Universidade Federal de São Carlos
Centro de Ciências Biológicas e da Saúde
Departamento de Genética e Evolução
Laboratório de Genética de Populações e Evolução

COMPARAÇÃO DE TRANSCRIPTOMAS POR SEQUENCIAMENTO DE PRÓXIMA GERAÇÃO EM TECIDOS DE CABEÇA DE DUAS ESPÉCIES DE MOSCAS-DAS-FRUTAS, *Anastrepha fraterculus* e *Anastrepha obliqua*

Candidato: Victor Borges Rezende

Orientador: Prof. Dr. Reinaldo Alves de Brito

São Carlos-SP

2014

VICTOR BORGES REZENDE

COMPARAÇÃO DE TRANSCRIPTOMAS POR SEQUENCIAMENTO DE PRÓXIMA GERAÇÃO EM TECIDOS DE CABEÇA DE DUAS ESPÉCIES DE MOSCAS-DAS-FRUTAS, *Anastrepha fraterculus* e *Anastrepha obliqua*

Dissertação apresentada ao Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do Título de Mestre em Genética e Evolução

Orientador: Prof. Dr. Reinaldo Alves de Brito

SÃO CARLOS-SP

2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

R467ct

Rezende, Victor Borges.
Comparação de transcriptomas por sequenciamento de próxima geração em tecidos de cabeça de duas espécies de moscas-das-frutas, *Anastrepha fratercules e Anastrepha obliqua* / Victor Borges Rezende. -- São Carlos : UFSCar, 2015.
72 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Genética e evolução. 2. Anastrepha. 3. Transcriptoma. 4. Expressão gênica. I. Título.

CDD: 575 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA MOLECULAR

"COMPARAÇÃO DE TRANSCRIPTOMAS POR SEQUENCIAMENTO DE PRÓXIMA GERAÇÃO EM TECIDOS DE CABEÇA DE DUAS ESPÉCIES DE MOSCAS-DAS FRUTAS, *Anastrepha fraterculus* e *Anastrepha obliqua*"

Dissertação de Mestrado de

VICTOR BORGES REZENDE

Banca Examinadora

Prof. Dr. Reinaldo Alves de Brito

.....

Profa. Dra. Sónia Cristina da Silva Andrade

.....

Prof. Dr. Guilherme Targino Valente

.....

“For millions of years, mankind lived just like the animals.
Then something happened which unleashed the power of our imagination.
We learned to talk.”

Stephen Hawking (1993)

AGRADECIMENTOS

Agradeço profundamente ao orientador Prof. Dr. Reinaldo Alves de Brito, pela confiança depositada, conhecimento compartilhado, dedicação, incentivo e amizade em todos os momentos.

Aos meus pais, Rosângela, Hamilton e Marinete, que sempre me apoiaram em todas as minhas decisões e sempre estão presente nas horas que eu mais preciso. Aos meus irmãos, Igor e Aline, obrigado por compartilharem sua amizade e cumplicidade sempre.

À minha noiva Natália que sempre esteve ao meu lado em todos os momentos. Obrigado pelo companheirismo, apoio e companhia nesses quase 6 anos juntos. Cada conquista pessoal minha está diretamente relacionada à felicidade que é viver ao seu lado.

Aos pós-doutorandos Iderval e Samira, Felipe e Mário pela assistência intelectual e amizade sempre presente na vizinhança.

Aos amigos de laboratório Aline, André, Carlos, Cris, Emeline, Baratinha, Janaína, Lívia, Manu, Nancy, Paulo e Andréa pelas incontáveis e gratas horas que passamos juntos nesta etapa de nossas vidas. Por terem compartilhado comigo um pouco de cada um de vocês tornando até os momentos mais difíceis um bom motivo para boas gargalhadas.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos e à “mãe” FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo - por suporte financeiro que subsidia grande parte da infraestrutura e material utilizado no laboratório.

Ao programa de Pós-graduação em Genética Evolutiva e Biologia Molecular da UFSCar pelo oferecimento do curso.

SUMÁRIO

LISTA DE TABELAS

CAPÍTULO 3 – MANUSCRITO 1

CAPÍTULO 4 – MANUSCRITO 2

LISTA DE FIGURAS

CAPÍTULO 3 – MANUSCRITO 1

CAPÍTULO 4 – MANUSCRITO 2

LISTA DE ANEXOS

RESUMO

Investigamos os padrões de expressão gênica em tecidos cefálicos de duas espécies de moscas-das-frutas do gênero *Anastrepha* (Diptera: Tephritidea), *A. fraterculus* e *A. obliqua*, proximamente relacionadas, identificando SNPs com alto grau de diferenciação entre as espécies e realizando análises evolutivas com o objetivo de encontrar genes candidatos relacionados ao recente processo de separação dessas espécies. Para isso, utilizamos as novas tecnologias de sequenciamento em larga escala (NGS) com a metodologia de RNA-Seq em tecidos cefálicos das duas espécies de moscas em diferentes fases da vida reprodutiva dos dois sexos. Após processamento e retirada das sequências com baixa qualidade utilizamos mais de 140 milhões de sequências paired-end para montarmos um transcriptoma conjunto das espécies, com mais de 154 mil contigs, e outros dois separados por espécie. Estes resultados estão apresentados em dois manuscritos distintos, um primeiro que descreve as bibliotecas produzidas para as diferentes espécies e um segundo que investiga padrões de expressão e genes envolvidos na diferenciação das espécies. Estes dados revelaram 1991 genes com expressão diferencial em pelo menos uma comparação de fase de vida reprodutiva entre as espécies, sendo que encontramos duas vezes mais genes com diferença na expressão entre as espécies em machos do que em fêmeas. Diversos destes genes foram associados a genes relacionados ao olfato, como a família gênica das *Obp*s (*odorant-binding protein*) e o gene *visgun*, o que pode indicar uma mudança comportamental na preferência por alimento, parceiros ou sítios para cópula e oviposição. Encontramos também dois conjuntos de genes que são diferencialmente expressos entre as espécies, sendo um conjunto super-expresso em *A. obliqua* e sub-expresso em *A. fraterculus* e outro conjunto com um padrão revertido. Análises de padrão de seleção nestes genes sugerem sete que apresentam indícios de seleção positiva e ao menos um SNP com altos índices de diferenciação entre as espécies.

PALAVRAS-CHAVE: *Anastrepha fraterculus*; *Anastrepha obliqua*; Transcriptoma; Next-Generation Sequencing; Expressão Gênica Diferencial; SNP Calling.

