.1148	0.1645	0	0	1	0	0
24	0	0.9912	0	0	0	0	0	0	0	0.0114	0	0	0	0	0.1001	0.9924	0	0	0.293	0	0	0.9999	0	1	0.9633
25	0	0.9548	0	0	0	0.0094	0	0	0	0.0481	0.0069	0	0	0	0.0951	0.9559	0	0.0367	0.2751	0	0	0.9633	0	0.9633	1

Note: Each number represent a nuclear gene: 1)*ADAR, 2) AP1GAMMA, 3) ART7, 4) CAP60A, 5) CG2812, 6) CG3209, 7) CG4386, 8) CG4585, 9) CG4797, 10) CG7033, 11) CG14435, 12) CG42265, 13) CLN3, 14) GCLC, 15) GLCT1, 16) GMER, 17) MCM3, 18) MP20, 19) NAP1, 20) PRL1RB, 21) RAB35, 22) TER94, 23) VPS20, 24) VPS35, 25) WRAPPER.*

Table 5: Summary of the results recovered in the individual BAs performed for the restricted dataset in the first-stage of the BCA, in relation to positioning of *D. incompta* and the posterior probability associated with the respective clade

Gene	Posterior probability (PP)			
	D. incompta plus *D. virilis*	*D. incompta* plus *D. mojavensis*	*D. incompta* plus *D. grimshawii*	Other topologies
ADAR	0.9985	*	*	*
AP1GAMMA	1	*	*	*
ART7	1	*	*	*
CAP60A	1	*	*	*
CG2812	1	*	*	*
CG3209	1	*	*	*
CG4386	1	*	*	*
CG4585	0.9501	*	*	*
CG4797	1	*	*	*
CG7033	1	*	*	*
CG14435	0.8375	*	*	*
CG42265	*	*	*	*D. incompta* plus *D. simulans* = (0.6285)
CLN3	0.8407	*	*	*
GLCT1	1	*	*	*
GCLC	1	*	*	*
GMER	1	*	*	*
MCM3	1	*	*	*
MP20	0.95	*	*	*
NAP1	0.9941	*	*	*
PRL1RB	*	*	*	Polytomy of *D. incompta*, *D. grimshawii* and *D. mojavensis* = (0.8322)
RAB35	0.9997	*	*	*

TER94	0.9915	*	*	*
VPS20	0.9521	*	*	*
VPS35	1	*	*	*
WRAPPER	1	*	*	*
ATP6	*	0.6883	*	*
ATP8	*	*	*	Polytomy of *D. incompta*, *D. grimshawi* and *D. virilis* = (0.7111)
COI	*	0.5973	*	*
COII	*	*	*	(*D. incompta*, *D. mojavensis*, (*D. virilis*, *D. grimshawii*)) = (1.00)
COIII	*	*	0.9287	*
CYTB	*	*	*	(*D. incompta*, (*D. virilis*, *D. grimshawii*)) = (0.9075)
ND1	*	0.9895	*	*
ND2	*	0.6557	*	*
ND3	*	0.6562	*	*
ND4	*	0.9947	*	*
ND4L	*	*	*	Entire polytomy
ND5	*	0.9943	*	*
ND6	0.6496	*	*	*
RNA12S	*	0.9944	*	*
RNA16S	*	*	0.7685	*

Note: The "*" indicates that the phylogenetic relationship concerned was not recovered by the respective marker.

5. Supplementary material

Table 1S: RCFV values for 13 mitochondrial loci in 19 *Drosophila* species

	RCFV values																		
	Species																		
		gri	*pse*	*moj*	*mel*	*ana*	*ere*	*mau*	*ele*	*rho*	*per*	*musca*	*yak*	*wil*	*sim*	*vir*	*lit*	*inc*	*bia*	*sec*
Mitochondrial loci	All loci	0.0009	0.0007	0.001	0.001	0.0003	0.0003	0.0006	0.0007	0.001	0.0006	0.0002	0.0006	0.0004	0.0005	0.0007	0.0026	0.0004	0.001	0.0004
	ATP6	0.0018	0.0009	0.0019	0.002	0.0006	0.0016	0.0005	0.0011	0.0014	0.0004	0.0007	0.0016	0.0012	0.0012	0.0016	0.0037	0.0013	0.0009	0.0004
	ATP8	0.002	0.0011	0.0038	0.0017	0.0031	0.0012	0.002	0.0011	0.0012	0.0011	0.0045	0.0011	0.0021	0.0015	0.0013	0.0011	0.0028	0.0029	0.0011
	COIII	0.0033	0.0008	0.0014	0.0015	0.001	0.0008	0.0008	0.0009	0.0018	0.0006	0.0029	0.0015	0.0012	0.0008	0.0008	0.0024	0.0016	0.0013	0.0009
	COII	0.0008	0.001	0.0004	0.0009	0.002	0.0012	0.0004	0.0009	0.0006	0.0009	0.001	0.0009	0.001	0.0007	0.0006	0.0019	0.0018	0.001	0.0005
	COI	0.0005	0.0005	0.002	0.0015	0.0007	0.0008	0.0012	0.0008	0.0012	0.0004	0.0007	0.001	0.0004	0.0003	0.0004	0.0027	0.0008	0.0003	0.0008
	CYTB	0.0008	0.0005	0.0009	0.0006	0.0006	0.0005	0.001	0.001	0.0016	0.0005	0.001	0.0004	0.0008	0.0005	0.0017	0.0036	0.001	0.0011	0.0006
	ND1	0.0007	0.0009	0.001	0.0011	0.0005	0.001	0.0011	0.0014	0.0006	0.0001	0.0004	0.0014	0.0009	0.001	0.0012	0.0014	0.0013	0.001	0.0011
	ND2	0.0005	0.001	0.0014	0.0019	0.0018	0.0011	0.001	0.0002	0.0013	0.0011	0.0005	0.0006	0.0006	0.0004	0.0003	0.0018	0.0008	0.0001	0.0006
	ND3	0.0017	0.0013	0.0026	0.0016	0.0005	0.0007	0.0011	0.0023	0.0014	0.0007	0.0011	0.001	0.0015	0.001	0.0011	0.0013	0.0011	0.0008	0.0007
	ND4L	0.0024	0.0029	0.0009	0.0015	0.0006	0.0014	0.0008	0.001	0.0013	0.0019	0.0033	0.0007	0.0009	0.0022	0.0022	0.0027	0.0016	0.0014	0.0018
	ND4	0.0011	0.0013	0.0008	0.0013	0.0005	0.0003	0.0003	0.0005	0.0008	0.001	0.0005	0.0009	0.0013	0.0004	0.0009	0.0023	0.0007	0.0054	0.0006
	ND5	0.0016	0.0012	0.0005	0.0007	0.0006	0.0002	0.0007	0.001	0.001	0.0014	0.0012	0.0009	0.0009	0.0009	0.0004	0.0038	0.0012	0.0002	0.0009
	ND6	0.0019	0.0009	0.0015	0.0012	0.0007	0.001	0.0014	0.0007	0.0008	0.0017	0.0014	0.0008	0.0034	0.0018	0.0026	0.0032	0.0009	0.0014	0.0011

Note: Taxa abbreviations: yak (*D. yakuba*), ere (*D. erecta*), mau (*D. mauritiana*), ele (*D. elegans*), mel (*D. melanogaster*), rho (*D. rhopaloa*), sim (*D. simulans*), sec (*D. sechellia*), per (*D. persimilis*), pse (*D. pseudoobscura*), vir (*D. virilis*), wil (*D. willistoni*), inc (*D. incompta*), ana (*D. ananassae*), moj (*D. mojavensis*), gri (*D. grimshawii*), bia (*D. biarmipes*), lit (*D. littoralis*), musca (*Musca domestica*).

Table 2S: RCFV values for 25 nuclear loci in 19 *Drosophila* species

	RCFV values																		
	Species																		
	pse	*sec*	*mir*	*sim*	*inc*	*musca*	*per*	*bia*	*rho*	*gri*	*yak*	*ana*	*ere*	*vir*	*eug*	*moj*	*ele*	*wil*	*mel*
All loci	0.0008	0.003	0.0015	0.0034	0.0031	0.0099	0.0006	0.004	0.0019	0.0028	0.0027	0.0021	0.0029	0.0031	0.0009	0.0019	0.0037	0.0069	0.0025
AP1GAMMA	0.0008	0.004	0.0009	0.0042	0.0023	0.0151	0.0007	0.0047	0.0033	0.002	0.0036	0.0032	0.0039	0.0017	0.0011	0.0014	0.004	0.01	0.0031
ART7	0.0005	0.0046	0.0001	0.0043	0.0011	0.0135	0.0006	0.0042	0.0009	0.0012	0.003	0.0016	0.0029	0.001	0.0038	0.0002	0.0021	0.0085	0.0039
CAP60A	0.0012	0.0045	0.0007	0.0058	0.011	0.0062	0.0012	0.0045	0.0042	0.0039	0.0039	0.0026	0.0043	0.0097	0.0014	0.0056	0.0069	0.0038	0.0042
CG14435	0.001	0.0027	0.0075	0.003	0.0048	0.0127	0.0012	0.0051	0.0031	0.0061	0.0022	0.0038	0.003	0.0049	0.0012	0.0052	0.0043	0.0057	0.0016
CG2812	0.0034	0.0081	0.0031	0.0085	0.0094	0.0095	0.0034	0.0068	0.0022	0.0041	0.0091	0.0008	0.0083	0.009	0.0005	0.0061	0.006	0.0107	0.0084
CG3209	0.0019	0.0054	0.0052	0.0059	0.0073	0.006	0.0018	0.0038	0.0012	0.0046	0.0053	0.0011	0.0054	0.0077	0.0011	0.0039	0.0028	0.0072	0.0056
CG42265	0.0034	0.0006	0.0042	0.0009	0.0025	0.0105	0.0035	0.0026	0.0014	0.0029	0.0005	0.0055	0.0017	0.0015	0.0033	0.0045	0.0015	0.0056	0.0016
CG4386	0.0059	0.0021	0.0053	0.0018	0.0006	0.0061	0.0062	0.0018	0.0018	0.0003	0.0022	0.0044	0.0011	0.0011	0.004	0.0062	0.0016	0.0076	0.0021
CG4585	0.0066	0.0091	0.0039	0.0091	0.011	0.0018	0.0042	0.007	0.0046	0.0086	0.0073	0.0016	0.0075	0.0089	0.0008	0.0091	0.0082	0.0082	0.0083
CG4797	0.0009	0.0049	0.0007	0.0057	0.0014	0.0111	0.0005	0.0013	0.0025	0.004	0.0043	0.0025	0.0047	0.0024	0.0028	0.0027	0.0019	0.0042	0.0039
CG7033	0.0009	0.0033	0.0035	0.0037	0.0026	0.0113	0.0008	0.0069	0.0047	0.0059	0.0043	0.0035	0.0037	0.0036	0.0012	0.0036	0.0058	0.0081	0.0032
G1CT1	0.0008	0.0025	0.0007	0.0029	0.0017	0.011	0.0006	0.0079	0.0039	0.0062	0.0024	0.0029	0.0029	0.0013	0.0018	0.002	0.0054	0.0122	0.0024
GMER	0.0009	0.0036	0.0024	0.0037	0.0007	0.0114	0.0005	0.0041	0.0008	0.0036	0.0019	0.0009	0.0033	0.0007	0.0026	0.003	0.0028	0.0057	0.0037
GClC	0.0045	0.0009	0.003	0.0014	0.0029	0.0113	0.0044	0.0061	0.0023	0.0025	0.0014	0.0055	0.001	0.0029	0.0043	0.0012	0.0022	0.0086	0.0011
MP20	0.0011	0.0035	0.0006	0.0021	0.001	0.0066	0.0015	0.0036	0.0015	0.0036	0.0033	0.0023	0.0035	0.0034	0.0006	0.0021	0.0027	0.0064	0.0016
MCM3	0.0035	0.0023	0.0035	0.0028	0.0026	0.0131	0.0033	0.0065	0.001	0.0031	0.0027	0.0036	0.0019	0.0026	0.0061	0.0008	0.0028	0.0056	0.0014
NAP1	0.0037	0.0073	0.0035	0.0075	0.0082	0.0091	0.0035	0.0067	0.0019	0.0036	0.0076	0.0024	0.0066	0.0079	0.0028	0.0073	0.0054	0.0046	0.007
PRL1RA	0.001	0.0007	0.0046	0.0007	0.0059	0.0022	0.0021	0.0021	0.0012	0.0013	0.0013	0.0029	0.0006	0.0013	0.0016	0.0015	0.0004	0.0022	0.0009
RAR35	0.0015	0.0054	0.0041	0.0054	0.0065	0.0152	0.0013	0.0065	0.0044	0.0037	0.0056	0.005	0.0059	0.0042	0.002	0.0045	0.0045	0.0114	0.0035
TER94	0.002	0.0008	0.0028	0.001	0.002	0.0136	0.0023	0.0038	0.0036	0.001	0.001	0.0029	0.0016	0.0017	0.0007	0.0043	0.0058	0.0066	0.001
VPS20	0.0029	0.0041	0.0038	0.004	0.0048	0.0126	0.0026	0.0052	0.0022	0.004	0.0033	0.0011	0.0035	0.0022	0.0023	0.0057	0.003	0.0076	0.0046
VPS35	0.0026	0.0009	0.0015	0.0008	0.0012	0.0111	0.0028	0.0024	0.0026	0.0011	0.0006	0.0011	0.0009	0.0011	0.001	0.0015	0.0049	0.0064	0.0012
ADAR	0.0014	0.0018	0.002	0.0016	0.0019	0.0035	0.0016	0.0005	0.0018	0.0009	0.0016	0.0043	0.0006	0.0024	0.0023	0.0018	0.0021	0.0014	0.0023
CLN3	0.0043	0.0021	0.0044	0.0017	0.001	0.0042	0.0045	0.0033	0.0004	0.0022	0.0013	0.0009	0.0006	0.001	0.0031	0.0033	0.0012	0.0059	0.0011
WRAPPER	0.0043	0.0011	0.006	0.0017	0.0011	0.0091	0.0038	0.0023	0.0038	0.0029	0.0015	0.0008	0.0018	0.0013	0.0015	0.0011	0.0009	0.0104	0.0008

Nuclear loci

Note: Taxa abbreviations: yak (*D. yakuba*), ere (*D. erecta*), mau (*D. mauritiana*), ele (*D. elegans*), mel (*D. melanogaster*), rho (*D. rhopaloa*), sim (*D. simulans*), sec (*D. sechellia*), per (*D. persimilis*), pse (*D. pseudoobscura*), vir (*D. virilis*), wil (*D. willistoni*), inc (*D. incompta*), ana (*D. ananassae*), moj (*D. mojavensis*), gri (*D. grimshawii*), bia (*D. biarmipes*), lit (*D. littoralis*), musca (*Musca domestica*).

6. References

AKAIKE, H. A new look at the statistical model identiWcation. IEEE Trans. Automat. Contr. 19, 716–723, 1974.

ANÉ, CÉCIL et al. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution, v. 24, n. 2, p. 412-426, 2007.

ANÉ, C. et al. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24 (2), 412-426. Abstract and Erratum, 2007.

AUBERT, J. SOLIGNAC, M. Experimental evidence for mitochondrial DNA introgression between *Drosophila* species. Evolution, p. 1272-1282, 1990.

BÄCHLI, G. TaxoDros: The Database on Taxonomy of Drosophilidae, v. 1.03, Database 2009/04. http://taxodros.unizh.ch/. Last accessed on 01/03/2016.

BACHTROG, DORIS et al. Extensive introgression of mitochondrial DNA relative to nuclear genes in the Drosophila yakuba species group. Evolution, v. 60, n. 2, p. 292-302, 2006.

BERGSTEN, J. A review of long-branch attraction. Cladistics, v. 21, n. 2, p. 163-193, 2005.

BERNT, M. et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol, 69 (2), 313-319, 2013.

BRENNAN, I. G. et al. Mitochondrial introgression via ancient hybridization, and systematics of the Australian endemic pygopodid gecko genus Delma. Molecular phylogenetics and evolution, v. 94, p. 577-590, 2016.

BRNCIC, D. Ecological and cytogenetic studies of *Drosophila flavopilosa*, a Neotropical species living in *Cestrum* flowers**.** Evolution; 20:16-29, 1966.

BROWN, W. M. et al. Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences, v. 76, n. 4, p. 1967-1971, 1979.

CASTRESANA, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17, 540-552, 2000.

CLARK, A. G. et al. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–218, 2007.

CLARY, D. O., WOLSTENHOLME, D. R. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. Journal of Molecular Evolution, v. 22, n. 3, p. 252-271, 1985.

CHAN, K. M. A, LEVIN, S. A. Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. Evolution, v. 59, n. 4, p. 720-729, 2005.

CHEVREUX, B., WETTER, T., SUHAI, S. Genome sequence assembly using trace signals and additional sequence information. In: German Conference on Bioinformatics. p. 45-56, 1999.

DE RÉ, F. C. et al. Characterization of the complete mitochondrial genome of flower-breeding *Drosophila incompta* (Diptera, Drosophilidae). Genetica, v. 142, n. 6, p. 525-535, 2014.

DELSUC, F., BRINKMANN, H., PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics, v. 6, n. 5, p. 361-375, 2005.

DE RÉ, F. C., LORETO, E. L. S., ROBE, L. J. Gene and species trees reveal mitochondrial and nuclear discordance in the *Drosophila cardini* group (Diptera: Drosophilidae). Invertebrate biology, v. 129, n. 4, p. 353-367, 2010.

DUNN, C. W. et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature, v. 452, n. 7188, p. 745-749, 2008.

GALTIER, N., GOUY, M. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci, 92:11317-11321, 1995.

GEE, H. Evolution: ending incongruence. Nature, v. 425, n. 6960, p. 782-782, 2003.
HAHN, C., BACHMANN, L., CHEVREUX, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Research, v. 41, n. 13, p. e129-e129, 2013.

HOFMANN PRP. Variabilidade genética em espécies de nível ecológico restrito. Ciênc.Cult. 37:579-581, 1985.

JEFFROY, O. et al. Phylogenomics: the beginning of incongruence? TRENDS in Genetics, v. 22, n. 4, p. 225-231, 2006.

JETZ, W. et al. The global diversity of birds in space and time. Nature, v. 491, n. 7424, p. 444-448, 2012.

KECK, B. P., NEAR, T. J. Geographic and temporal aspects of mitochondrial replacement in Nothonotus darters (Teleostei: Percidae: Etheostomatinae). Evolution, v. 64, n. 5, p. 1410-1428, 2010.

KUCK P., MEUSEMANN K. FASconCAT, Version 1.0, Zool. Forschungsmuseum A. Koenig, Germany, 2010.

LEWIS, R. L., BECKENBACH, A.T., MOOERS, A. O. The phylogeny of the sobgroups within the *melanogaster* species group: Likelihood tests on *COI* and *COII* sequences and a Bayesian estimate of phylogeny. Mol Phylogenet Evol 37:15-24, 2005.

MADDISON, W. P. Gene trees in species trees. Systematic biology, v. 46, n. 3, p. 523-536, 1997.

NYLANDER, J. A. A. MrModeltest v2. Program Distributed by the Author. Evolutionary Biology Centre, Uppsala University, 2004.

O'GRADY, P.M., KIDWELL, M.G. Phylogeny of the subgenus *Sophophora* (Diptera: Drosophilidae) based on combined analysis of nuclear and mitochondrial sequences. Mol. Phylogenet. Evol. 22 (3), 443–453, 2002.

OKONECHNIKOV K., OLGA G., MIKHAIL F. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28.8: 1166-1167, 2012.

PAMILO, P., NEI, M. Relationships between gene trees and species trees. Molecular biology and evolution, v. 5, n. 5, p. 568-583, 1988.

POLLARD, D. A. et al. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet, v. 2, n. 10, p. e173, 2006.

POWELL, J. R. Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from Drosophila. Proceedings of the National Academy of Sciences, v. 80, n. 2, p. 492-495, 1983.

RAMBAUT A. FigTree version 1.3.1 [computer program] http://tree.bio.ed.ac.uk. 2009.

ROBE, L.J. et al. Molecular phylogeny of the subgenus *Drosophila* (Diptera, Drosophilidae) with an emphasis on Neotropical species and groups: a nuclear versus mitochondrial gene approach. Molecular Phylogenetics and Evolution. 36:623-640. 2005.

ROBE, L.J., LORETO, E.L.S., VALENTE, V.L.S. Radiation of the *Drosophila* subgenus (Drosophilidae, Diptera) in the Neotropics. Journal of Zoological Systematics and Evolutionary Research. DOI 10.1007/s10709-009-9432-5, 2010a.

ROBE, L.J.; VALENTE, V.L.S.; LORETO, E.L.S. Phylogenetic relationships and macroevolutonary patterns within the *Drosophila tripunctata* "radiation" (Diptera: Drosophilidae). Genetica, v.138, p.725-735, 2010b.

RODRÍGUEZ-TRELLES, F., TARRÍO, R., AYALA, F. J. Evidence for a high ancestral GC content in Drosophila. Molecular Biology and Evolution, v. 17, n. 11, p. 1710-1717, 2000.

ROKAS, A. et al. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature, v. 425, n. 6960, p. 798-804, 2003.

RONQUIST, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology, v. 61, n. 3, p. 539-542, 2012.

SCHAWAROCH, V. Phylogeny of a paradigm lineage: the Drosophila melanogaster species group (Diptera: Drosophilidae). Biological Journal of the Linnean Society, v. 76, n. 1, p. 21-37, 2002.

SCHIERUP, M. H., HEIN, J. Consequences of recombination on traditional phylogenetic analysis. Genetics, v. 156, n. 2, p. 879-891, 2000.

SILVA-BERNARDI, et al. Phylogenetic relationships in the Drosophila fasciola species subgroup (Diptera, Drosophilidae) inferred from partial sequences of the mitochondrial cytochrome oxidase subunit I (COI) gene. Genetics and Molecular Biology, v. 29, n. 3, p. 566-571, 2006.

STRUCK, T. H. TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. Evolutionary Bioinformatics Online, 10, 51–67, 2014.

SWOFFORD, D.L. PAUP: Phylogenetic Analysis using Parsimony (and other methods). Version 4. Sinauer Associates, Massachusetts, 2003.

TAMURA, K. et al. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution, 2011.

TATARENKOV A., AYALA F. J. Phylogenetic relationships among species groups of the *virilis-repleta* radiation of *Drosophila.* Mol Phylogenet Evol 21:327-331, 2001.

THROCKMORTON, L. H. The phylogeny, ecology and geography of *Drosophila*. In: King, R. C. (ed) Handbook of Genetics. Plenun, New York, pp 421-469. 1975.

TOEWS, D. P. L., BRELSFORD, A. The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, v. 21, n. 16, p. 3907-3930, 2012.

TOURASSE, N. J., W-H. LI. Performance of the relative-rate test under non stationary models of nucleotide substitution. Mol. Biol. Evol. 16:1068–1078, 1999.

ZACHOS, F. E. Gene trees and species trees–mutual influences and interdependences of population genetics and systematics. Journal of Zoological Systematics and Evolutionary Research, v. 47, n. 3, p. 209-218, 2009.

ZHENG, Yuchi; WIENS, JOHN J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Molecular phylogenetics and evolution, v. 94, p. 537-547, 2016.

ZHONG, M. et al. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis for terebelliform annelids. BMC Evol. Biol. 11, 369, 2011.

WEIGERT, A. et al. Illuminating the base of the annelid tree using transcriptomics. Molecular Biology and Evolution, p. msu080, 2014.

WEISROCK, D. W. Concordance analysis in mitogenomic phylogenetics. Molecular phylogenetics and evolution, v. 65, n. 1, p. 194-202, 2012.

WONG, A. et al. Phylogenetic incongruence in the *Drosophila melanogaster* species group. Molecular phylogenetics and evolution, v. 43, n. 3, p. 1138-1150, 2007.

4 ARTIGO 3 - REPERTOIRE OF OLFACTORY AND GUSTATORY RECEPTOR GENES IN *Drosophila incompta*, A HIGHLY SPECIALIZED DROSOPHILIDAE SPECIES

Repertoire of olfactory and gustatory receptor genes in *Drosophila incompta*, a highly specialized Drosophilidae species

De Ré, F.C.,[1] Robe, L.J.,[1,2] Wallau, G.L.,[1,3] Loreto, E.L.S.[1,4*]

1. Programa de Pós Graduação em Biodiversidade Animal, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil.
2. Programa de Pós-Graduação em Biologia de Ambientes Aquáticos Continentais Universidade Federal do Rio Grande, Rio Grande, Rio Grande do Sul, Brazil.
3. Departamento de Entomologia, Centro de Pesquisas Aggeu Magalhães - FIOCRUZ-CPqAM, Recife, PE, Brazil.
4. Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil.

*Corresponding author: Elgion L. S. Loreto (elgion@base.ufsm.br)

Keywords: Duplication; Gene family evolution; gene loss; pseudogenization.

Abstract

Drosophila incompta is a highly specialized *Drosophila* species, which relies on chemosensation to detect nutrient-rich foods, select mates and deposit eggs in hospitable zones, all activities performed on flowers of *Cestrum* (Solanaceae). The olfactory and gustatory receptors (OR and GR, respectively) help in this process by transferring signals from the external environment to the nervous system. Here, we analyzed the repertoire of gustatory and olfactory receptor genes of *D. incompta* and characterized the predominant type of evolutionary force shaping their divergence patterns in this species. For this, the whole genome of *D. incompta* was sequenced and genes for OR and GR were identified and assembled using orthologous protein sequences from *D. melanogaster* as seed. Our results suggest that *D. incompta* has only 28 and 12 OR and GR genes, respectively, setting this species as the *Drosophila* species with the lower number of genes as yet described for both gene families, which may reflect an adaptive response to niche specialization. Nevertheless, the low average ratio between the number of nonsynonymous substitutions per nonsynonymous site and the number of synonymous substitutions per synonymous site presented in the comparisons involving *D. incompta* and each of the other 12 *Drosophila* species for which the complete genome was previously characterized suggest each of these chemosensory genes are mainly under the influence of purifying selection.

1. Introduction

Insects use chemical signals to identify host plants, which suggests that chemosensory perception could be a target of natural selection during host specialization (Whiteman and Pierce, 2008). In this sense, along the coevolutionary process, compounds that once served to deter attacks by plant enemies are often co-opted as feeding or oviposition stimulants by specialists that have evolved the ability to detoxify them (Wheat et al. 2007). The olfactive

receptors (ORs) and gustative receptors (GRs) can assist in this process, producing internal responses of stimuli from the external environment. These gene families are well characterized in *Drosophila melanogaster* (Clyne et al. 1999; Gao e Chess, 1999; Vosshall et al. 1999; Scott et al. 2001), which presents 60 genes in each of the GR and OR classes, encoding 62 and 68 proteins, respectively (Robertson et al. 2003). However, due to the high variability in behavioral and ecological strategies presented by *Drosophila*, the characterization of the genetical factors linked to chemoreception in different species could certainly provide a deeper understanding of the role played by these genes along the evolution of each species.

In fact, recent studies involving drosophilids have revealed interesting patterns of molecular evolution for different genes associated with niche specialization, like chemoreception, body pigment, metabolism and detoxification genes, both in regard to the number and location of genetic factors, and also to the rate of non-synonymous to synonymous substitutions (Gilbert et al. 2007; Gardiner et al. 2008; McBride, 2007; Matute et al. 2009, Wittkopp et al. 2009, Wittkopp et al. 2010, Matzkin et al. 2006). For chemosensory gene families, for example, specialist species have shown a fivefold faster rate of gene loss in GR (McBride and Arguello, 2007) and a higher dn/ds ratios for OR and GR genes. Nevertheless, according to Gardiner et al. (2008), endemism rather than niche specialization may account for much of this large scale chemosensory gene loss.

Drosophila incompta is a species of the *Drosophila flavopilosa* group (Bächli, 2016) that uses flowers of *Cestrum* (Solanaceae) as unique sites for oviposition, larval development, and feeding (Brncic, 1966; Hofmann, 1985). This is one of the most widely distributed species of the *flavopilosa* group, occurring from Mexico to the South of Brazil and Argentina (Bächli, 2016), and showing a larger niche breadth in comparison to other species of the group (Robe et al., 2013). The restricted ecological patterns, the reduced levels of endemism and the host

plant preference makes *D. incompta* a particularly interesting species to study the GR and OR gene families, especially in comparison with other generalist species of *Drosophila*, like *D. ananassae*, *D. melanogaster* and *D. simulans*, or other specialist species which present different levels of endemism, like *D. grimshawi* and *D. sechellia*. So, this study aims to identify the number of GR and OR genes in *D. incompta*, while analyzing the type of selection that may be acting on these.

2. Material and Methods

2.1. Sequencing and genes recovery

We performed a comparative genomic analysis of genes known to be involved in olfactory and taste behavior in *Drosophila*. For this, the entire genome of *D. incompta* was sequenced by the Fasteris DNA Sequencing Service with a Solexa-Illumina HiSeq 2000 Next Generation Sequencing (NGS) device. A single-end approach with a read-size of approximately 100 bp was employed. Once the draft genome of *D. incompta* was available, the 60 gustatory and olfactory receptors gene sequences of *D. melanogaster* Robertson et al. (2003) were downloaded from FlyBase (Attrill et al. 2015) and submitted to Blastx tool (NCBI website) in order to recover *D. melanogaster* protein sequences. These sequences were arranged in two separate matrices, one containing OR and the other with GR protein sequences. Each of these files was then used as a seed in the search and assembly of orthologous gene sequences in the reads of *D. incompta* genome, through the use of aTRAM software (Allen et al. 2015) with default parameters, except for the commands -complete (which tells aTRAM to stop assembling after the seed is entire covered) and -protein (for the use of proteins as seed) settings.

2.2. dn/ds ratio estimates

After assembly of the olfactory and gustative receptors of *D. incompta,* all nucleotide sequences had their identity confirmed by BLAST (NCBI website), and were used in the search for orthologous sequences in FlyBase (Attrill et al. 2015), considering the 12 genomes previously sequenced (Clark et al. 2007): *D. simulans*, *D. erecta, D. sechellia, D. yakuba, D. pseudoobscura, D. perimilis, D. willistoni, D. virilis, D. mojavensis, D. grimshawii, D. ananassae* and *D. melanogaster* (Table 1). The coding sequences obtained in this search were aligned individually for each gene using the algorithm Muscle (Edgar, 2004), with predicted amino acid sequences as templates, as implemented in Mega 5.0 software (Tamura et al. 2011). Extents of synonymous and non-synonymous substitutions were determined for all pairwise comparisons among the 13 species using SNAP (Synonymous Non-synonymous Analysis Program), as available on the HIV database website (http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html) (Korber 2000). This program estimates the numbers of synonymous substitutions per synonymous site (ds) and non-synonymous substitutions per non-synonymous site (dn) based on the method of Nei and Gojobori (1986), and incorporating a statistic developed in Ota and Nei (1994). Briefly, statistically significant values of dn/ds < 1, = 1 and > 1 imply the occurrence of purifying selection, neutral evolution and positive selection, respectively.