ABSTRACT

We studied patterns of gene expression in two closely related species of fruit flies of the genus *Anastrepha* (Diptera: Tephritidea), *A. fraterculus* and *A. obliqua*, with the goal of finding candidate genes related to the recent differentiation process between these species. In order to do this, we used the Next-generation sequencing (NGS) with RNA-Seq methodology in head tissues of the two species of flies at different stages of the reproductive life for both sexes. After processing and removal of low quality reads we retained over 140 million paired-end reads. These sequences were assembled into individual transcriptomes for each species and a pooled transcriptome, with 154,787 contigs, representing both species. Based on the results of the assemblies, annotation and mapping we prepared two separate manuscripts, one describing the libraries for each species and a combined analysis and a second that investigate the contigs involved with species differences and their patterns of expression. These data reveal 1991 genes with differential expression in at least one comparison among different reproductive stages. It is noteworthy that we observed twice as many genes with differential expression when contrasting males than with females. Several of these genes were associated to odour, such as *Odorant Binding proteins* and *visgum*, suggesting that behavioral changes in food sources, mating choice, or breeding and oviposition sites might be involved with species differences.We also identified two large sets of genes that are differentially expressed between the species, one being under-expressed in *A. obliqua* and overexpressed in *A. fraterculus* and the other that had a reverse pattern. We used a differentiation index of SNPs per contig ($\overline{D}$) and pairwise tests of positive selection between the sequences, and analysis of the substitution amino acid type caused by SNPs to identify 7 genes there are candidates to be related to the speciation process between *A. fraterculus* and *A. obliqua*.

KEY-WORDS: *Anastrepha fraterculus*; *Anastrepha obliqua*; Transcriptome; Next-Generation Sequencing; Differential Gene Expression; SNP Calling.

PREFÁCIO

Anastrepha fraterculus e *Anastrepha obliqua* são espécies proximamente relacionadas pertencentes ao grupo *fraterculus* (Diptera, Tephritidae) que são importantes pragas da cultura de frutos carnosos por causarem um enorme prejuízo econômico. Estudos que ajudem a diferenciar estas espécies são de grande valia para os setores da agricultura lesados com as atividades dessas pragas.

Visando verificar a hipótese de que estas duas espécies do gênero *Anastrepha* divergiram recentemente, porém já possuem caracteres morfológicos e comportamentais que as separam como espécies, o presente estudo apresenta dois manuscritos que descrevem uma enorme quantidade de dados de sequenciamento e busca auxiliar na elucidação da diferenciação entre *A. fraterculus* e *A. obliqua*.

Os dois manuscritos são precedidos por uma breve introdução, no capítulo 1, que tem a função de introduzir tanto os assuntos comuns aos dois manuscritos, quanto as teorias especificas que baseiam as discussões de cada um.

O capítulo 3 é o manuscrito do artigo científico intitulado "**Characterization of Head Transcriptomes in Different Life Stages and Analysis of Gene Expression of Two Species of *Anastrepha* Fruit Flies (Tephritidae, Diptera)**" e descreve os resultados do sequenciamento de próxima geração (NGS) do tecido cefálico de *A. fraterculus* e *A. obliqua*. Este artigo também detalha o tratamento realizado por *softwares* nos dados digitais e resultados da análise de expressão gênica diferenciais entre as espécies. Os resultados de sequenciamento, *assembly*, mapeamento e anotação são os mesmos utilizados para a análise evolutiva realizada no manuscrito 2.

Com o título de "**Candidate genes involved in differentiation between two non-model species of fruit flies (*Anastrepha,* Tephritidae) screened from head transcriptomes**", o manuscrito apresentado no Capítulo 4 utiliza grande parte dos dados descritos no primeiro manuscrito em uma análise evolutiva que tem o objetivo de selecionar genes candidatos envolvidos no processo de diferenciação entre *A. fraterculus* e *A. obliqua*. Para alcançar este objetivo, utilizamos uma metodologia de busca de SNP (Single Nucleotide Polimorphism) que apresentaram um alto nível de polimorfismo entre as espécies e um teste evolutivo que é capaz de identificar seleção positiva por substituição de aminoácido nas sequências de nucleotídeos

alinhadas par a par.

Como os manuscritos são derivados de um mesmo conjunto de dados, pode se dizer que os dois capítulos se complementam. A fim de sobrepor os resultados, o Capítulo 5 elabora "Considerações Finais" que consta de uma sinopse de assuntos relevantes aos dois capítulos, inclusive de sugestões para estudos futuros. Com o intuito de ilustrar e detalhar algumas das principais considerações finais, na última sessão do trabalho apresentamos tabelas e figuras em anexo necessários para a realização de ambos os artigos.

Como os capítulos 3 e 4 dizem respeito a manuscritos a serem submetidos à publicação independentemente, estes são acompanhados das suas respectivas referências bibliográficas já devidamente formatadas para os periódicos a serem submetidos, o que deixa a dissertação com três listas de referência. Por outro lado, a fim de promover uma leitura mais fluida e simples, as tabelas e figuras foram mantidas no decorrer dos textos, ficando assim, fora dos padrões dos periódicos.

CAPÍTULO 1 – INTRODUÇÃO

1.1 – Especiação e genes candidatos

Eventos de surgimento de novas espécies no planeta Terra são raros em comparação a eventos de desaparecimento de espécies (MAY, 1988), mas ao longo dos últimos 3,5 bilhões de anos, desde o aparecimento da vida, processos de especiação e de extinção têm modelado a diversidade da vida na Terra. Alguns autores definem a especiação como sendo o processo evolutivo de formação de novas espécies ocasionado pelo acúmulo de diferenças genéticas de modo que haja o estabelecimento de uma barreira contra o fluxo gênico (WU; TING, 2004). A causa mais comum de isolamento reprodutivo, ou seja, a impossibilidade de troca de genes entre dois organismos são as barreiras à fecundação entre as espécies. Assim processos que levam ao isolamento reprodutivo são mais prováveis de levar a especiação e a ausência de fluxo gênico facilitaria tal processo (BUTLIN; RITCHIE, 2001).

No entanto, o processo de especiação não é um evento que dependa necessariamente de um isolamento genético para ocorrer, uma vez que não é um evento exclusivamente associado à ocorrência de fluxo gênico. A diferenciação genética que ocorre entre populações e resulta na especiação está principalmente associada à seleção natural, à deriva genética, e ao balanço dessas duas forças, embora todas as outras forças evolutivas possam causar diferenciação genética e levar a especiação (TEMPLETON, 1989).

Um importante passo para se entender o processo de especiação é a identificação e o isolamento de genes que causem o isolamento reprodutivo. Estudos dessa natureza já têm sido desenvolvidos há bastante tempo (COYNE, 2004), e têm sido bem sucedidos em identificar genes e regiões no genoma envolvidos com a diferença entre espécies. Em função do isolamento reprodutivo, as espécies deixam de compartilhar material genético devido à ocorrência do isolamento geográfico e passam a acumular diferenças genéticas por deriva genética e seleção natural. Com o aumento no tempo de divergência entre as espécies, cada vez mais as diferenças vão se acentuando o que torna mais difícil distinguir os genes que foram originalmente os responsáveis pelo isolamento reprodutivo daqueles que se diferenciaram após a divergência inicial. Assim, a identificação de genes candidatos que estão relacionados diretamente ao

processo de especiação é facilitada pelo estudo das diferenças genéticas de espécies irmãs ou proximamente relacionadas (ZENG et al., 2000).