3. Results and discussion

3.1. GR and OR repertoires

The availability of complete genome sequences for 12 *Drosophila* species, enabled us to examine the evolutionary diversification of gustatory (GR) and olfactory (OR) receptors genes that mediate insect-plant interactions within Drosophilidae, and the molecular patterns related to the restricted ecology of *D. incompta*. In this sense, using *D. melanogaster* protein sequences as reference for search and gene assembly within *D. incompta* draft genome, we were able to recover only 28 odorant receptors (OR) and 12 gustatory receptors (GR) for this

species (Table 1). The recovered fragments in this search vary in size between 103bp (OR67B) and 2214bp (OR46A) and between 189bp (GR61A) and 1098bp (GR98A) for the OR and GR gene, respectively (Table 1), and all the set of chemoreceptor genes found in *D. incompta* present orthologous copies in the genomes of the other 12 *Drosophila* species (see materials and methods for details). Moreover, *D. incompta* seems to present a duplication for GR98A. Although further studies are necessary to confirm this pattern, duplication of GR or OR genes is not a rare phenomenon within Drosophilidae (Gardiner et al. 2008).

Previous comparisons of the OR and GR families between distantly related insects have uncovered dramatic changes in gene family size and content, fueling the suspicion that these genes evolve rapidly (Hill et al 2002; Robertson et al. 2003; Robertson and Wanner, 2006). In *Anopheles gambiae*, for example, there are a total of 79 and 76 ORs and GRs, respectively (Hill et al. 2002), whereas *Bombyx mori* has 48 ORs (Wanner et al. 2007), and *Linepithema humile* reaches 367 genes for this complex (Smith et al. 2011a). The repertoire of ORs genes also seems to have expanded in *Apis mellifera*, which possesses 170 olfactory loci, putatively related to the remarkable olfactory abilities presented by this species, including perception of several pheromone blends, kin recognition signals, and diverse floral odors. Despite this, the honeybee genome has only 10 GR genes (Robertson and Wanner, 2006), a number similar to that presented by *D. incompta*.

Nevertheless, even at shallower phylogenetic scales the complete repertories of chemoreceptor genes can vary, and this is a remarkable pattern within *Drosophila*. In fact, until this study, the GR family size in *Drosophila* ranged from 52 (in *D. mojavensis*) to 83 genes (in *D. grimshawi*), whereas the OR family size ranged from 60 (in *D. erecta* and *D. sechellia*) up to 83 genes (in *D. grimshawi*) (Gardiner et al. 2008) (Table 2). This study presents *D. incompta* as the *Drosophila* species with the lower number of genes as yet described for both, the OR and the GR families, which may reflect an adaptive response to niche specialization. Such a pattern was

previously presented by McBride and Arguello (2007), which showed that specialization on novel host plants along both the *D. sechellia* and *D. erecta* lineages coincides with a dramatic contraction on the Gr family. Furthermore, according to McBride (2007), the specialist *D. sechellia* has a high fraction of Ors and Grs exhibiting lack-of-function (LOF) mutations that clearly render them pseudogenes, resulting in a rate of gene loss 9–10 times higher than that of its generalist sister species, *D. simulans*. Nevertheless, Gardiner et al. (2008) presented contradictions to this pattern and found that besides specialization, endemism may also account for the contractions in the size of chemoreceptor gene families.

Although the reduced number of ORs and GRs detected here for *D. incompta* could be explained by high rates of gene loss in face of the reduced need to detect and respond to as a wide variety of compounds as generalist taxa, the results could be hardly conciliated with the levels of endemism presented by this species. Moreover, we cannot rule out the putative effect of artifacts related to the analysis, which may have prevented the detection of too divergent genes and pseudogenes. Additionally, as for several genes only a fragment of the coding region was recovered, there is also the possibility that some of the detected genes will confirm to be pseudogenes in the future. Therefore, additional analyzes are necessary to better assess this scenario of gene losses in *D. incompta*, since this study only refers to the number of genes found in the genome with the methodology applied here.

3.2. Selection analysis

In order to investigate the patterns of molecular evolution of OR and GR genes in *D. incompta*, we estimated the average ratio between the number of non-synonymous substitutions per non-synonymous site (dn) and the number of synonymous substitutions per synonymous site (ds) in the comparisons involving *D. incompta* and each of the other 12 included *Drosophila* species. Our results revealed that all 28 olfactive and the 12 gustative receptors included in the analysis are under purifying selection (ie, $dn/ds < 1$) (Table 1).

These results are in agreement with those previously published by Vieira, Sánchez-Gracia and Rozas (2007) and Lavagnino et al. (2012), which detected dn/ds ratios compatible either with purifying selection or with neutral evolution for different chemosensory genes. Nevertheless, Gardiner et al. (2008) found that few loci exhibit statistically significant evidence of positive selection. So, in general, the evolutionary forces acting on the chemosensory gene repertoire of *Drosophila* remain controversial, and the signals may vary between genes and between species. In the case of *D. incompta*, only more robust analyses can clearly elucidate the role played by random genetic drift, positive or purifying selection in the reduced set of OR and GR genes.

Table 1: Repertoire of GR and OR genes in *D. incompta*, with the recovered fragment size, the average of dn, ds and dn/ds ratio and the Flybase identification numbers for *D. melanogaster* orthologous sequence

Gene	Fragment size assembled for *D. incompta* (bp)	Averages of all pairwise comparisons (dn)	Averages of all pairwise comparisons (ds)	Average ratio dn/ds	FlyBase ID from *D. melanogaster*
GR2A	992	0.2259	1.1657	0.19378914	FBgn0265139
GR21A	231	0.0394	0.9735	0.040472522	FBgn0041250
GR28A	1040	0.0778	1.4683	0.052986447	FBgn0041247
GR28B	212	0.2717	1.7038	0.159467074	FBgn0045495
GR32A	277	0.1424	1.7362	0.082018201	FBgn0041246
GR39B	269	0.1946	1.2933	0.150467796	FBgn0041245
GR43A	492	0.1135	1.7207	0.065961527	FBgn0041243
GR57A	410	0.2679	1.7177	0.155964371	FBgn0041240
GR61A	189	0.2009	1.6398	0.122514941	FBgn0035167
GR63A	253	0.0536	1.6231	0.033023227	FBgn0035468
GR66A	206	0.0704	1.5587	0.045165843	FBgn0035870
*GR98A**	1098	0.3425	1.7998	0.190298922	FBgn0039520
OR9A	650	0.2412	1.1454	0.210581456	FBgn0030204
OR10A	310	0.2642	1.5998	0.165145643	FBgn0030298
OR22C	373	0.1597	1.2903	0.123769666	FBgn0026396
OR23A	215	0.3704	1.4303	0.25896665	FBgn0026395
OR30A	243	0.0835	1.8965	0.044028474	FBgn0032096
OR33C	788	0.3023	1.6046	0.188395862	FBgn0026390
OR35A	298	0.1771	0.9919	0.178546224	FBgn0028946
OR42A	204	0.1774	1.6382	0.108289586	FBgn0033041
OR43A	290	0.1872	1.1706	0.159917991	FBgn0026389
OR46A	2214	0.3613	1.8503	0.195265633	FBgn0026388
OR49A	353	0.231	1.662	0.13898917	FBgn0033727
OR49B	152	0.1359	1.5225	0.089261084	FBgn0028963
OR56A	279	0.1579	2.0238	0.078021544	FBgn0034473
OR59A	698	0.1961	1.7493	0.112101984	FBgn0026384
OR63A	171	0.1944	1.1993	0.162094555	FBgn0035382
OR67B	103	0.1073	1.5889	0.067530996	FBgn0036019
OR67C	249	0.0981	1.3456	0.072904281	FBgn0036078
OR67D	501	0.2957	1.8983	0.155770953	FBgn0036080
OR69A	785	0.2667	1.728	0.154340278	FBgn0041622
OR74A	304	0.1924	1.6482	0.116733406	FBgn0036709
OR82A	203	0.1917	1.8622	0.102942756	FBgn0041621
OR83A	535	0.0982	1.3618	0.072110442	FBgn0037322
OR83B	745	0.0503	1.3091	0.038423344	FBgn0037324
OR83C	457	0.3159	1.9328	0.163441639	FBgn0037399
OR85E	346	0.2078	1.4639	0.141949587	FBgn0026399
OR88A	744	0.1883	1.4913	0.126265674	FBgn0038203
OR94A	284	0.1245	1.4401	0.08645233	FBgn0039033
OR98A	328	0.2199	1.7543	0.125349142	FBgn0039551

The "*" indicates that there is evidence of gene duplication in *D. incompta.*

Table 2: Number of gustatory (GR) and olfactory (OR) receptors in 12 *Drosophila* species.

Species	Number of GR genes	Number of OR genes
D. melanogaster	60	61
D. simulans	65	61
D. sechellia	65	60
D. yakuba	63	62
D. erecta	58	60
D. ananassae	71	71
D. pseudoobscura	55	71
D. persimilis	55	69
D. willistoni	73	80
D. mojavensis	52	63
D. virilis	53	64
D. grimshawi	83	83

Note: This table was modified from Gardiner et al. (2008).

4. References

ALLEN, Julie M. et al. aTRAM-automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. BMC bioinformatics, v. 16, n. 1, p. 1, 2015.

ATTRILL, Helen et al. FlyBase: establishing a Gene Group resource for Drosophila melanogaster. Nucleic acids research, p. gkv1046, 2015.

BÄCHLI, G. TaxoDros: The Database on Taxonomy of Drosophilidae, v. 1.03, Database 2009/04. http://taxodros.unizh.ch/. Last accessed on 01/03/2016.

BRNCIC, D. Ecological and cytogenetic studies of *Drosophila flavopilosa*, a Neotropical species living in *Cestrum* flowers. Evolution; 20:16-29, 1966.

CLARK, A. G. et al. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–218, 2007.

CLYNE, Peter J. et al. A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in Drosophila. Neuron, v. 22, n. 2, p. 327-338, 1999.

EDGAR, ROBERT C. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32 (5), 17-97, 2004.

GAO, Qian; CHESS, Andrew. Identification of candidate Drosophila olfactory receptors from genomic DNA sequence. Genomics, v. 60, n. 1, p. 31-39, 1999.

GARDINER, Anastasia et al. Drosophila chemoreceptor gene evolution: selection, specialization and genome size. Molecular Ecology, v. 17, n. 7, p. 1648-1657, 2008.

GILBERT, J. M., PERONNET, F., SCHLOTTERER, C. Phenotypic plasticity in *Drosophila* pigmentation caused by temperature sensitivity of a chromatin regulator network. PLoS Genetics 3: e 30, 2007.

HILL, Catherine A. et al. G protein-coupled receptors in Anopheles gambiae.Science, v. 298, n. 5591, p. 176-178, 2002.

HOFMANN, P. R. P. Variabilidade genética em espécies de nível ecológico restrito. Ciência e Cultura. 37:579-581, 1985.

KORBER B: HIV signature and sequence variation analysis. In Computational and evolutionary analysis of HIV molecular sequences Volume 4. Edited by: Rodrigo AG, Learn GH. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000:55-72.

LAVAGNINO, Nicolás et al. Evolutionary genomics of genes involved in olfactory behavior in the Drosophila melanogaster species group. Evolutionary Bioinformatics, v. 8, p. 89, 2012.

MCBRIDE, Carolyn S.; ARGUELLO, J. Roman. Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics, v. 177, n. 3, p. 1395-1416, 2007.

MATZKIN, L. M. et al. Functional genomics of cactus host shifts in Drosophila mojavensis. Molecular Ecology, v. 15, n. 14, p. 4635-4643, 2006.

MATUTE, D. M., et al. Temperature-based extrinsic reproductive isolation in two species of *Drosophila.* Evolution: 63:595–612, 2009.

MCBRIDE, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proceedings of the National Academy of Sciences of the United States of America; 104 (12):4996-5001, 2007.

ROBE L. J., et al. The *Drosophila flavopilosa* species group (Diptera, Drosophilidae): An Array of exciting questions. Fly, 7:59 – 69, 2013.

ROBERTSON, Hugh M.; WARR, Coral G.; CARLSON, John R. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proceedings of the National Academy of Sciences, v. 100, n. suppl 2, p. 14537-14542, 2003.

ROBERTSON, Hugh M.; WANNER, Kevin W. The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family. Genome research, v. 16, n. 11, p. 1395-1403, 2006.

SCOTT, Kristin et al. A chemosensory gene family encoding candidate gustatory and olfactory receptors in Drosophila. Cell, v. 104, n. 5, p. 661-673, 2001.

SMITH, C.D. et al. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). Proceedings of the National Academy of Sciences of the United States of America 108: 5673-5678, 2011a.

TAMURA, Koichiro et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution, v. 28, n. 10, p. 2731-2739, 2011.

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

VOSSHALL, Leslie B. et al. A spatial map of olfactory receptor expression in the Drosophila antenna. Cell, v. 96, n. 5, p. 725-736, 1999.

WANNER, K. W. et al. Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, *Bombyx mori.* Insect molecular biology 16.1: pp. 107-119, 2007.

WHEAT, Christopher W. et al. The genetic basis of a plant–insect coevolutionary key innovation. Proceedings of the National Academy of Sciences, v. 104, n. 51, p. 20427-20431, 2007.

WHITEMAN, Noah K.; PIERCE, Naomi E. Delicious poison: genetics of Drosophila host plant preference. Trends in ecology & evolution, v. 23, n. 9, p. 473-478, 2008.

WITTKOPP, P. J., BELDADE, P. Development and evolution of insect pigmentation: genetic mechanisms and the potential consequences of pleiotropy. Seminars in Cell & Developmental Biology, 20, 65–71, 2009.

WITTKOPP, P. J. et al. Local adaptation for body color in *Drosophila americana.* Heredity: [Epub ahead of print], july 2010.

5 ARTIGO 4 – PHYLOGEOGRAPHIC PATTERNS IN *Drosophila incompta* (DIPTERA, DROSOPHILIDAE)

Phylogeographic patterns in *D. incompta* (Diptera, Drosophilidae), a Neotropical species with restricted ecology

De Ré, F.C.,[1] Loreto, E.L.S.[1,2] Robe, L.J.,[1,3]

7. Programa de Pós Graduação em Biodiversidade Animal, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil.
8. Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil.
9. Programa de Pós-Graduação em Biologia de Ambientes Aquáticos Continentais Universidade Federal do Rio Grande, Rio Grande, Rio Grande do Sul, Brazil.

*Corresponding author: Lizandra Jaqueline Robe (lizbiogen@gmail.com)

Keywords: Populational expansion, *Drosophila flavopilosa* group, specialist species phylogeography, quaternary.

Abstract

Drosophila incompta belongs to the *flavopilosa* group of *Drosophila* and like the other species of this group uses *Cestrum* flowers as unique sites for feeding and breeding. Glacial and interglacial Quaternary periods changed the climate and the vegetation of the Neotropical region, and may have impacted *D. incompta* in direct and indirect ways. After all, both, in both evolutionary and ecological perspectives, the distribution of the flies is associated with the distribution of their host. Here we evaluate the intraspecific diversity and structure patterns presented by different populations of *D. incompta*, while trying to infer the historical demography of this species. For this, our sample includes 123 individuals collected throughout the southern region of Brazil. Phylogeographic analyzes were performed based on the mitochondrial COI (Cytochrome Oxidase Subunit I) and COII (Citochrome Oxidase Subunit II) genes and on the nuclear Hb (Hunchback) gene. In general, the two sets of markers seem to have recovered signals from two different though not mutually exclusive moments of the species history: mitochondrial genes indicated that the species underwent a population expansion event, whereas nuclear genes presented signs that this may have been preceded by a bottleneck. The expansion event was dated to the Quaternary period, and seem to have occurred about 175,000 and 100,000 years before present, although no signs of populational expansion/contraction were detected to the last glacial maximum, where the abiotic conditions were extreme. Moreover, *D. incompta* presents significant levels of genetic diversity but no defined geographical structure, which is compatible with a strategy of recurrent but scattered gene flow along alternate host flourishing seasons as a mean of overcoming periods of scarcity of resources.

1. Introduction

Drosophila incompta belongs to the *flavopilosa* group of *Drosophila*, which was first proposed by Wheeler et al. (1962) as a Neotropical group of species strictly adapted to flowers of *Cestrum* both in terms of oviposition and feeding. This species detaches among the other species of the *flavopilosa* group by having one of the most wide distributions, presenting registers from South America to Central America (Bachli, 2016). Even so, the species seems to be specially adapted to the environmental conditions encountered in the Southern region of Brazil (Robe et al. 2013), where it has been encountered only in association with flowers of eight species of *Cestrum* (*C. amictum, C. calycinum, C. corymbosum, C. intermedium, C. nocturnum, C. parqui, C. schlechtendalii* and *C. sendtnerianum* and *Sessea brasiliensis* (Santos & Vilela, 2005). According to Napp and Brncic (1978) this species lays its eggs in open flowers, scarifying the internal surface of the petals, which characterizes an ecological segregation in regard to *D. cestri*, frequently encountered in sympatry and syntopy with *D. incompta* (Robe et al. 2013).

Due to this narrow ecology profile, the distribution of *D. incompta* and the other species of the *flavopilosa* group is completely dependent on the distribution of their hosts, which seem to be abundant in the Neotropics (Hofmann, 1985). In Brazil, more specifically, the greatest diversity of *Cestrum* species is located in the Atlantic Forest and Cerrado Biomes (Soares, Silva and Mentz, 2007). Nevertheless, it is acknowledged that the distribution of these resources is somewhat discontinuous along the year (Sepel et al. 2000), and the populational dynamics able to explain the persistence of the species under these special circumstances still remains to be addressed (Robe et al. 2013). Frequent bottlenecks, alternate host flourishing patterns associated with migration and even diapause are among the factors invoked to explain the maintenance of *D. incompta* throughout the less favorable conditions. Nevertheless, the high levels of genetic diversity previously detected for this species (De Ré

et al. 2014b) suggest that the subjacent evolutionary scenario may be even more complex than previously thought.

So, given this close insect-plant relationship, we seek to identify the past demographic events able to explain the current patterns of genetic diversity revealed for *D. incompta*. After all, obtaining answers to the demographic issues involving populations of *D. incompta* is fundamental to understanding the true history and the evolutionary potential of the group taken as a whole.

2. Material and Methods

2.1. Samples

Collections of *Cestrum* (Solanaceae) flowers were performed in two of the southern Brazilian states: Rio Grande do Sul (RS) and Paraná (PR) (Figure 1, Table 1). After collection, the flowers were taken to the laboratory, where they were kept until the eclosion of adult drosofilids. After hatching the flies were separated by genus and by sex through external morphology, and then the males of the *Drosophila flavopilosa* group were identified through their internal morphology and genitalia patterns, according to the pictures provided by Wheeler et al. (1962) and Danko Brncic (unpublished data).

2.2. DNA isolation and sequencing

Total DNA was extracted from each individual using the NucleoSpin Tissue XS kit (MACHEREY-NAGEL). Fragments of the mitochondrial Cytochrome Oxidase c Subunit I (*COI*) and *Cytochrome Oxidase* c *Subunit II* (*COII*) genes were amplified using the primer pairs TYJ1460 and C1N2329 (as modified by Bolzan et al. 2011 from Simon et al. 1994), and TL2J3037 and TKN3785, respectively (Simon et al. 1994), whereas fragments of the nuclear *Hunchback (hb)* gene were amplified using primers HB106F and HB903R (Mota et al. 2008). The amplification properties were identical for *COI, COII* and *Hb* markers and followed the

general protocol described by De Ré et al. (2014a), with the exception of the annealing temperature which was set to 58ºC, 60ºC and 55ºC, respectively. PCR reactions were carried out using 10-50ng of DNA, 1x buffer, 0.2µM of each primer, 0.25 mM of each dNTP, 2.5 mM of $MgCl_2$, and 1U of *Taq* DNA polymerase.

In order to verify whether the amplification was successful, 5 ul of the PCR product were applied to 0.8% agarose gel, stained with a solution of gel loading buffer containing GelRed™ (Uniscience). Then, the obtained amplicons were purified with the use of a solution containing 13% PEG and 1.6M NaCl and directly sequenced. Sequencing was performed on a MegaBACE 500 automatic sequencer using the DYEnamic ET® kit (Amersham), according to the protocol provided by the manufacturer using the same primers described above.

2.3 Data analysis

First of all, the sequence chromatograms were inspected and assembled using the Staden Package Gap4 Program (Staden, 1996). The consensus sequences so obtained had their identity confirmed by BLASTN (NCBI website), and were then aligned using the algorithm ClustalW, as implemented in Mega 5.0 software (Tamura et al. 2011). To ensure the accuracy of the analyses while excluding possible sequencing artifacts, each polymorphic site detected in the intraspecific alignments was individually checked in the respective chromatograms. Finally, 10 *COI* sequences of *D. incompta* specimens whose precedence was traceable were downloaded from GenBank (Tabela 2). With respect to the nuclear gene, in order to avoid biases related to the occurrence of recombination, the phase of diploid unphased data was first reconstructed in DNAsp5 (Librado and Rozas, 2009). This software was also used to estimate the recombination parameter (R) (Hudson, 1987) and the minimum number of recombination events (Rm) (Hudson and Kaplan 1985) affecting *Hb* sequences.

2.3.1. Phylogeographic analyses

Measures of genetic diversity [nucleotide diversity (π), number of haplotypes (h), haplotype diversity (Hd), and number of polymorphic sites (s)], and neutrality [Tajima's D (Tajima, 1989), Fu's D, Li's D, Fu's F and Li's F (Fu, 1997)] were obtained individually for each mitochondrial and nuclear gene in DNAsp 5 (Librado and Rozas, 2009). Differentiation levels between sampling localities were estimated by FST indices in Arlequin v. 3.5, with 100 permutations (Excoffier and Lischer, 2010). A Mantel test, with 1000 permutations, was also carried out in Arlequin v. 3.5 to assess whether there is a significant positive correlation between genetic and geographical distances, implying the suitability of an "isolation by distance" model. For this test, we used the linear geographic distance between populations calculated in DIVA-GIS program 7.1 (Hijmans et al. 2005) based on the geographical coordinates of each sampled location and FST values as representative of the genetic distances. A SAMOVA test was further implemented in SAMOVA 2.0 (Dupanloup, Schneider and Excoffier (2002) in order to define groups of populations that are geographically homogeneous and maximally differentiated from each other. Finally, the relationships between the resulting haplotypes were inferred by median joining in the Network software (Bandelt et al. 1999). In this case, it is important to note that before reconstructing the network, a weight value of 0 was assigned to each recombinant site in the nuclear DNA matrix, aiming to reduce the effect of loops in the haplotypes tree.

The size distribution of the *D. incompta* population over time was evaluated by a Bayesian Skyline analysis individually for each mitochondrial gene, as in runs of BEAST 1.8.0 program (Drummond et al. 2012). To perform this analysis, the evolutionary rates were set to 7.42×10^{-3} and 1.47×10^{-2}, as previously estimated by De Ré et al (2014a) for *COI* and *COII*, respectively. Evolutionary models suitable for each gene were evaluated by the Akaike Information Criterion (AIC) (Akaike, 1974), as implemented in MrModelTest 2.3 (Nylander, 2004), with the aid of PAUP 4.0a147 (Swofford, 2003), and encompassed HKY and

GTR+I+G, respectively. The number of generations in MCMC (Markov Chain Monte Carlo) was 50 million, sampled every 1,000 interactions. The samples obtained in BEAST were evaluated in Tracer 1.5 (Rambaut and Drummond, 2009), where the ESS values (Effective Sample Size) of each parameter and the convergence throughout each run were evaluated. Moreover, Mismatch Distributions were also carried out in DNAsp 5 for all three genes to evaluate the occurrence of population expansion, with significance evaluated through 1,000 random permutations.

3. Results

3.1 Collection and sequence analysis

After adult eclosion, 123 individuals of *D. incompta* were identified across 16 collection points (Table 1, Fig. 1). Regarding the two mitochondrial markers, about 660 base pairs (bp) of the *COI* gene were sequenced for 66 individuals, whereas 539 bp of the *COII* gene were characterized for 24 specimens. In addition, for *Hb*, 33 sequences with 609 bp were obtained.

3.2 Genetic diversity

Table 3 shows the diversity indices presented by *D. incompta* for each of the three genes analyzed. Considering the two mitochondrial genes, the analysis of *COI* for 66 individuals showed the presence of 18 haplotypes distributed along 15 sampling sites, whereas for *COII*, 11 haplotypes were sampled for only 24 individuals collected in three points. In fact, despite the reduced sample size, *COII* appears to be more variable than *COI*, presenting higher values of haplotypic (0.895±0.038 vs. 0.824±0.030, respectively) and nucleotide diversities (0.00597±0.00081 vs. 0.00352±0.00035, respectively) (Table 3). Unexpectedly, *Hb* showed even higher diversity values, totaling 24 haplotypes in 33 sampled individuals, distributed in 11 sampling sites. This marker also showed high haplotypic (Hd/DP = 0.916/0.027) and

nucleotide diversity measures (0.01021±0.0011), and presented recombination signs between 5 sites (Rm = 5), namely: 178 - 219; 219 - 385; 459 - 460; 460 - 464; 464 - 470 (Table 3).

3.3 Population structure

The networks recovered by the tree markers here employed evidenced starlike patterns, with no apparent geographic structure and with a high number of exclusive haplotypes/sequences (14, 8 and 19 for COI, *COII* and *Hb*, respectively) (Fig. 2A - B and Fig. 3). For *COI*, from the remaining four shared haplotypes, haplotype 4 (H4) was found in 21 individuals from 10 of the 15 sampling sites (Fig. 2A), whereas haplotypes $H_{8,}$ H_9 and H_7 presented narrower distribution ranges, being sampled in six, seven and three of the sampling localities, respectively (Fig. 2A). For *COII*, the three shared haplotypes were sampled in at least two of the tree sampling sites, with haplotype 3 (H_3) being both, the most frequent and widely distributed (Fig. 2B). All the *COI* and *COII* shared haplotypes were sampled in the population of Curitiba/PR, which is the collection point further north used in our analysis. Moreover, for both genes, all unique haplotypes were separated from a shared haplotype by a maximum of 3 mutational steps (Fig. 2A and B). Although for *Hb* the maximum number of mutational steps separating any two sequences was higher, reaching values of 10, the predominance of a single haplotype in terms of frequency and distribution was notable (Fig. 3). In fact, haplotype H_1 was shared by seven of the analyzed populations, although it was not encountered in the population of Curitiba/PR. Even so, the second most common haplotype (H_3) was detected in Curitiba/PR and Santiago/RS.

From the 105 Fst pairwise comparisons recovered with *COI* sequences, only seven were significant, and from these, five referred to Itaara, a population from the center of the Brazilian Rio Grande do Sul State which revealed moderate to pronounced levels of genetic differentiation ($0.14046 < Fst < 0.63746$) in regard to populations located to the South (Canguçu and Pelotas) or the North of the same State (Frederico Westphalen, Montenegro and

Porto Alegre). The other two *COI* Fst comparisons presenting significant values occurred between Cruz Alta and Porto Alegre (Fst = 0.24186) and Santa Maria and Pelotas (Fst = 0.53125) (Tab. 4). With respect to COII, only the comparison involving Curitiba and São João do Polêsine presented a significant, but moderate, Fst value (Fst = 0.09165) (Tab. 5). These two populations also presented significant, but pronounced levels of genetic differentiation for *Hb* (Fst = 0.54286), and at least one of them were involved in seven of the 11 significant Fst comparisons obtained for this marker. The other four cases of significant Fst values presented for Hb occurred between Itaara in relation to Rio Grande (Fst = 0.13015) or Montenegro (Fst = 0.44954) and Frederico Westphalen in relation any of these two localities (Fst = 0.14504 and 0.29707, respectively).

The absence of a visible geographical pattern in these significant Fst comparisons was further confirmed by the SAMOVA and Mantel Tests results. In fact, the SAMOVA revealed that, in all cases, higher Fct values tend to be recovered when each sampling locality is considered as a single unit, although even at these cases, at least 83% of the genetic variability is found within populations. In regard to the Mantel test, correlation coefficients ranged from -0.11 and 0.03 for *Hb* and *COI*, respectively, to 0.51 for *COII*, although in no case the estimates were significant.

3.4 Effective population size and demographic history

In general, significant negative values were found for *D. incompta* in regard to the *COI* mitochondrial gene for all three performed neutrality tests (Table 3). Statistically significant negative values indicate an excess of rare polymorphisms in a population, which is consistent with demographic events of population expansion. Conversely, for *Hb*, Fu and Li`s D presented significant positive values (Table 3), indicating an excess of intermediate-frequency alleles resulting either from balancing selection or from population bottlenecks.

The results of the neutrality tests and the starlike pattern recovered for both mitochondrial genes agree with the Bayesian skyline plots, which confirmed the occurrence of a population expansion event in populations of *D. incompta*. Although signs of population expansion were recovered for both, *COI* and *COII*, the impact was more subtle for the last gene, probably as a result of the lower sample size. According to our data, this expansion started between 175,000 and 100,000 years before present (Fig. 4). In fact, the mismatch distribution recovered for *COI* presented a unimodal smooth distribution consistent with a more ancient population expansion (Fig. 5a), although signs of demographic equilibrium (Excoffier et al. 1992) were recovered by this test for *COII* and *Hb*, for which a bimodal ragged distribution was presented (Figure 5b and c).

10. Discussion

Overall, the mitochondrial and nuclear markers employed here suggested two distinct evolutionary/demographic scenarios for *D. incompta*: the first genes consistently support the occurrence of a population expansion, whereas the second reports a reduction in population size (bottleneck). As the time to coalescence may vary between markers (Templeton, 2006), such a disagreement may stem from the fact that mitochondrial and nuclear genes are reporting different though not mutually exclusive moments of the species history. Alternatively, this result may also reflect the action of balancing selection in *Hb*, which increases the genetic variability in relation to neutrality patterns (Aguilar et al. 2004). As the high levels of genetic variability detected for *D. incompta* in regard to Hb are compatible with the results obtained for cytogenetics (Brncic, 1962) and isozyme markers (Napp and Brncic, 1978), we follow here the reasoning that incongruence stems mainly from differences in coalescent times, although the definitive source of this inconsistence can only be achieved with the inclusion of additional nuclear sequences,

In this case, whereas the significantly negative estimates of the neutrality tests, the starlike pattern of the network, the unimodal distribution of the mismatch analysis and the Bayesian skyline presented for at least one of the mitochondrial markers indicate that this species has undergone a population expansion event, the significantly positive estimates of Fu and Li's D neutrality test showed for *Hb* suggest that this expansion was possibly preceded by a bottleneck. According to the demographic analysis, the expansion event occurred between 175,000 and 100,000 years before present and is possibly linked to the constant and regular climate changes occurred along the Quaternary period, where repeated glacial and interglacial periods widely affected the distribution of living organisms (Bennett, 1997; Hewitt, 2000). In fact, studies with species in the Northern Hemisphere reveal a pattern of contraction and geographic isolation of populations in the glacial periods, followed by population expansion in the interglacial periods (Hewitt, 2000, Alexandrino et al. 2002, Kotlik et al. 2004). Although the oscillations were quite more subtle along the Southern Hemisphere, during the Quaternary, this region experienced a period of alternating dry and wet conditions, with forested and non-forested biomes constantly changing in distribution, becoming potential refugia, and then expanding again (Haffer, 1997). Thus, the population expansion event detected for *D. incompta* must have occurred at a time when climatic conditions were favorable to the establishment of the *Cestrum* host plants. Previous to this, when the abiotic factors necessary for the survival of flowers were absent or limited, there was a reduction of the available resources and a consequent reduction in the population size of the associated drosophilids, which is compatible with the historical bottleneck signals detected for *Hb*.