1.2 – Expressão gênica em tecido específico

Toda célula viva, ao realizarem seus processos biológicos, coordenam os diferentes subconjuntos de seus genes a serem expressos em momentos diferentes. A expressão de genes específicos em determinado estágio e abundância são cruciais para o funcionamento adequado da célula e consequentemente do organismo. A mensuração dos níveis da expressão gênica em diferentes estágios do desenvolvimento, diferentes tecidos do corpo, diferentes condições ambientais e até mesmo de diferentes espécies, é fundamental para a compreensão de processos biológicos e essas informações podem ser úteis em estudos de caracterização de genes, efeitos de tratamentos experimentais e na compreensão de muitos outros processos moleculares (BEN-DOR et al., 1999). Muito embora grande parte da teoria evolutiva considere que mutações, deriva e seleção natural sejam os principais determinantes do processo evolutivo e especiação, a análise da expressão gênica tem revelado um papel importante para mudanças na quantidade de proteínas expressas (WHITEHEAD; CRAWFORD, 2006).

Perfis de expressão gênica têm a capacidade de evidenciar fenótipos ocultos a outras abordagens tradicionais gerando o potencial de fornecer novas perspectivas sobre especiação. Isso porque o entendimento sobre evolução está constantemente relacionado a habilidade de definir fenótipos, que devem ser precisos quanto a medição. Características morfológicas e comportamentais normalmente são de fácil e precisa medição, o que não acontece, por exemplo em estudos de fisiologia, o que reflete uma relativa pobreza de efeitos fisiológicos para explicar eventos de especiação ecológica (NEVINS; POTTI, 2007; PAVEY et al., 2010). Essencialmente, a expressão gênica nos permitir contornar esses limites, desmascarando fenótipos escondidos que possuam relevante potencial ecológico e fenótipos de difícil percepção (PAVEY et al., 2010). Neste panorama, novas tecnologias de produção de dados para análise de expressão gênica têm permitido o estudo de órgãos e tecidos específicos e revelado fenótipos ocultos que podem estar sendo expressos em condição específica. O poder dos perfis de expressão gênica em desvendar fenótipos está bem estabelecido em estudos de doenças genéticas humanas (LIOTTA; PETRICOIN, 2000; NEVINS; POTTI, 2007), mas precisa continuar a ser implementada em estudos de especiação ecológica.

Estudos de expressão gênica são baseados na criação de bibliotecas de cDNA, que são obtidas a partir do RNA mensageiro pelo processo de transcriptase reversa e são fundamentais quando se quer estudar apenas os genes que estão sendo expressos em um determinado organismo, tecido ou condição (CORTNER; WOUDE, 1997). Sabe-se que o sistema nervoso central é o responsável, em conjunto com estruturas receptoras sensoriais, em modificar o comportamento relacionado as diferentes funções biológicas, como por exemplo, a busca por alimento, parceiro sexual ou sítios de oviposição (ROBACKER; HEATH, 1997; NUNES et al., 2013).

A cabeça de uma mosca é composta por vários tecidos, incluindo olhos compostos, órgãos sensoriais responsáveis pelo olfato, paladar e audição, o corpo gorduroso (funcionalmente análogo ao fígado em cordados) e o cérebro (HUGHES et al., 2012). O olfato e o paladar desempenham um papel extremamente importante em praticamente todos os estágios da vida dos insetos, representando uma das principais interfaces entre o indivíduo e o meio ambiente. As mais importantes proteínas envolvidas no olfato incluem as *obp* (proteínas ligantes a odores), *csp* (proteínas quimiosensoriais), *or* (receptores de olfato) e *gr* (receptores gustativos) que são codificadas por famílias gênicas altamente divergentes e de evolução rápida (GALINDO; SMITH, 2001; MITCHELL et al., 2012; ZHU et al., 2012). O paladar está diretamente ligado à saliva que tem um papel importante na ingestão de alimentos e na interação entre um inseto e seu hospedeiro. Glândulas salivares labiais são o tipo mais comum de glândulas salivares em inseto (DELAY et al., 2012). Componentes do feromônio volátil podem ser produzidos e estocados nas glândulas salivares e em outras partes corporais (LIMA et al., 1996; LIMA et al., 2001; GONÇALVES et al., 2006).

A existência de genes em tecidos não-reprodutivos, como cabeça, que estejam sob seleção positiva são interessantes por promover a identificação de fenótipos que possam também estar afetando o comportamento sexual. Por outro lado, estes genes também podem estar influenciando funções sensoriais que diferem entre espécies pela existência de diferenças em suas características ecológicas (GREENSPAN; FERVEUR, 2000). Sinais de seleção positiva em tecido de cabeça entre espécies de separação recente de *Drosophila* foram relatados em genes do sistema olfatório (*Or9*) e ligantes a RNA (*Rlb*), além de diversos genes com função desconhecida (*CG9284* e *CG18869*) (JAGADEESHAN, 2005).

A expressão e regulação de genes específicos para um dos sexos na cabeça podem ser relacionadas a cascata de diferenciação dos sexos e a estados fisiológicos distintos, como observado nos genes *dsx* e *fru* de *Drosophila* (FUJII; AMREIN, 2002; CHANG et al., 2011). Estes genes são expressos no corpo gorduroso e provavelmente desempenhem funções específicas em comportamento sexuais ou estejam estritamente ligados ao ritmo circadiano de cada espécie (CLARIDGE-CHANG et al., 2001; FUJII; AMREIN, 2002). Estudos de expressão gênica de genes circadianos em *Drosophila* já existem há mais de 12 anos, normalmente usando-se microarranjos e mais recentemente RNA-Seq, para identificação de transcritos cíclicos em tecido cefálico das moscas nas mais diversas condições de luz e comparativa a linhagens mutantes (MCDONALD; ROSBASH, 2001; CLARIDGE-CHANG et al., 2001; HUGHES et al., 2012). Os principais genes conhecidos do ritmo circadiano como *per*, *tim*, *Clk*, *cry* e *vri* são constantemente relacionados a genes associados à visão, ativados pela presença ou ausência de luz (PLAUTZ, 1997).

Devido a estas importantes funções e à escassez de estudos para o gênero *Anastrepha*, é possível que parte dos genes envolvidos na determinação das diferenças interespecíficas dos grupos morfológicos esteja sendo expressa na cabeça.

1.3 – Sequenciamento de próxima geração (NGS)

A principal metodologia para se medir expressão gênica de um tecido até pouco tempo atrás se resumia às tecnologias de microarranjos (MARIONI et al., 2008; MATSUMURA et al., 2010), porém com o desenvolvimento das tecnologias de sequenciamento de próxima geração (do inglês Next-Generation Sequencing – NGS) a forma de se estudar expressão gênica mudou drasticamente (SHENDURE, 2008). As NGS têm facilitado a obtenção de dados de expressão gênica pela utilização da metodologia de sequenciamento de RNA (RNA-Seq), que utiliza as bibliotecas de cDNA para a criação de transcriptomas (MARIONI et al., 2008; WANG et al., 2009; RIESGO et al., 2012; VIJAY et al., 2013; XIAO et al., 2013).

As NGS geram gigabases de sequência de DNA em um curto espaço de tempo e a custos mínimos utilizando plataformas como Illumina, Roche 454 e ABI SOLiD. Isso significa que genomas inteiros podem agora ser sequenciados a partir do zero, dentro dos limites financeiros normais de uma pesquisa científica (GOMPERT et al., 2010). Uma infinidade de aplicações está sendo desenvolvida para essas plataformas, nas mais diversas áreas do

conhecimento (DAVEY; BLAXTER, 2010; GOMPERT et al., 2010; JACKSTADT et al., 2013; KOBOLDT et al., 2013), o que têm auxiliado na elucidação de questões em diversas áreas da biologia como genética molecular, genética de populações, filogeografia, filogenia e ecologia molecular (DAVEY; BLAXTER, 2010; EMERSON et al., 2010; HOLSINGER, 2010; BOKULICH et al., 2012; BRAS et al., 2012). Dentre estas questões, as NGS estão facilitando grandemente o processo de mapeamento genético, permitindo a geração rápida de mapas de ligação compostos por milhares de marcadores sequenciados, criando mapas de ligação de alta densidade, de tal forma que as sequências úteis ligada a um gene de interesse podem ser identificados em um único experimento (BAXTER et al., 2011). Este novo panorama também tem permitindo a utilização de organismos não-modelo para se estudar a base genética de diversas características importantes (RIESGO et al., 2012), que antes era restrita a organimos modelo. As novas tecnologias de próxima geração estão sendo aplicadas em organismos interessantes por suas propriedades ecológicas, fisiológicas, de desenvolvimento ou evolutiva e não pela complexidade da informação genética disponível neles (EMERSON et al., 2010; RIESGO et al., 2012), podendo, portanto ser aplicada a praticamente qualquer organismo de interesse.

A maior eficiência em tempo e custo das NGS advém do uso da clonagem *in vitro* e de sistemas de suporte sólido para as unidades de sequenciamento, que dispensa grande parte do intensivo trabalho laboratorial de produção de clones bacterianos, da montagem das placas de sequenciamento e da separação dos fragmentos em géis. A clonagem *in vitro* em suporte sólido permite que milhares de leituras possam ser produzidas de uma só vez com a plataforma 454, Solexa ou SOLiD (CARVALHO; SILVA, 2010; KU; ROUKOS, 2013). Porém, existe a preocupação de que essas novas tecnologias de sequenciamento tendem a ter taxas mais altas de erro do que o tradicional sequenciamento de Sanger, e pode levar à superestimação dos níveis de polimorfismo molecular (GOMPERT et al., 2010), que ainda é motivo de questionamentos (HARISMENDY et al., 2009).

1.4 – A família Tephritidae

Diferentemente das amplamente estudadas mosca de frutas da família Drosophilidae, as moscas-das-frutas verdadeiras, da família Tephritidae, ainda são pouco estudadas, apesar de sua importância econômica. Tefritídeos são as pragas mais importantes da agricultura em todo o

mundo por causarem danos à cultura de frutos carnosos, promovendo prejuízo econômico de bilhões de dólares todos os anos (JEYASANKAR, 2009). A maior parte dos prejuízos ocorre pela inviabilização dos frutos causado pelo consumo do endocarpo pelas larvas, e também pelo processo mecânico da oviposição, que permite a invasão de outros organismos como fungos causadores de podridão, provocando o amadurecimento precoce dos frutos (DUARTE; MALAVASI, 2000). Além de reduzir a produção do fruto hospedeiro, as culturas infestadas também provocam o aumento do uso de inseticidas causando a exposição dos trabalhadores e consumidores aos produtos agrotóxicos usados no combate da praga (JEYASANKAR, 2009).

Um foco importante dos estudos em tefritídeos ocorre no campo da evolução (FEDER et al., 1998; SOBRINHO; BRITO, 2010; SOBRINHO; BRITO, 2012) pela recente e rápida divergência de vários grupos da família (BERLOCHER, 2000; CLARKE et al., 2005; CONDON et al., 2008; VIRGILIO et al., 2008). Em muitas espécies deste grupo é comum a ocorrência de especiação simpátrica exatamente em virtude desta rápida divergência e adaptação a diferentes hospedeiros (BUSH, 1975; FEDER et al., 1988). Especificamente em algumas espécies do gênero *Anastrepha* pertencentes ao grupo fraterculus, a existência de polimorfismos intra-específicos e a geração de prole viável de alguns cruzamentos interespecíficos em laboratório (SANTOS et al., 2001; SELIVON et al., 2005; HENNING; MATIOLI, 2006), sugerem que essas espécies divergiram recentemente e que os mecanismos de isolamento reprodutivo provavelmente ainda não estão consolidados, por não terem acumulado diferenças genéticas suficientes (FERNANDES, 2010).

1.4.1 – O gênero *Anastrepha*

Existem 481 gêneros de tefritídeos descritos (ALUJA; NORRBOM, 2000), porém o que mais se destaca é o gênero *Anastrepha* (Schiner) por ser o mais importante economicamente e o maior em número de espécies, apenas nas Américas tropicais e subtropicais possuem mais de 237 espécies descritas (ALUJA, 1994; NORRBOM et al., 2000; NORRBOM; KORYTKOWSKI, 2009; NORRBOM; KORYTKOWSKI, 2011). *Anastrepha* foi dividido em 17 grupos baseado em caracteres morfológicos como coloração do corpo, padrão de desenhos nas asas e tamanho do acúleo das fêmeas (NORRBOM et al., 2000). No Brasil, existe registro de pelo menos 95 espécies de *Anastrepha* de 13 grupos (URAMOTO et al., 2004), destes, destacamos o "grupo *fraterculus*" composto por 29 espécies, algumas delas crípticas, devido ao

compartilhamento de semelhanças morfológicas e genéticas (NORRBOM et al., 2000; NORRBOM; KORYTKOWSKI, 2009). *Anastrepha obliqua* e *Anastrepha fraterculus* são espécies proximamente relacionadas deste grupo e pela sua ampla gama de hospedeiros, vasta distribuição e escassez de estudos são duas das principais espécies do grupo (ZUCCHI, 2000; SOLFERINI; MORGANTE, 1987).