Studies with other drosophilids, restricted to xeric environments also reflected population growth due to Pleistocene events in Brazil (Franco et al. 2010, Franco and Manfrin, 2013; Moraes et al. 2009) and in the North Hemisphere (Pfeiler et al. 2007), but contrary to what is hypothesized here for *D. incompta*, in such cases the expansion seems to have been associated

with periods of greater aridity, since those species use cacti as resource. However, changes in the effective population size along the Quaternary are not restricted to specialist species. Recently, De Ré et al. (2014a) showed that *D. maculifrons*, a generalist species, went through a period of recent population expansion, subsequent to the period of the Last Glacial Maximum, revealing extremely low genetic diversity patterns.

Interestingly, after the expansion signal between 175,000 and 100,000 years before present, there were no evidences of further reduction/expansion of *D. incompta* populations, which seem to have reached an equilibrium. Thus, our results suggest that this species has not been affected by the climate changes of the Last Glacial Maximum, which occurred about 20-18 thousand years ago (Ledru et al. 1998a). As in Brazil, the greatest diversity of *Cestrum* is located in the Atlantic Forest and Cerrado Biomes (Soares, Silva and Mentz, 2007), switching between forested and not forested biomes may have affected the establishment of *D. incompta* host only in subtler levels, thus promoting resource availability for the maintenance of the fly populations. Nevertheless, even under this scenario of resource availability, different species of the *flavopilosa* group may have responded in different manners to the environmental oscillations. In fact, *D. incompta* seems to be more tolerant than some other species of the group that are sympatric to the same region in relation to environmental variables such as temperature and precipitation (Robe et al. 2013), and this is another interesting aspect that may have favored the establishment of equilibrium conditions after the population expansion event.

Although the occurrence of successive bottlenecks was previously invoked to explain the persistence of *D. incompta* during periods of resource scarcity, the detection of signals of demographic equilibrium and the high levels of genetic variability detected for this species by all markers employed here, and also by cytogenetic (Brncic, 1962) and isozyme analysis (Napp and Brncic, 1978; Brncic and Napp, 1980) does not support such a scenario. In fact,

bottlenecks are expected to cause a reduction in genetic variability, especially in regard to haplotypic diversity levels. These high diversity levels and the absence of geographically structured genetic differentiation patterns given by the negative results recovered for SAMOVA and Mantel, in spite of some punctual moderate to high significant Fst measures, best support a scenario of recurrent but scattered gene flow along alternate host flourishing seasons as the strategy adopted by *D. incompta* to overcome periods of scarcity of resources. Nevertheless, despite these interesting demographic and genetic diversity results found for the populations of *D. incompta* in the southern region of Brazil, only an increased sample, especially concerning the nuclear gene, will be able to add confidence to the scenarios here reported.

Figure 1: Geographical distribution of the *D. incompta* sampling localities in the Rio Grande do Sul (RS) and Paraná (PR) Brazilian states. 1) Cachoeira do Sul, 2) Canguçu, 3) Cruz Alta, 4) Curitiba, 5) Frederico Westphalen, 6) Itaara, 7) Montenegro, 8) Pelotas, 9) Porto Alegre, 10) Rio Grande, 11) Saldanha Marinho, 12) Santa Maria, 13) Santiago, 14) São João do Polêsine, 15) São Sepé.

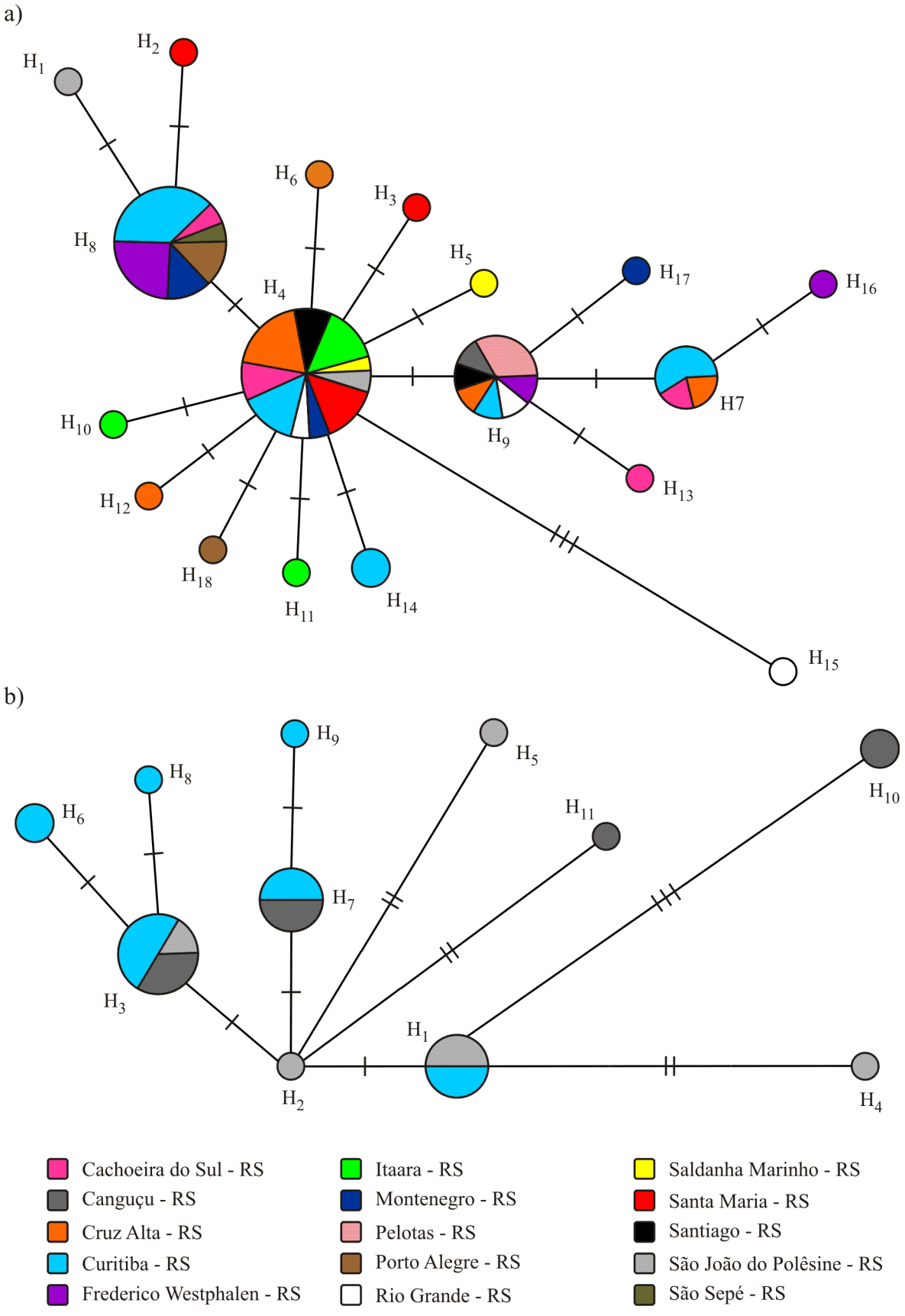


Figure 2: Median-joining network of *D. incompta* based on COI (A) and COII (B). Lines correspond to mutacional steps connecting haplotypes, represented by circles, whose size is proportional to frequency and whose colors refer to geographic origins, given in the Legend and in Figure 1.

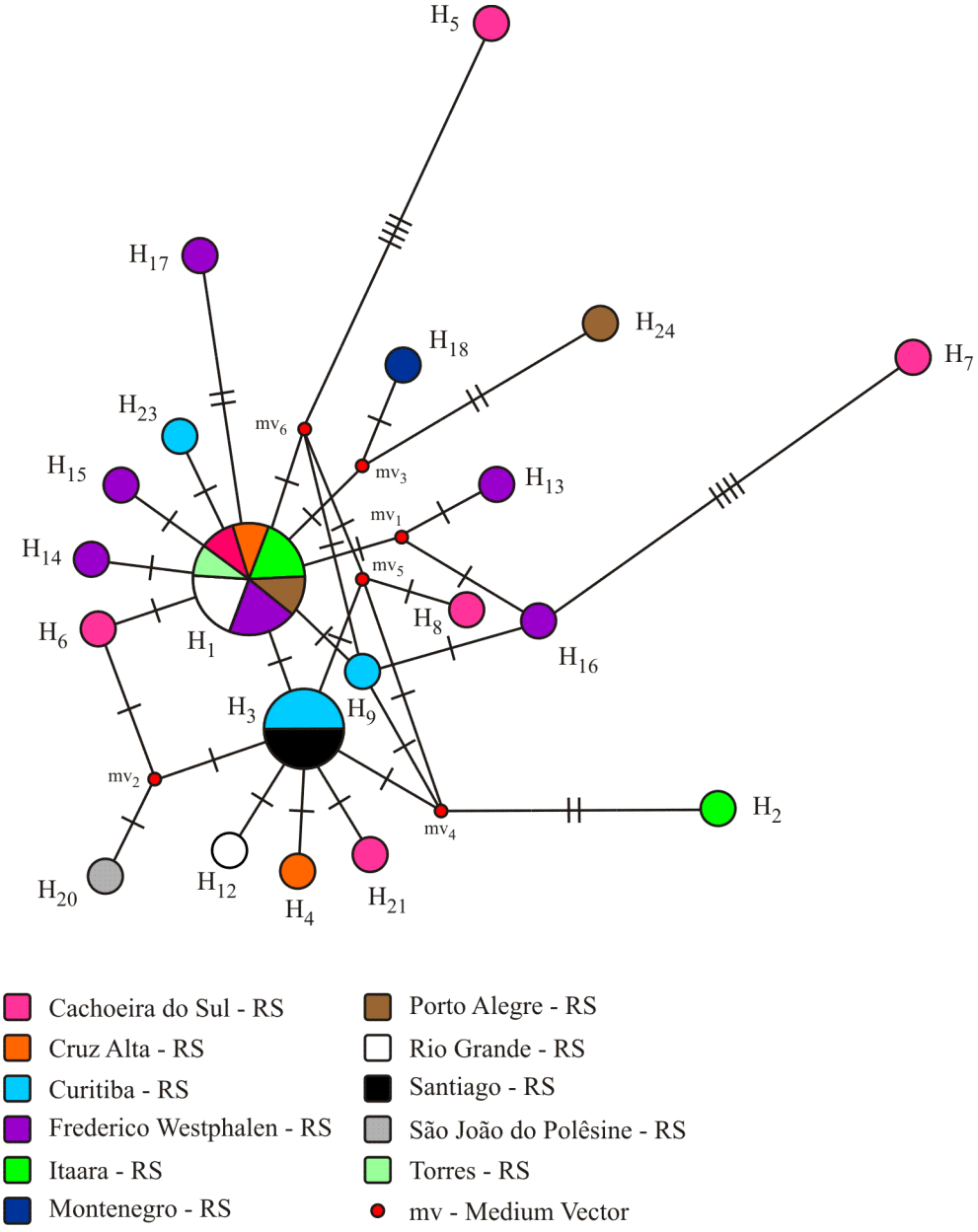
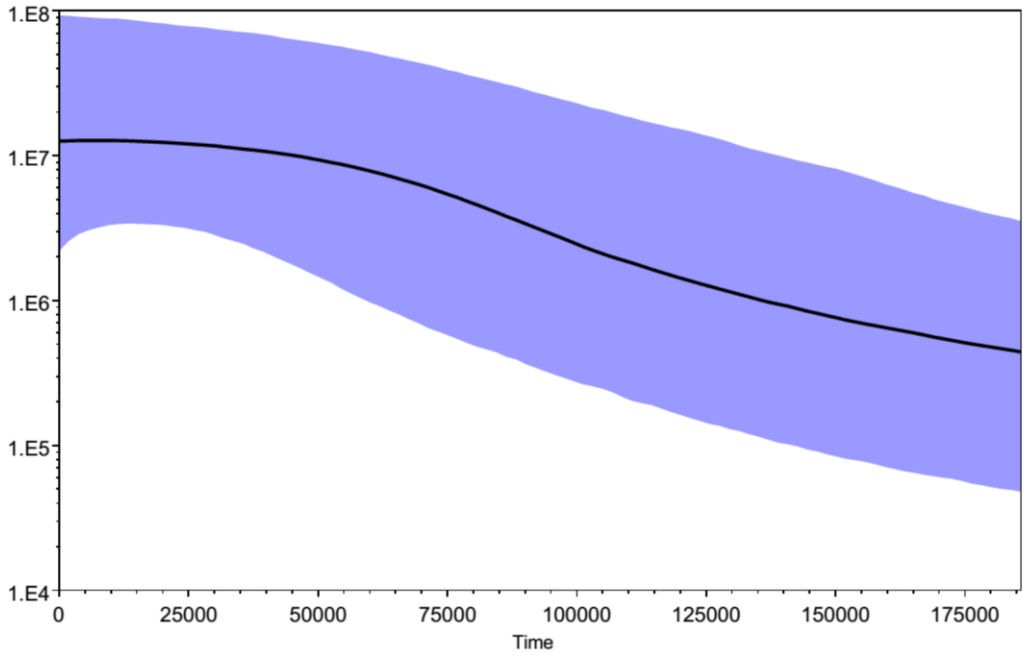


Figure 3: Median-joining network of *D. incompta* based on Hb sequences. Lines correspond to mutacional steps connecting haplotypes, represented by circles, whose size is proportional to frequency and whose colors refer to geographic origins, given in the Legend and in Figure 1.

a)



b)

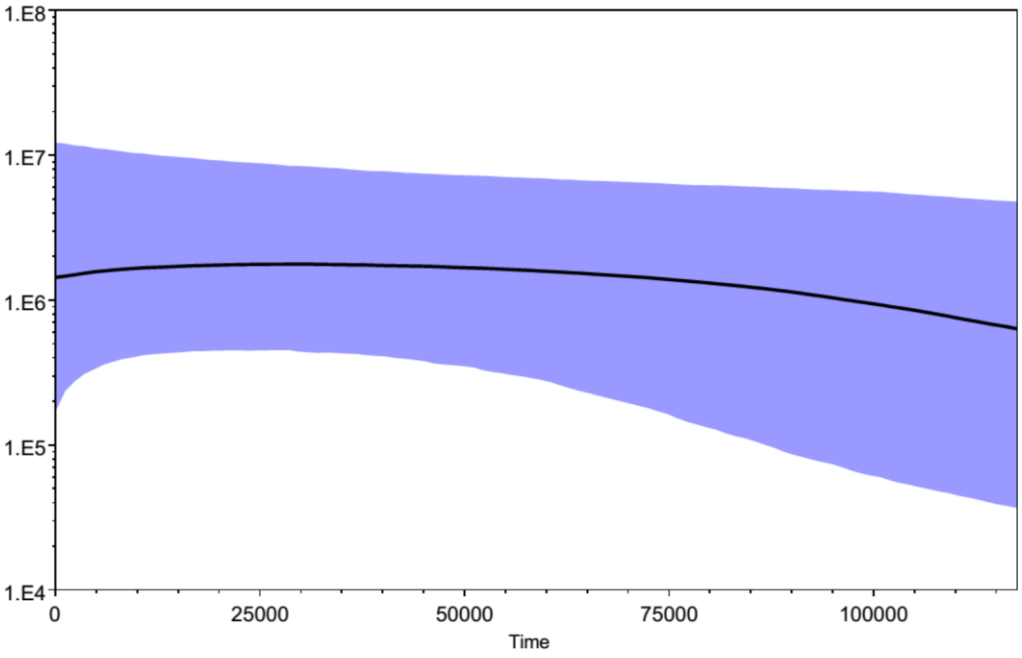


Figure 4: Graph showing the variation in effective population size inferred for *D. incompta* over time by BSP analysis. The graph was generated using *COI* (A) and *COII* (B) data sets.