Apesar de estarem proximamente relacionadas, as espécies do grupo *fraterculus* apresentam algumas diferenças morfológicas, dentre as quais, o tamanho do ápice do acúleo, as manchas presentes no subescutelo e o padrão de coloração e morfologia da asa (SELIVON, 2000; MALAVASI; ZUCCHI, 2000), e comportamentais, como a preferência por determinados frutos para a oviposição e o horário de cópula e oviposição na natureza (HEATH et al., 2000). Alguns autores afirmam que em insetos fitófagos, o polimorfismo associado ao hospedeiro pode representar o primeiro estágio de especiação (MALAVASI; MORGANTE, 1982; FEDER et al., 1998). Essas características analisadas independentemente não nos permitem identificar com precisão as espécies, pois há variação ao longo da distribuição geográfica e também entre os exemplares obtidos em um mesmo hospedeiro (ARAUJO; ZUCCHI, 2006). Portanto, apesar das espécies estarem proximamente relacionadas, há evidencias de que esses marcadores morfológicos e comportamentais não correspondam aos padrões de variação e divergência genética dessas moscas (MORGANTE et al., 1980; MALAVASI; MORGANTE, 1982). Estudos moleculares realizados até o momento também não conseguiram identificar marcadores específicos de cada espécies ou grupos monofiléticos que separem as espécies, utilizando marcadores mitocondriais (SMITH-CALDAS et al., 2001; ALUJA, 1994) ou nucleares (RUIZ et al., 2007; SARNO et al., 2010; SOBRINHO; BRITO, 2010; SOBRINHO; BRITO, 2012).

Aliando à escassez de dados genéticos que possibilitem a precisa identificação de genes ligados a diferenciação evolutiva entre *A. fraterculus* e *A. obliqua* e as novas e revolucionárias plataformas de sequenciamento de próxima geração, nós justificamos o trabalho baseando-se na oportunidade de analisar e entender, ao menos parcialmente, a evolução e diferenciação não só entre *A. fraterculus* e *A. obliqua*, mas também de todo o gênero *Anastrepha*. Este tipo de trabalho também é uma janela para facilitar o entendimento dos processos envolvidos na especiação o que, consequentemente, poderia permitir a utilização dos marcadores associados à especiação nestas espécies, ou de marcadores equivalentes, para espécies próximas ou testá-los em espécies com histórias evolutivas e comportamentais semelhantes. Finalmente, a alta quantidade de dados

genéticos gerados pelas novas tecnologias de sequenciamento abre um leque de possibilidades de pesquisa para diversos aspectos da biologia das espécies estudadas.

CAPÍTULO 2 – OBJETIVOS

2.1 – Objetivo geral

Comparação da expressão gênica diferencial entre tecidos de cabeça de *A. fraterculus* e *A. obliqua* para identificação de genes candidatos a estarem envolvidos no processo de diferenciação entre estas duas espécies.

2.2 – Objetivos específicos

a. Criação de transcriptomas de tecidos cefálicos em diferentes fases da vida reprodutiva de *A. fraterculus* e *A. obliqua* por sequenciamento de próxima geração, utilizando a estratégia de RNA-Seq.

b. Construção de um banco de dados de transcriptoma de cabeça de *A. fraterculus* e *A. obliqua* devidamente anotado.

c. Investigação da expressão gênica diferencial dentro dos contrastes criados nas fases da vida reprodutiva entre as espécies.

d. Busca por genes com padrão de expressão diferenciado entre as espécies, o que pode estar associado a diferentes preferências comportamentais ou ambientais como dieta, reprodução, ritmo circadiano e odores, dentre diversos outros aspectos que diferem entre as espécies, bem como com o processo de especiação em si.

e. Promover uma análise evolutiva, baseada na busca por SNPs com frequências altamente diferenciadas e testes de seleção positiva, que possa identificar genes candidatos a estar envolvidos no processo de diferenciação de *A. fraterculus* e *A. obliqua*.

CAPÍTULO 3 – MANUSCRITO 1

Characterization of Head Transcriptomes at Different Life Stages and Analysis of Gene Expression in Two Species of *Anastrepha* Fruit Flies (Tephritidae, Diptera)

Victor Borges Rezende[1]
Email: victorrez85@yahoo.com.br

Carlos Congrains Castillo[1]
Email: carlos_congrains@outlook.com

André Luís A. Lima[1]
Email: andreluisbio@gmail.com

Emeline Boni Campanini[1]
Email: emelinebc@gmail.com

Aline Minali Nakamura[1]
Email: alinemnk@gmail.com

Janaína Lima Oliveira[1]
Email: janajpb@hotmail.com

Samira Chahad-Ehlers[1]
Email: schahad@gmail.com

Iderval Sobrinho Junior[1]
Email: iderval_jr@yahoo.com

Reinaldo Alves de Brito[1*]
[*]Corresponding author
Email: brito@power.ufscar.br

[1]Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, Brazil

Introduction

Every living cell has to coordinate different subsets of genes being expressed at different times in order to perform biological processes. The expression and abundance of specific genes at a certain stage are crucial for the proper functioning of the cell and, consequently, the organism. Measurements of gene expression levels at different stages of development, different body tissues, different environmental conditions and even different species, are crucial for understanding molecular and biological processes, characterize genes, and perform experimental treatments (1). Although much of evolutionary theory considers that mutations, drift, and natural selection are the main determinants of the evolutionary and speciation process, the analysis of gene expression has revealed an important role for changes in the amount of protein expressed as well (2).

Since expression patterns may change rapidly across species, such changes may exert great morphological and behavior differentiation even when there is little genetic difference between species, as it seems to be the case of species in the *fraterculus* group of the genus *Anastrepha* (Diptera: Tephritidea). This is the largest genus in the family Tephritidae, with over 237 species, but some of the most economically important species in the genus belong to the closely related and recently diverged *fraterculus* group. These fruit flies encompass some of the most economically devastating fruit pests in the world because of the damage they inflict to the culture of diverse fleshy fruits. For several species in the group, the mechanisms of reproductive isolation are not yet consolidated (3–5), which facilitates the study of species differences and the identification of important markers that may help establishing pest control strategies. This is the case of the closely related species *Anastrepha obliqua* and *A. fraterculus*, both of which are somewhat generalists, and are widely distributed, making them some of the main species of the genus.

Several of the genes involved with the differentiation in the *fraterculus* group may be expressed in the head, where we find several sensory receptor structures, as well as the central nervous system. Together, these structures are responsible for several behaviors and biological functions that have been associated with species differences, such as host preference, choice of mating and oviposition sites (6,7). The events of copulation and oviposition are responsible for changes in behavior and physiology that affect their reproductive successes (8). Several, if not

the majority of reproductive changes along their life are directly related to the expression of different genes, which are also responsible for activating complex cascades of differentiation in the head and the rest of the body (9).

The head of a fly is composed of several tissues, including compound eyes, sense organs responsible for smell, taste and hearing, the fat body (functionally analogous to the liver in chordates) and the brain (10). The smell and taste play an extremely important role in virtually all life stages of insects, representing one of the main interfaces between the individuals and their environment. The most important proteins involved in olfaction include *Obp* (odorant binding proteins), *csp* (chemosensory proteins), *or* (olfactory receptors) and *gr* (gustatory receptors) that are encoded by highly divergent gene families and undergoes rapid evolution (11–13). The palate is directly connected to the saliva and plays an important role in food intake and in the interaction between an insect and its host. Labial salivary glands are the most common type of insect salivary glands (14).

NGS (Next Generation Sequencing) has facilitated the production of gene expression data by the method of RNA sequencing (RNA-Seq) using cDNA libraries for creating transcriptomes (15–19). Here we sought to identify genes involved in the differentiation of *A. fraterculus* and *A. obliqua*, two of the most economically important species of the genus, by focusing on expressed genes in different reproductive stages in the head tissue, which includes several important tissues that are involved in interactions with environmental activities, such as finding food, mates and sites for breeding and oviposition.

Methods

Samples

Flies were obtained in the field from guava and Jocote (Anacardiaceae) fruits, collected respectively in midwest (16^{o} 41' 58" S, 49 ^{o} 16' 35" W) and southeastern (22^{o} 01' 03" S, 47 ^{o} 53' 27" W) regions of Brazil. Populations of the two species used in this work have been established in the Population Genetics and Evolution lab at the Federal University of São Carlos, Brazil, for over a year now, in a controlled environment room at 26 ^{o}C ± 5^{o}C (60–90% humidity), and natural photoperiod. The newly emerged adults were morphologically identified as *Anastrepha obliqua* and *A. fraterculus*, respectively, and their decedents were kept in acrylic cages supplied with water and a mixture of hydrolyzed protein, vitamins, and dry sucrose as food. Flies were

allowed to mate after they became sexually mature and their proliferation was performed using mango fruit introduced into the cages for oviposition. Each fruit was then transferred to a new cage filled with vermiculite into which mature larvae migrated and pupated. Finally, pupae were sieved from the vermiculite and transferred to a new cage, establishing a new generation. In order to reduce inbreeding, these populations were maintained at a size of at least 100 mating adults at a time.

Different transcriptome profiles were established considering different genders and reproductive stages (virgin, post-mating, and post-oviposition) for each species (*A. obliqua* and *A. fraterculus*), totaling 10 different profiles, with replication. Fly heads from 10 individuals per profile were used to form pools of individuals that would make each library, adding to two hundred individuals in total. Virgin flies had their heads collected 10 days after emergence to ensure their sexual maturity. Post mating males and females were collected between 15 and 20 hours after a successful copulation, while heads from post oviposition females were collected immediately after oviposition.

RNA extraction, cDNA library construction and sequencing

Total RNA was extracted from pools of five heads using the Trizol/chloroform protocol (20). The RNA quality was visually inspected in agarose gel and quantified in a Qμbit fluorometer and a Nanodrop spectrophotometer. RNA-seq libraries were constructed from 4 μg of total RNA using the TruSeq Stranded Total RNA Sample Prep Kit (Illumina) protocol according to the manufacturer's instructions. Libraries were sequenced on an Illumina HiSeq2000 (Laboratory of Functional Genomics Applied to Agriculture and Agri-energy, ESALQ-USP, Brazil) on flow cells with runs of 2 x 100 bp paired-end reads. We used Illumina's HiSeq Control Software and CASAVA v1.8.2 software (Illumina, Inc.) for base calling and sample demultiplexing.

Sequence cleaning and assembly

All reads were trimmed for quality and length using the software SeqyClean, available at https://bitbucket.org/izhbannikov/seqyclean. We only kept reads that had a minimum sequence length of 50, a minimum of 0.01 for the parameter 'max-avg-error' and 0.05 for 'max-error-at-ends', and an average Phred quality score ≥ 20. SeqyClean also scans and removes from the sequences any remaining adapter sequences. This software analyses the two complementary sequences (paired-end) together and discards both sequences if only one does not fulfill the

filtering requirements established. This feature is important for the software that assembles the contigs, because all reads are paired, producing a more efficient assembly, especially in the absence of a reference genome.

Processed reads were assembled using the Trinity short read assembler (release 2013-02-25) (21), using default parameters (22) on a Dell T610 server (24 cores, 128 G of memory) at the Population Genetics and Evolution Laboratory at the Federal University of São Carlos. We processed all 10 libraries from both species into a single pooled assembly, which was used as a reference transcriptome for mapping and gene expression data construction.

Functional annotation

In order to annotate the assembled transcripts, we first investigated for the presence of open reading frames (ORFs) using Trinotate (http://trinotate.sourceforge.net/), a Trinity software for functional annotation of ORFs that uses the search for homologs (NCBI-BLAST). After having identified the best ORF candidates, we searched these ORFs using the BLASTp tool in the package 'Standalone BLAST Setup for Unix' (23) against the gene ontology terms (GO), an annotated database of ESTs of *Drosophila melanogaster* (BDGP5/dm3) and an annotated database of ESTs of *Ceratitis capitata*. We also mapped the gene ontology terms of the transcripts in categories with the GOTermMapper tool (24).

Mapping reads against reference transcriptome and analysis of gene expression

We followed a Trinity protocol (21,25) that includes analyses using Bowtie aligner (26), RSEM (27) and EdgeR (16,28), to align the reads back to the assembled reference transcriptome and perform the differential gene expression analysis. The Bowtie software (version 1.0.0) was used to align the RNA-seq Illumina reads back to the reference transcripts assembled by Trinity, which was done for each of the library and its replicate. Because of a lack of an available reference genome, we used the RSEM software to associate reads to genes and isoforms and to estimate the abundance of transcripts in the assemblies with the information provided by the mapping. EdgeR (Bioconductor's package of R) was used to identify differentially expressed transcripts based and clustering transcripts according to expression profiles and also to correct for potential biases caused by differences in RNA output across different samples. This was achieved since EdgeR adopts a TMM (trimmed mean of M values) method (29) to calculate a normalization factor, using *calcNormFactors* function. Contigs were considered to be significantly differentially expressed at the $P<0.05$ level if the corrected FDR (false discovery

rate) was below <0.001 (30). EdgeR is also responsible for ploting the graphics that enable the visualization of such results and to cluster differentially expressed genes by similar patterns of expression. Bowtie, RSEM and EdgeR were used with the default parameters as implemented in the Trinity software package, available for consultation on the website http://trinityrnaseq.sourceforge.net/analysis/diff_expression_analysis.html (21).

Clusters of genes with different patterns of expression were analyzed for classes of enriched GO terms in GOrilla, which investigates evaluates if the list of GO terms created from a set of genes differentially expressed differs significantly from a background list (31). We compared each cluster of genes against a background lists of GO terms created from all genes expressed in the combined analysis of both species, *A. fraterculus* and *A. obliqua*.

Results and Discussion

Illumina Paired-end Sequencing and *de novo* Assembly

The cDNA libraries generated 155,940,826 paired-end reads of a hundred base pair (bp) after adaptor removal from 20 libraries (5 profiles per species with replicates). This represents an average of almost 7,8 million raw reads per library profile. The library with most reads was one of the replicates of the post mating males of *A. fraterculus* with 11,3 million reads whereas the library with less reads was one of the replicates of virgin males of *A. obliqua* with 6,19 million reads. The distribution of the raw data generated by Illumina broken down by each library profile by species is shown in supplementary material (Table S1). The raw reads were then filtered by quality, resulting in 140,493,653 paired-end reads and over 28 Giga base pairs distributed in more than 7 million reads per replicate in *A. fraterculus* and 6,68 millions for *A. obliqua* with an average of 95 base pairs per read for both species. SeqyClean software removed from the analysis between 6.41% and 15.9% of the paired end reads due to low quality. The number of filtered reads, percentage of discarded reads and number of bases kept in the analysis are shown in Table S2 of the supplementary material. We processed this data in Trinity to create a pooled assembly of different genders, states and species which had a total of 154,787 contigs. The assembly length distribution of these data includes an N50 of 2012 and 213 contigs with more than 10,000 bp. The average for all contigs was 1,027 bp with a length of assembled contigs that ranged between 201 and 25,704 bp. The results of sequencing by Illumina and assembly by Trinity are shown in Table 1.