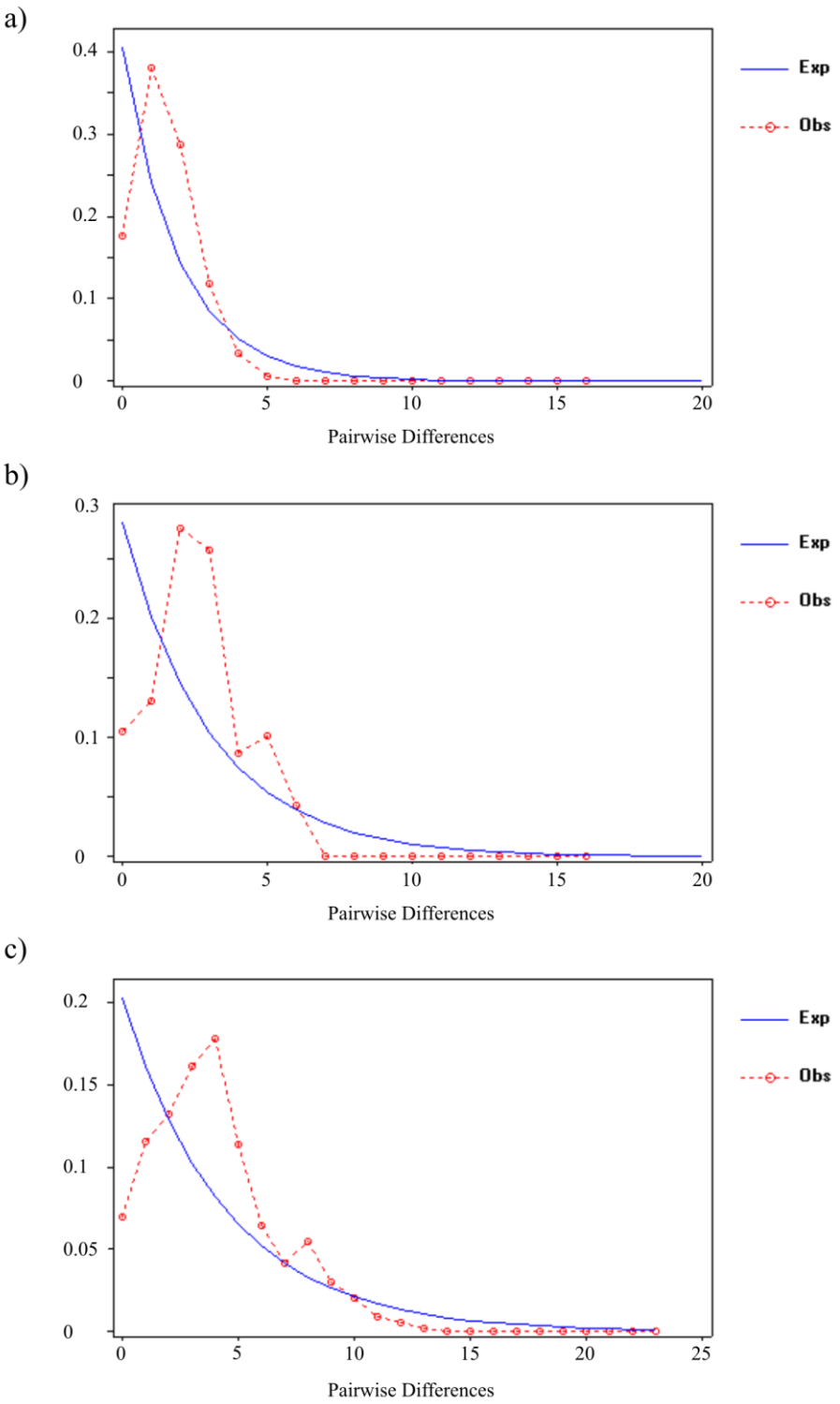


Figure 5: Mismatch Distribution of *D. incompta* based on *COI* (A), *COII* (B) and *Hb* (C) sequences.

Table 1: List of sampling points, with their respective date of collection, number of individuals sequenced and geographical coordinates

	Collection Points	Number of individuals sequenced for COI/COII/Hb	Collection date	South Coordinates (S)	West Coordinates (W)
RS	Cachoeira do Sul	5/0/6	14/04/2011	30º02’01.25”	52º53’35.60”
	Canguçu	2/7/0	24/08/2011	31º23’47.32”	52º40’43.63”
	Cruz Alta	8/0/2	26/09/2010	28°34'06.95"	53°37'20.72"
	Frederico Westphalen	6/0/7	10/04/2013	27º21’32.84”	53º23’46.81”
	Itaara	5/0/3	26/09/2010	29º35’27.36”	53º45’31.00”
	Montenegro	4/0/1	23/07/2012	29º40’58.49”	51º28’06.30”
	Pelotas	3/0/0	30/05/2012	31°46'02.05"	52°26'55.34"
	Porto Alegre	3/0/2	12/07/2010	30°04'35.56"	51°07'27.70”
	Rio Grande	3/0/4	14/08/2012	32°03'02.24"	52° 05'37.73"
	Saldanha Marinho	2/0/0	26/09/2010	28º23’40.45”	53º05’51.09”
	Santa Maria	5/0/0	21/09/2010	29°41'17"	53°48'21.73"
	Santiago	3/0/1	04/08/2010	29º11’29.26”	54º51.’59.51”
	São João do Polêsine	2/6/1	29/07/2010	29°38'59.73"	53°31'00.33"
	São Sepé	1/0/0	10/06/2012	30º12’11.70”	53º36’24.71”
	Torres	0/0/2	05/02/2012	29°44’40.44”	57°04’53.68”
PR	Curitiba	14/11/4	25/02/2013	25º25’27.94”	49º15’55.37”

Table 2 - Additional sequences downloaded from GenBank for the COI mitochondrial gene with their respective access numbers

Espécie	Ponto de Coleta	COI
1 .*D. incompta*	Canguçu	JX993107
2. *D. incompta*	Pelotas	JX993106
3. *D. incompta*	São Sepé	JX993105
4. *D. incompta*	Curitiba	JX993104
5. *D. incompta*	Cruz Alta	JX993103
6. *D. incompta*	Saldanha Marinho	JX993102
7. *D. incompta*	Saldanha Marinho	JX993101
8. *D. incompta*	Santa Maria	JX993100
9. *D. incompta*	Santa Maria	JX993099
10. *D. incompta*	São João do Polêsine	JX993098

Table 3: Diversity values, neutrality tests and recombination results for *D. incompta* in regard to *COI*, *COII* and *Hb*

	COI	***COII***	***Hb***
N	66	24	33
Nps	*	*	66
S	19	15	34
h	18	11	24
Hd (DP)	0.824 (0.030)	0.895 (0.038)	0.916 (0.027)
π% (DP)	0.00352 (0.00035)	0.00597 (0.00081)	0.01021 (0.00110)
Tajima's D	-1.91 [p < 0.05]	-1.27 [P > 0.10]	-1.58 [0.10 > P > 0.05]
Fu and Li's D	-4.56 [p < 0,02]	-1.38 [P > 0.10]	1.94 [p < 0,02]
Fu and Li's F	-4.30 [p < 0,02]	-1.58 [p > 0.10]	0.74 [P > 0.10]
R	*	*	14
Rm	*	*	5
Recombination between sites	*	*	(178,219) (219,385) (459,460) (460,464) (464,470)

N = number of individuals; S = number of polimorphic sites; h = haplotype number; Hd (SD) = haplotypic diversity (standard deviation); π% (SD) = nucleotide diversity (standart deviation); Rm = minimum number of recombination events; R = estimate of the recombination parameter; *Nps* = number of sequences after phasing; "*" indicates data not calculated for a particular marker.

Table 4: Fst values based on *COI* sequences are below diagonal and *p* values based on *COI* sequences are above diagonal

	São João do Polesine	Santa Maria	Saldanha Marinho	Itaara	Santiago	Cruz Alta	Cachoeira do Sul	Curitiba	Rio Grande	Monte Negro	Canguçu	São Sepé	Frederico Westphalen	Porto Alegre	Pelotas
São João do Polesine		0.66+-0.03	0.99+-0.00	0.36+-0.05	0.40+-0.03	0.28+-0.04	0.49+-0.05	0.64+-0.06	0.75+-0.03	0.77+-0.03	0.34+-0.02	0.99+-0.00	0.71+-0.03	0.74+-0.03	0.09+-0.02
Santa Maria	0		0.56+-0.04	0.99+-0.00	0.81+-0.03	0.28+-0.06	0.76+-0.03	0.21+-0.03	0.26+-0.03	0.29+-0.03	0.09+-0.02	0.99+-0.00	0.19+-0.04	0.42+-0.04	0.009+-0.009
Saldanha Marinho	0	0		0.69+-0.05	0.72+-0.03	0.70+-0.05	0.50+-0.06	0.31+-0.03	0.90+-0.02	0.54+-0.05	0.34+-0.04	0.99+-0.00	0.23+-0.04	0.37+-0.04	0.09+-0.02
Itaara	0.17197	0	0.0411		0.79+-0.02	0.38+-0.05	0.55+-0.04	0.05+-0.02	0.31+-0.05	0.04+-0.01	0.02+-0.01	0.99+-0.00	0.04+-0.02	0.009+-0.00	0.02+-0.01
Santiago	0.11765	0	0.04545	0		0.99+-0.00	0.99+-0.00	0.49+-0.04	0.99+-0.00	0.63+-0.03	0.32+-0.04	0.99+-0.00	0.24+-0.03	0.40+-0.05	0.36+-0.06
Cruz Alta	0.14024	0.02992	0	0.01906	0		0.72+-0.03	0.15+-0.04	0.27+-0.04	0.23+-0.04	0.18+-0.02	0.99+-0.00	0.06+-0.02	0.01+-0.01	0.07+-0.02
Cachoeira do Sul	0	0.01316	0	0.07143	0	0		0.49+-0.05	0.63+-0.03	0.77+-0.04	0.29+-0.06	0.99+-0.00	0.44+-0.03	0.23+-0.03	0.31+-0.05
Curitiba	0	0.02396	0.08197	0.12258	0	0.07356	0		0.10+-0.01	0.82+-0.02	0.18+-0.02	0.99+-0.00	0.57+-0.03	0.33+-0.05	0.05+-0.02
Rio Grande	0	0.06327	0	0.09263	0	0.02662	0	0.13696		0.54+-0.04	0.51+-0.04	0.99+-0.00	0.18+-0.04	0.44+-0.02	0.35+-0.03
Monte Negro	0	0	0.04	0.14046	0	0.07726	0	0	0.02362		0.36+-0.03	0.99+-0.00	0.90+-0.02	0.81+-0.02	0.10+-0.02
Canguçu	0.4	0.45386	0.5	0.55224	0.32258	0.24393	0	0.28475	0.06557	0.27273		0.99+-0.00	0.47+-0.05	0.09+-0.03	0.41+-0.04
São Sepé	0	0	0.33333	0.42857	0.5	0.27473	0	0	0	0	0.6		0.99+-0.00	0.99+-0.00	0.16+-0.02
Frederico Westphalen	0	0.11274	0.16746	0.25419	0.09091	0.17862	0	0	0.14286	0	0.23348	0		0.67+-0.03	0.18+-0.05
Porto Alegre	0	0.0583	0.19512	0.27711	0.25	0.24186	0.07346	0	0.14286	0	0.51471	0	0		0.06+-0.03
Pelotas	0.64706	0.53125	0.76923	0.63746	0.5	0.30343	0.11765	0.32562	0.2	0.36842	0.25	1	0.33898	0.66667	

Table 5: Fst values based on *COII* sequences are below diagonal and *p* values based on *COII* are above diagonal

	São João do Polêsine	Curitiba	Canguçu
São João do Polêsine		0.04505+-0.0244	0.36937+-0.0459
Curitiba	0.09165		0.23423+-0.0364
Canguçu	0.02999	0.04423	

Table 6: Fst values based on *Hb* sequences are below diagonal and *p* values based are above diagonal

	Itaara	Santiago	Cruz Alta	Cachoeira do Sul	Curitiba	Torres	Rio Grande	Frederico Westphalen	Montenegro	São João do Polêsine	Porto Alegre
Itaara		0.45+-0.04	0.23+-0.03	0.40+-0.05	0.42+-0.03	0.33+-0.04	0.02+-0.01	0.09+-0.02	0.02+-0.01	0.09+-0.02	0.13+-0.02
Santiago	0.0625		0.27+-0.03	0.45+-0.04	0.63+-0.05	0.15+-0.03	0.66+-0.05	0.14+-0.02	0.29+-0.03	0.35+-0.03	0.14+-0.03
Cruz Alta	0.11111	0.11111		0.18+-0.03	0.27+-0.03	0.32+-0.04	0.23+-0.03	0.18+-0.02	0.07+-0.02	0.09+-0.02	0.20+-0.05
Cachoeira do Sul	0.01241	-0.05912	0.05882		0.13+-0.03	0.16+-0.03	0.05-0.02	0.09+-0.02	0.06+-0.01	0.00+-0.00	0.09+-0.03
Curitiba	0.02123	-0.08185	0.05398	0.0655		0.37+-0.04	0.19+-0.05	0.009+-0.009	0.02+-0.01	0.02+-0.01	0.02+-0.01
Torres	0.16245	0.11111	0.16667	0.06694	0.03614		0.71+-0.04	0.09+-0.03	0.09+-0.03	0.11+-0.03	0.20+-0.04
Rio Grande	0.13015	-0.06667	0.05882	0.1169	0.04762	-0.07975		0.00+-0.00	0.12+-0.03	0.02+-0.01	0.05+-0.02
Frederico Westphalen	0.06712	0.11579	0.11111	0.06004	0.10099	0.12752	0.14504		0.01+-0.01	0.02+-0.01	0.05+-0.02
Montenegro	0.44954	1	0.68	0.18514	0.43389	0.59322	0.48387	0.29707		0.31+-0.03	0.27+-0.04
São João do Polêsine	0.58621	1	0.75758	0.28419	0.54286	0.68	0.54286	0.50148	1		0.06+-0.01
Porto Alegre	0.24919	0.44186	0.33333	0.14019	0.25234	0.33333	0.27572	0.1586	0.44186	0.73626	

5. References

AGUILAR, Andres et al. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proceedings of the National Academy of Sciences of the United States of America, v. 101, n. 10, p. 3490-3494, 2004.

AKAIKE, H. A new look at the statistical model identicafition. IEEE Trans. Automat. Contr. 19, 716–723, 1974.

ALEXANDRINO J., ARNTZEN J. W., FERRAND N. Nested Clade Analysis and the genetic evidence for population expansion in the phylogeography of the golden-striped salamander, *Chioglossa lusitanica* (Amphibia: Urodela). Heredity 88: 66 – 74, 2002.

BÄCHLI, G. TaxoDros: The Database on Taxonomy of Drosophilidae, v. 1.03, Database 2009/04. http://taxodros.unizh.ch/. Last accessed on 01/03/2016.

BANDLET, H., FORSTER, P., ROHL, A. Median-joining network for inferring intraspecific phylogenies. Molecular Biology and Evolution, v.161), p.37-48, 1999.

BENNETT, K.D. Evolution and ecology. The pace of life. Cambridge studies in ecology. University Press, Cambridge. 241 pp. 1997.

BOLZAN, A.R. DNA barcode de drosofilídeos micófagos pertencentes aos gêneros *Hirtodrosophila*, *Mycodrosophila* e *Zygothrica*. 2011. Dissertação (Mestrado em Biodiversidade Animal)-Universidade Federal de Santa Maria, Santa Maria, 2011.

BRNCIC, D. Ecological and cytogenetic studies of *Drosophila flavopilosa*, a Neotropical species living in *Cestrum* flowers. Evolution; 20:16-29, 1966.

DE RÉ, et al. Brazilian populations of *Drosophila maculifrons* (Diptera: Drosophilidae): low diversity levels and signals of a population expansion after the Last Glacial Maximum. Biological Journal of the Linnean Society, v. 112, n. 1, p. 55-66, 2014a.

DE RÉ, F. C. et al. Characterization of the complete mitochondrial genome of flower-breeding Drosophila incompta (Diptera, Drosophilidae). Genetica, v. 142, n. 6, p. 525-535, 2014b.

DRUMMOND, Alexei J. et al. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution, v. 29, n. 8, p. 1969-1973, 2012.

DUPANLOUP, Isabelle; SCHNEIDER, Stefan; EXCOFFIER, Laurent. A simulated annealing approach to define the genetic structure of populations.Molecular Ecology, v. 11, n. 12, p. 2571-2581, 2002.

EXCOFFIER, Laurent; SMOUSE, Peter E.; QUATTRO, Joseph M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, v. 131, n. 2, p. 479-491, 1992.

EXCOFFIER, L., LISCHER H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analysis under linux and windows. Molecular Ecology Resources, 10, 564-567. 2010.

FRANCO, F. F. et al. Intra- and interspecific divergence in the nuclear sequences of the clock gene period in species of the *Drosophila buzzatii* cluster. Journal of Zoological Systematics and Evolutionary Research 33, 225-223, 2010.

FRANCO, F. F., MANFRIN, M. H. Recent demographic history of cactophilic *Drosophila* species can be related to Quaternary palaeoclimatic changes in South America. Journal of Biogeography, 40: 142–154, 2013.

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

HAFFER, J. Alternative models of vertebrate speciation in Amazonia: an overview. Biodiversity and Conservation, v.6 (3), p.451-476, 1997.

HEWITT,G. M. The genetic legacy of the Quaternary ice ages. Nature 405,907–913, 2000.

HIJMANS, R. J. et al. DIVA-GIS Version 5.2. Manual. http://www.diva-gis.org. 2005a.

HOFMANN, P. R. P. Variabilidade genética em espécies de nível ecológico restrito. Ciência e Cultura. 37:579-581, 1985.

HUDSON, Richard R. Estimating the recombination parameter of a finite population model without selection. Genetical research, v. 50, n. 03, p. 245-250, 1987.

HUDSON, Richard R.; KAPLAN, Norman L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics, v. 111, n. 1, p. 147-164, 1985.

KOTLIK, P. B. N., EKMEKCI, F. Circum black sea phylogeography of Barbus freschwater fishes: divergence in the Pontic glacial refugium. Molecular Ecology 13 (1): 87-95, 2004.
LEDRU, M. P. *et al*. Absence of last glacial maximum records in lowland tropical forests. Quaternary Research 49, 233– 237, 1998a.

LIBRADO, P., ROZAS J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics, 25, 1451-1452, 2009.

MORAES, E.M. et al. Phylogeography of the cactophilic species *Drosophila gouveai*: demographic events and divergence timing in dry vegetation enclaves in eastern Brazil. Journal of Biogeography, v.36, p.2136-2147, 2009.

NAPP, M., BRNCIC, D. Eletrophoretic variability in two closely related brazilian species of the *flavopilosa* species group of *Drosophila*. Brazilian Journal of Genetics. 1978; 1:1-10

NYLANDER, J. A. A. MrModeltest v2. Program distributed by the author. Evolutionary Biology Center, Uppsala University 2004.

PFEILER, E., et al. Genetic differentiation and demographic history in *Drosophila pachea* from the Sonoran Desert. Hereditas 144:63_74, 2007.

RAMBAUT, A. DRUMMOND, A. J. Tracer v1.5. 2009. Available from http://beast.bio.ed.ac.uk/Tracer (Accessed dezember, 2016).

ROBE L. J. et al. The *Drosophila flavopilosa* species group (Diptera, Drosophilidae): An Array of exciting questions. Fly, 7:59 – 69, 2013.

SANTOS, R. C. O, VILELA, C. R. Breeding sites of Neotropical Drosophilidae (Díptera). Living and fallen flowers of *Sessea brasiliensis* and *Cestrum* spp. (Solanaceae). Revista Brasileira de Entomologia; 49:544-551, 2005.

SEPEL, L. M. N. et al. Seasonal fluctuations of *D. cestri* and *D. incompta*, two species of the *flavopilosa* group. DIS; 83:122-126, 2000.

SIMON, C. et al. Evolution, weighting and phylogenetic utility of mitochondrial genes sequences and a compilation of conserved polymerase chain reaction primers. Annals of the Entomological Society of America 87: 651–701, 1994.

STADEN, R. The Staden Sequence Analysis Package. Molecular Biotechnology 5, 233-241, 1996.

SWOFFORD, D.L. PAUP: Phylogenetic Analysis using Parsimony (and other methods). Version 4. Sinauer Associates, Massachusetts, 2003.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595. 1989.

TAMURA, K. et al. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution, 2011.

TEMPLETON, A. R. Population Genetics and Microevolutionary Theory. John Wiley & Sons, New Jersey, 705 pp, 2006.

WHEELER, M. R., TAKADA, H., BRNCIC, D. The *flavopilosa* species group of *Drosophila*. Studies in Genetic II. Univ Texas Publ; 6 205:396-412, 1962.

6 CONCLUSÕES GERAIS E PERSPECTIVAS

Drosophila incompta pertence ao grupo *flavopilosa* de *Drosophila,* e se caracteriza por apresentar padrões ecológicos bastante restritos, sendo adaptada a exploração de flores de *Cestrum* como única fonte de recursos alimentares e sítios de oviposição. Esse padrão ecológico especializado foi claramente evidenciado em nossas coletas de *Cestrum*, que foram utilizados como meios para a obtenção dos exemplares de *D. incompta*, espécie que não pode ser mantida em laboratório. Os indivíduos coletados e identificados foram divididos em dois grupos no momento da extração do DNA: (1) No primeiro grupo, a extração foi feita a partir de um conjunto de indivíduos pertencentes a uma mesma população, e esse DNA foi utilizado para o sequenciamento do genoma da espécie. A montagem e anotação destes reads permitiu não apenas a caracterização completa do genoma mitocondrial de *D. incompta* (Artigo I), como também um melhor entendimento dos processos evolutivos subjacentes ao posicionamento filogenético da espécie (Artigo II) e das adaptações moleculares associadas à sua especialização ecológica (ArtigoIII). (2) No segundo grupo, a extração do DNA foi feita individualmente de cada espécime, e estes foram utilizados no sequenciamento de marcadores mitocondriais (genes citocromo oxidase c subunidades I e II – *COI* e *COII*) e nucleares (gene *Hunchback – Hb*), utilizados em uma abordagem filogeográfica (Artigo IV).

- *Capítulo II*

O capítulo II desta tese refere-se à caracterização e anotação do genoma mitocondrial completo de *D. incompta.* Esse genoma apresenta um total de 15.641 pares de bases, sendo constituído por 13 genes codificadores de proteínas, 22 genes de *tRNAs*, 2 genes de *rRNAs* e uma região rica em A+T. A organização destes genes está em perfeita sintenia com os outros genomas mitocondriais publicados de *Drosophila.* O resultado mais interessante e surpreendente, nesse caso, foi o grau de polimorfismo encontrado ao longo desse genoma, revelando níveis pronunciados de variação intra-populacional. Entretanto, como os níveis de diversidade nucleotídica variaram entre diferentes regiões do genoma, a escolha do marcador mitocondrial mais adequado para estudos de estrutura populacional se revelou um importante ponto a ser considerado.

- *Capítulo III*

O capítulo III mostra que há incongruências entre os conjuntos de dados mitocondriais e nucleares no que diz respeito ao posicionamento de *D. incompta* dentro do gênero *Drosophila.* Neste sentido, a árvore de concordância primária recuperada através da análise de concordância bayesiana (BCA) para os 13 genes mitocondriais codificadores de proteínas agrupou *D. incompta* e *D. mojavensis*, com fator de concordância (FC) de 0.58, o que revela que apenas 58% dos genes mitocondriais (que devem compartilhar uma história única) suportam a informação de que essas duas espécies são irmãs. Os fatores de concordância gerados na BCA foram maiores para os genes nucleares, que recuperaram *D. incompta* como sendo espécie irmã de *D. virilis*, com FC de 0.98, de modo que 98% dos genes nucleares contam a mesma história evolutiva para este clado. Embora nossos resultados sugiram um efeito importante da saturação nas inconsistências recuperadas pelos dados mitocondriais e nucleares, defendemos, neste capítulo, que a incongruência entre esses dois conjuntos de dados é provavelmente derivada das histórias evolutivas independentes apresentadas pelos dois conjuntos de genes. Neste caso, como a maioria dos genes nucleares suportam a mesma resolução topológica no que diz respeito ao posicionamento de *D. incompta*, concluímos que esta é a verdadeira história da espécie. Acreditamos que história evolutiva diferente apresentada pelos genes mitocondriais pode ser um artefato da introgressão do genoma mitocondrial entre os ancestrais dos grupos *flavopilosa* e *repleta*. Outro resultado não convencional apresentado neste capítulo diz respeito ao posicionamento filogenético de *D. willistoni*, comumente alocada como basal dentro do gênero ou parte do subgênero *Drosophila.* Tal resultado pode ser atribuído ao viés composicional dessa espécie e também a atração de ramos longos.

- *Capítulo IV*

Neste capítulo, com base em sequências dos genes mitocondriais *COI* e *COII*, foi possível evidenciar que as populações de *D. incompta* da região Sul do Brasil passaram por um evento de expansão populacional entre 175 e 100 mil anos atrás seguido de um longo período de estabilidade demográfica. Além disso, resultados obtidos para o gene nuclear Hb permitem especular que esse evento tenha sido precedido por uma redução do tamanho populacional. Embora ambos os eventos devam estar correlacionados às oscilações climáticas do Quaternário, essa espécie parece não ter sido amplamente afetada pelo último máximo glacial. Como no Brasil, as plantas de *Cestrum* distribuem-se tanto no Bioma Mata Atlântica

quanto no Cerrado, acredita-se que as alterações de clima e vegetação típicas deste período não tenham influenciado na sua distribuição e, consequentemente, na dinâmica populacional desses drosofilídeos de ecologia restrita. Por fim, embora *D. incompta* tenha apresentado níveis consideráveis de diversidade genética, não há evidencias de um padrão geográfico bem definido, o que pode indicar a ocorrência de fluxo gênico entre populações como resposta as situações de limitação de recursos. Como esta necessidade de migração se impõe de forma diferente para diferentes localidades, é possível que *D. incompta* apresente um contingente de populações com hábitos migratórios diversos.

- *Capítulo V*

No capítulo V abordamos o tamanho do repertório gênico de duas famílias de quimioreceptores no genoma de *D. incompta*: receptores olfativos (*OR*) e receptores gustativos (*GR*). Além disso, buscamos compreender qual é a principal forma de seleção atuante nesses genes ao longo do genoma. De uma forma geral, podemos concluir que há a presença de pelo menos 28 ORs e apenas 12 Grs no genoma de *D. incompta*. Provavelmente, esse limitado número de genes, em especial para os Grs, está relacionado ao padrão de ecologia restrita já conhecido para a espécie, uma vez que esses genes desempenham um papel importante no processo de especialização à planta hospedeira. Afinal, como o recurso explorado por *D. incompta* é exclusivo, não parece surpreendente que essa espécie apresente um número de genes quimiorreceptores reduzido com relação a espécies generalistas que exploram diferentes sítios de oviposição e alimentação. Além disso, esses genes parecem estar sob efeito de seleção purificadora. Entretanto, análises complementares mais robustas são necessárias para confirmar esse resultado, facilitando assim, o melhor entendimento de como a seleção natural está atuando ao longo desses dois conjuntos gênicos, que auxiliam na adaptação ecológica em *D. incompta.*

As nossas perspectivas estão voltadas aos capítulos IV e V. Em relação ao capítulo IV, de filogeografia, sequências adicionais serão incluídas tanto para o gene *COI*, quanto para o gene *Hb*. A partir da atualização dos alinhamentos com as novas sequências, as análises serão feitas novamente, dessa vez com o tamanho amostral ampliado e compatível com uma boa publicação em revista científica. Considerando o capítulo V, análises mais robustas serão executadas para complementar as análises prévias apresentadas nessa tese, a fim de confirmar os resultados encontrados ou, ainda, encontrar outros igualmente interessantes. Nesse caso, as sequências nucleotídicas ortólogas obtidas em cada um dos genomas serão utilizadas na

obtenção da árvore de cada um dos genes com o uso do programa MrBayes 3.1.2 (Huelsenbeck e Ronquist, 2001). Estas árvores serão, então, utilizadas para detectar seleção natural em posições individuais dos códons através do pacote ADAPTSITE (Suzuki et al. 2001), que avalia as taxas de substituição não-sinônima por sítio não-sinônimo (rN) e as taxas de substituição sinônima por sítio sinônimo (rS) através da reconstrução de estados ancestrais ao longo dos nós internos da filogenia. Os processos evolutivos atuantes ao longo de cada códon dos genes escolhidos serão também avaliados com o uso da filogenia obtida e do programa CODEML, encontrado no pacote PAML (Yang, 1997). Neste caso, a verossimilhança dos modelos "invariante" (Goldman e Yang, 1994) (com um único parâmetro dN/dS), "neutro" (Nielsen e Yang, 1998) (com parâmetros p1 – proporção de sítios neutros – e p2 – proporção de sítios que apresentam seleção negativa), e de "seleção positiva" (Nielsen & Yang, 1998) (com parâmetro p3 – proporção de sítios que apresentam seleção positiva, além dos parâmetros p1 e p2) será analisada. Finalmente, sítios que apresentam evolução neutra, seleção negativa ou positiva serão individualmente identificados com o uso de um enfoque Bayesiano empírico. Neste caso, as modificações nos sítios sob seleção positiva serão situadas ao longo da filogenia de máxima verossimilhança, através das reconstruções de maior parcimônia obtidas com o auxílio do MacClade 3.0 (Maddison e Maddison 1992) e com ênfase na avaliação das autapomorfias apresentadas por *D. incompta.*

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRIANOV, B. et al. Comparative analysis of the mitochondrial genomes in *Drosophila virilis* species group (Diptera: Drosophilidae). **Trends in Evolutionary Biology,** 2(1), e4, 2010.