Table 1. Summary of the sequencing by Illumina and assembly by Trinity

Total of Paired-end reads	155,940,826
Filtered reads	140,493,653
Total number of contigs	154,787
N50	2,012
Contigs longer than 1000 bp	45,602
Contigs longer than 2000 bp	22,297
Contigs longer than 10000 bp	213
Average (bp)	1,027
Median length (bp)	494
Longest contig (bp)	25,704

Functional Annotation

All transcripts built by Trinity were blasted against gene ontology database (GO) for validation and annotation using BLASTx algorithm with an E-value threshold of 10^{-5}. By this approach, out of 154,787 transcripts, 55,990 genes (36,17% of all distinct sequences) returned an above cut-off BLAST result.

The annotated transcripts were also distributed in 40 different categories of gene ontology by GOTermMapper tool, belonging to 3 great classes. These 3 great ontologies cover the function of certain level of living units (biological process) , the activity of a gene product at the molecular level (molecular function), and the domains that belong to the parts of a cell or its extracellular environment (cellular component). The main category of biological process was “anatomical structure development”, associated with 33,45% of annotated transcripts, whereas the most sampled category of molecular function was “ion binding”, present in 24,76%, and 59,48% of transcripts in the cellular component class were ascribed to “cell” (Figure 1).

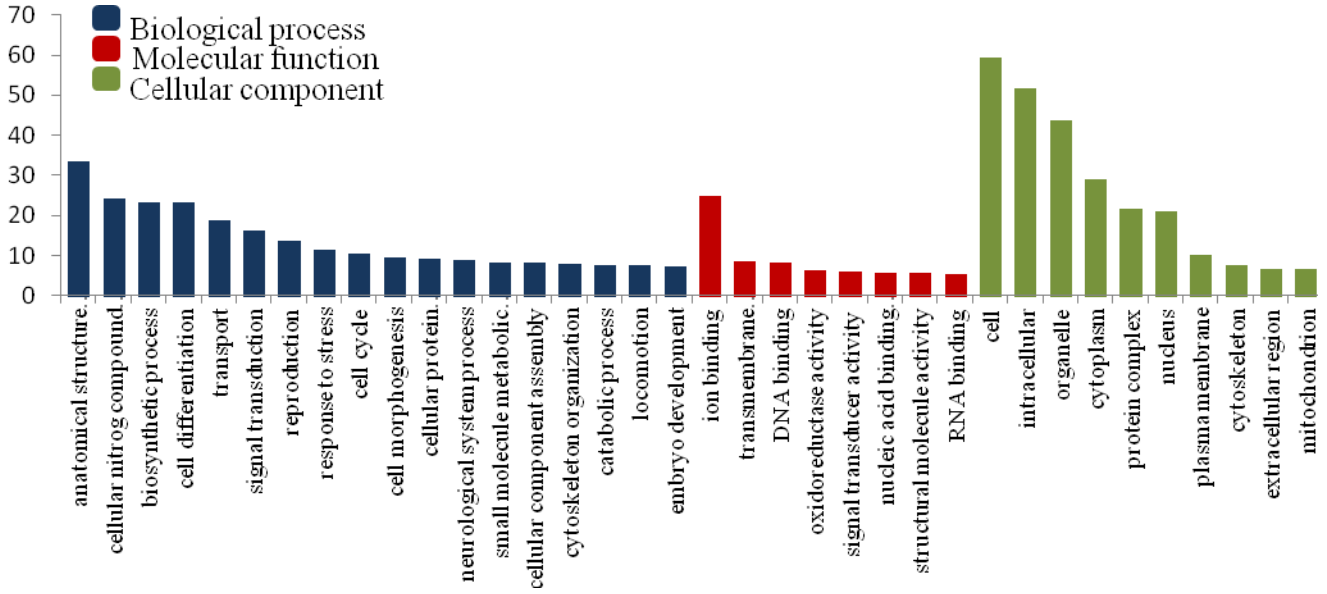


Figure 1. Histogram presentation of Gene Ontology classification. The results are summarized in three main classes: biological process, molecular function and cellular component. The y-axis indicates the percentage of the total of 55,990 of a specific category of genes in the main class.

Differential Gene Expression

We performed analyses of gene expression contrasting different life stages and species enabling a broad contrast of differences in patterns of gene expression. In order to compare different life stages within a species, we considered the two replicates for each library at each different life stage level as independents and this makes sense, since they were produced from different pools of individuals.

The comparisons of differentially expressed genes across life stages and species are shown in Figure 2 and indicate that a total of 1991 genes were differentially expressed in at least one profile between the species. The largest number of differentially expressed genes between *Anastrepha fraterculus* and *A. obliqua* was observed in virgin males, which showed 1560 differentially expressed genes between the species, a great contrast with the results found for post mating females, which showed the lowest number of differentially expressed genes. Between these two extremes there are virgin females (909); post mating males (612) and post oviposition females (84). Although there is a big difference in the total number of genes differentially expressed between the life stages, the proportion of 4 in 5 comparisons are close to 50% being more expressed in each of the species here studied, showing a consistent pattern of differentiation.

As the results clearly indicate, there are many more genes differentially expressed in males (2,172) than in females (1,018). This pattern is even more striking when you consider that the majority of these differentially expressed genes are expressed in virgin males. This interesting result may be a consequence of positive selection differentially affecting males, which has been indicated elsewhere for genes involved with reproduction (32). Combined, olfactory genes were three times more observed with higher expression in males (28 records) than in females (9 records). The *Obp* gene family, for example, had 5 annotated genes in females and 11 in males. Even more interestingly, the number of olfactory genes with differential expression is even more skewed towards males. The olfactory behavior genes, such as *Sip1* (*SRY interacting protein 1*), *Rtnl1* (*Reticulon-like1*), *Pino* (*Pinocchio*) and *scrib* (*scribbled*), showed differential expression for eight contrasts, but only when comparing different male profiles. The olfactory learning genes, such as *pst* (*pastrel*), *mol* (*moladietz*) and *vsg* (*visgun*), were recorded 9 times in males, 7 only by *vsg*, which also appeared 4 times in virgin female at all.