AVISE, J. C., **Phylogeography:** The History and Formation of Species. Harvard University Press, Cambridge, MA., 447 pp., 2000.

AVISE, J. C., et al. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. **Annual review of ecology and systematics**, 489-522, 1987.

AVISE, J. C., **Molecular Markers, Natural History and Evolution.** Springer US, 1994.

BÄCHLI, G. **TaxoDros: The Database on Taxonomy of Drosophilidae**, v. 1.03, Database 2009/04. http://taxodros.unizh.ch/. Last accessed on 26/09/2015.

BÄCHLI, G. **TaxoDros: The Database on Taxonomy of Drosophilidae**, v. 1.03, Database 2009/04. http://taxodros.unizh.ch/. Last accessed on 01/03/2016.

BALLARD, J. W. O. Comparative genomics of mitochondrial DNA in *Drosophila simulans*. **Journal of Molecular Evolution,** v. 51(1), 64-75, 2000.

BALLARD, J. W. O. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. **Journal of Molecular Evolution,** v. 51(1), 48-63, 2000.

BEHEREGARAY, L.B. Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere**. Molecular Ecology**. 17:3754-3574, 2008.

BOORE, J. L. Animal mitochondrial genomes. **Nucleic Acids Research** 27 (8), 1767-1780, 1999.

BRISSON, J.A. et al. Abdominal pigmentation variation in *Drosophila polymorpha*: geographic variation in the trait, and underlying phylogeography. **Evolution**, v.59(5), p.1046-1059, 2005.

BRNCIC D. Chromosomal structure of populations pf *Drosophila flavopilosa* studied in larvae collected in their natural breeding sites**. Chromosoma**; 13:183-195, 1962.

BRNCIC, D. Ecological and cytogenetic studies of *Drosophila flavopilosa*, a Neotropical species living in *Cestrum* flowers**. Evolution**; 20:16-29, 1966.

BRNCIC D. Chromosomal polymorphism in an ecologically restricted species of *Drosophila* living in Chile. **Cien e Cul**. 19:45-53, 1967.

BRNCIC D. The effects of temperature on chromosomal polymorphism of *Drosophila flavopilosa* larvae. **Genetics**. 59:427-432, 1968.

BRNCIC D. The *flavopilosa* group of species as an example of flower-breeding species. In: Ashburner M, Carson HL e Thompson JN (eds) **The genetics and biology of *Drosophila*.** V 3d., New York, Academic Press. 360-377, 1983.

BRODY, T. The Interactive Fly: gene networks, development and the Internet. **Trends in genetics**, v. 15, n. 8, p. 333-334, 1999.

CAMERON, S. L. Insect mitochondrial genomics: implications for evolution and phylogeny. **Annual Review of Entomology**, 59, 95-117, 2014.

CLARK, A. G. et al. Evolution of genes and genomes on the *Drosophila* phylogeny. **Nature.** 450:203–218, 2007.

CLARY, D. O., WOLSTENHOLME, D. R. The mitochondrial DNA molecule of Drosophila yakuba: nucleotide sequence, gene organization, and genetic code **Journal of Molecular Evolution,** 22(3), 252-271, 1985.

CLYNE, P. J., et al. A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in Drosophila. **Neuron** 22.2: 327-338, 1999.

De BRITO, R. A., MANFRIN, M. H., SENE, F. M. Nested cladistic analysis of Brazilian population of *Drosophila serido.* **Molecular Phylogenetics and Evolution**. 22:11-143, 2002a.

De BRITO, R. A.; MANFRIN, M.H.; SENE, F.M. Mitochondrial DNA phylogeography of Brazilian populations of *Drosophila buzzatii*. **Genetics and Molecular Biology**, v.25(2), p.161-171, 2002b.

DE RÉ, et al. Brazilian populations of *Drosophila maculifrons* (Diptera: Drosophilidae): low diversity levels and signals of a population expansion after the Last Glacial Maximum. **Biological Journal of the Linnean Society**, v. 112, n. 1, p. 55-66, 2014.

FRANCO, F. F., SENE, F. M.,MANFRIN, M. H. Low satellite DNA variability in natural populations of *Drosophila antonietae* involved in different evolutionary events. Journal of **Heredity** v. 101, p. 650-656, 2010.

FRANCO, F. F. et al. Intra- and interspecific divergence in the nuclear sequences of the clock gene *period* in species of the *Drosophila buzzatii cluster*. **Journal of Zoological Systematics and Evolutionary Research** 33, 225-223, 2010.

FRANCO, F. F., MANFRIN, M. H. Recent demographic history of cactophilic *Drosophila* species can be related to Quaternary palaeoclimatic changes in South America. Journal of Biogeography, 40: 142–154, 2013.

GARCIA, C. F. et al. Drosophilid assemblages at different urbanization levels in the city of Porto Alegre, state of Rio Grande do Sul, Southern Brazil. **Neotropical Entomology** 40(1): 32-41, 2012.

GARESSE, R. *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. **Genetics,** 118(4), 649-663, 1988.

GILBERT, J. M., PERONNET, F., SCHLOTTERER, C. Phenotypic plasticity in *Drosophila* pigmentation caused by temperature sensitivity of a chromatin regulator network. **PLoS Genetics** 3: e 30, 2007.

GOTTSCHALK, M. S., HOFMANN, P. R. P., VALENTE, V. L. S. Diptera, Drosophilidae: historical occurrence in Brazil. **Check List,** vol. 4, p. 485–518, 2008.

GRIMALDI, D. A. A phylogenetic, revised classification of the genera in the Drosophilidae (Diptera). **Bulletin of the American Museum of Natural History.** 197:1-139, 1990.

GRIMALDI, D., ENGEL, M., S. **Evolution of the Insects**. Cambridge University Press, 2005.

GUSTANI, E. C. et al. Demographic Structure and Evolutionary History of *Drosophila ornatifrons* (Diptera, Drosophilidae) from Atlantic Forest of Southern Brazil. **Zoological science**, v. 32, n. 2, p. 141-150, 2015.

HALLEM, E. A., CARLSON, J. R. The odor coding system of *Drosophila*. **Trends in Genetics,** *20*(9), 453-459, 2004.

HEBERT, P. D. N., RATNASINGHAM, S., DE WAARD J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. **Proceedings of the Royal Society of London.** Series B 7: 270 (Suppl 1) S96–S99, 2003b.

HEWITT, G. M. The genetic legacy of the Quaternary ice ages. **Nature** 405,907–913, 2000.

HOFMANN, P. R. P. Variabilidade genética em espécies de nível ecológico restrito. Ciência e Cultura. 37:579-581, 1985.

JONES, C. D. The genetics of adaptation in *Drosophila sechellia*. **Genetica** 123: 137–145, 2005.

KELLEHER, E. S., MARKOW, A. T. Duplication, selection and gene conversion in a Drosophila mojavensis female reproductive protein family. **Genetics** 181.4: 1451-1465, 2009.

LANG, B. F., GRAY, M. W., BURGER, G. Mitochondrial genome evolution and the origin of eukaryotes. Annual Review of Genetics. 33, 351– 397, 1999.

LAVROV, D. V. Mitochondrial Genomes in Invertebrate Animals. **In Molecular Life Sciences** (pp. 1-8). Springer, New York, 2014.

LUDWIG, A., et al. *Drosophila incompta* development without flowers. **DIS**; 85:40-41, 2002.

MATZKIN, L. M. et al. Functional genomics of cactus host shifts in Drosophila mojavensis. **Molecular Ecology**, v. 15, n. 14, p. 4635-4643, 2006.

MATZKIN, L. M. Ecological genomics of host shifts in *Drosophila mojavensis*. **Advances in Experimental Medicine and Biology**;781:233-247, 2014.

MANFRIN M. H., BRITO R. O. A., SENE F. M. Systematics and evolution of the *Drosophila buzzatii* (Diptera, Drosophilidade) cluster using mtDNA. **Annals of the Entomological Society of America,** 94:333-346, 2001.

MANFRIN, M. H., SENE, F. M. Cactophilic *Drosophila* in South America: a model fot evolutionary studies. **Genetica** 57-75, 2006.

MATUTE, D. M., et al. Temperature-based extrinsic reproductive isolation in two species of *Drosophila*. **Evolution:** 63*:*595–612, 2009.

MCBRIDE, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. **Proceedings of the National Academy of Sciences of the United States of America**; 104 (12):4996-5001, 2007.

MONTOOTH, K. L. et al. Comparative genomics of *Drosophila* mtDNA: novel features of conservation and change across functional domains and lineages**. Journal of Molecular Evolution**, 69 (1), 94-114, 2009.

MORAES, E. M., SENE, F. M. Microsatellite and morphometric variation in *Drosophila gouveai:* the relative importance of historical and current factors in shaping the genetic population structure. **Journal of Zoological Systematics and Evolutionary Research** 45, 336–344, 2007.

MORAES, E.M. et al. Phylogeography of the cactophilic species *Drosophila gouveai*: demographic events and divergence timing in dry vegetation enclaves in eastern Brazil. **Journal of Biogeography**, v.36, p.2136-2147, 2009.

NAPP, M., BRNCIC, D. Eletrophoretic variability in two closely related brazilian species of the *flavopilosa* species group of *Drosophila*. **Brazilian Journal of Genetics**. 1978; 1:1-10.

PEREIRA, M. A. Q. R., VILELA, C. R., SENE, F. M. Notes on breeding and feeding sites of some species of the *repleta* group of the genus *Drosophila* (Diptera, Drosophilidae). **Ciência e Cultura** 35(9): 1313–1319, 1983.

POLLARD, D. A. et al. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. **PLoS Genetics**, v. 2, n. 10, p. e173, 2006.

POPPE, J. L., VALENTE, V. L., SCHMITZ, H. J. Structure of Drosophilidae assemblage (Insecta, Diptera) in Pampa Biome (São Luiz Gonzaga/RS). **Papéis Avulsos de Zoologia.** 52(16):185-195, 2012.

POPPE, J. L. et al. High diversity of Drosophilidae (Insecta, Diptera) in the Pampas Biome of South America, with descriptions of new *Rhinoleucophenga* species. **Zootaxa** 3779(2):215–245, 2014.

POWELL, J. R. **Progress and Prospects in Evolutionary Biology: The *Drosophila* Model**, Oxford Univ. Press, New York.

REMSEN, J., O'GRADY P. O. Phylogeny of Drosophilinae (Diptera: Drosopilidae) with comments on combined analysis and character support**. Molecular Phylogenetics and Evolution,** 24:249-264, 2002.

ROBE, L.J. et al. Molecular phylogeny of the subgenus *Drosophila* (Diptera, Drosophilidae) with an emphasis on Neotropical species and groups: a nuclear versus mitochondrial gene approach. **Molecular Phylogenetics and Evolution.** 36:623-640. 2005.

ROBE, L.J., LORETO, E.L.S., VALENTE, V.L.S. Radiation of the *Drosophila* subgenus (Drosophilidae, Diptera) in the Neotropics. **Journal of Zoological Systematics and Evolutionary Research.** DOI 10.1007/s10709-009-9432-5, 2010a.

ROBE, L.J., VALENTE, V.L.S.; LORETO, E.L.S. Phylogenetic relationships and macroevolutonary patterns within the *Drosophila tripunctata* "radiation" (Diptera: Drosophilidae). **Genetica**, v.138, p.725-735, 2010b.

ROBE L. J., et al. The *Drosophila flavopilosa* species group (Diptera, Drosophilidae): An Array of exciting questions. **Fly**, 7:59 – 69, 2013.

ROBERTSON, H. M., WARR, C. G., CARLSON, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. **Proceedings of the National Academy of Sciences,** 100 (suppl 2), 14537-14542, 2003.

ROE, A., SPERLING, F. A. H. Patterns of evolution of mitochondrial cytochrome *c* oxidase I e II DNA and implications for DNA barcoding. **Molecular Phylogenetics and Evolution**, vol. 44, p. 325-345, 2007.

ROKAS, A. et al. Genome-scale approaches to resolving incongruence in molecular phylogenies. **Nature**, v. 425, n. 6960, p. 798-804, 2003.

ROQUE, F., MATA, R. A., TIDON, R. Temporal and vertical drosophilid (Insecta; Diptera) assemblage fluctuations in a neotropical gallery forest. Biodiversity and Conservation. 22, 657-672, 2013.

RUBIN, C. J. et al. Whole-genome resequencing reveals loci under selection during chicken domestication. **Nature**, v. 464, n. 7288, p. 587-591, 2010.

SAMBANDAN, D. et al. Dynamic genetic interactions determine odor-guided behavior in *Drosophila melanogaster.* **Genetics** 174*:* 1349–1363, 2006.

SANTOS, R. C. O, VILELA, C. R. Breeding sites of Neotropical Drosophilidae (Díptera). Living and fallen flowers of *Sessea brasiliensis* and *Cestrum* spp. (Solanaceae). Revista Brasileira de Entomologia; 49:544-551, 2005.

SMITH, C.D. et al. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). **Proceedings of the National Academy of Sciences of the United States of America** 108: 5673-5678, 2011a.

SEPEL, L. M. N. et al. Seasonal fluctuations of *D. cestri* and *D. incompta*, two species of the *flavopilosa* group**. DIS;** 83:122-126, 2000.

SIMON, C. et al. Evolution, weighting and phylogenetic utility of mitochondrial genes sequences and a compilation of conserved polymerase chain reaction primers. **Annals of the Entomological Society of America** 87: 651–701, 1994.

TAUTZ D. et al. Finger protein of novel structure encoded by *hunchback*, a second member of the gap class of *Drosophila* segmentation genes. **Nature** 327:383-389, 1987.

THROCKMORTON, L. H. The phylogeny, ecology and geography of *Drosophila*. In: King, R. C. (ed) **Handbook of Genetics**. Plenun, New York, pp 421-469. 1975.

TIDON, R., SENE, F. M. A trap that retains and keeps *Drosophila* alive. **Drosophila Information Service** 672: 89. 1988.

TODA, M. J. **DrosWLD-Species: Taxonomic Information Database for World Species of Drosophilidae**. http://bioinfo.lowtem.hokudai.ac.jp/db/modules/stdb/. Last accesed on 15/08/2015.

TORRES, T. T. et al. Expression profiling of Drosophila mitochondrial genes via deep mRNA sequencing. Nucleic Acids Res gkp856, 2009.

TURCHETTO-ZOLET, et al. **Guia prático para estudos filogeográficos**, 2013.

ZHANG, De-Xing; HEWITT, G. M. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. **Molecular ecology**, v. 12, n. 3, p. 563-584, 2003.

WANNER, K. W. et al. Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, Bombyx mori. **Insect molecular biology** 16.1: pp. 107-119, 2007.

WHEELER, M. R., TAKADA, H., BRNCIC, D. The *flavopilosa* species group of *Drosophila*. **Studies in Genetic II**. Univ Texas Publ; 6 205:396-412, 1962.

WHITWORTH, T. et al., DNA barcoding reliably identify species of the blowfly genus *Protocalliphora*. **Proceedings of the Royal Society**, vol. 274, p. 1731-1739, 2007.

WITTKOPP, P. J., BELDADE, P. Development and evolution of insect pigmentation: genetic mechanisms and the potential consequences of pleiotropy. **Seminars in Cell & Developmental Biology**, 20, 65–71, 2009.

WITTKOPP, P. J. et al. Local adaptation for body color in *Drosophila americana*. **Heredity**: [Epub ahead of print], july 2010.