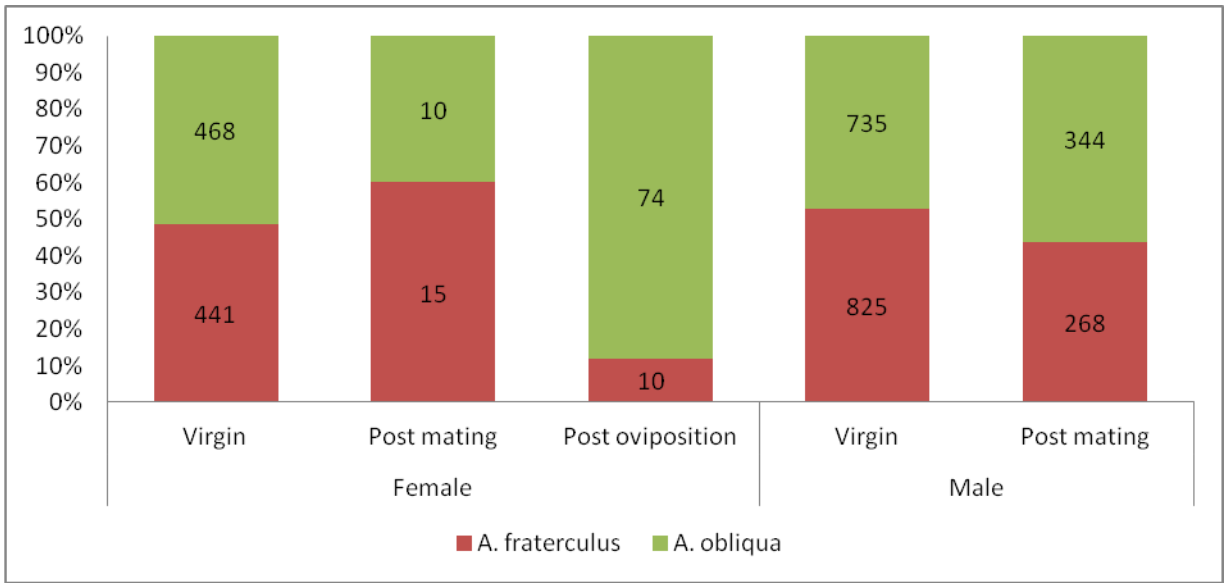


Figure 2. Proportions and numbers of differentially expressed genes in profiles between *A. fraterculus* (red bar) and *A. obliqua* (green bar).

EdgeR also plotted together all transcripts and analyzed the differential gene expression of all twenty libraries combining transcripts with similar expression patterns. Two of these graphics show an interesting pattern that seems to discriminate the two species for all their libraries (Figure 3). We identified 123 genes with a similar pattern of over-expression in *A.*

obliqua and under-expression in *A. fraterculus*, whereas there are 237 genes that are over-expressed in *A. fraterculus* and under-expressed in *A. obliqua*.

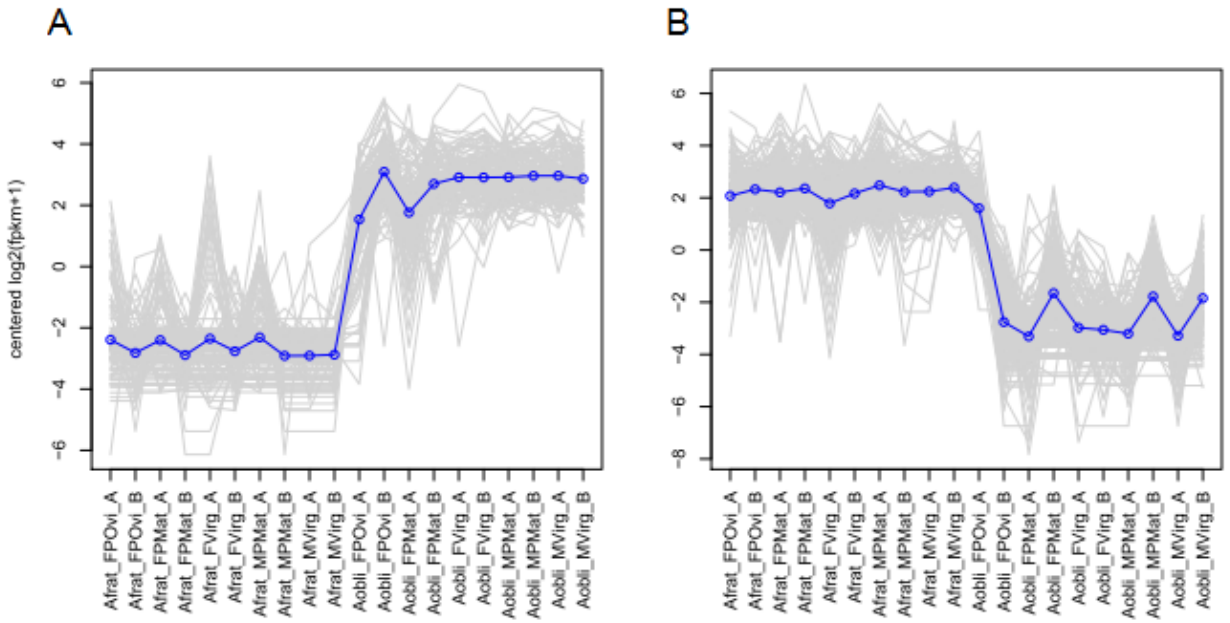


Figure 3. EdgeR plots of clusters of differentially expressed genes based in expression pattern using all profiles between *A. fraterculus* and *A. obliqua*. Each grey line represents the expression of one gene and the blue line represents the mean-centered expression patterns for each cluster. In X-axis we have the profiles replicates (Afrat = *A. fraterculus*; Aobli = *A. obliqua*; F = female; M = male; POvi = post oviposition; PMat = post mating; Virg = virgin; A and B are replicates). (A) Cluster of genes overexpressed in *A. obliqua* when compared to *A. fraterculus*. (B) Cluster of genes overexpressed in *A. fraterculus* when compared to *A. obliqua.*

The annotation of clusters of genes over- or under- expressed in one species when compared to the other revealed some important genes that may be involved with species differentiation. We are interested in genes, or classes of genes, that appear in only one of the clusters because they may provide hints regarding genes which might be involved with differences in species preferences, behaviors and activities.

Odorant-binding proteins, *Obps*, are present in both clusters of differentially expressed genes; however, different members of this gene family are found being differentially expressed. There are four *Obps in the A. fraterculus* overexpressed gene cluster, two *Obp56a*-like and two *Obp99b*-like, while in *A. obliqua* we find two *Obp56e*-like. On the other hand, two annotated genes related to pheromone-binding protein, "*Pbprp2*" and "*Pbprp3*", are exclusive to the *A. obliqua* gene cluster. Since odorant and pheromone binding proteins are involved in chemical recognition in the environment, the survival and reproduction success of an individual is directly

associated with a correct identification and a quick achievement of an especific physiological target (11,33). Thus, molecules of these different gene families found in different species could suggest changes in food habits, host preference or mating preference (34,35).

We have also identified several proteases and protease inhibitors, such as the serine protease inhibitor, *Spn43Ab,* and a peptidase related to the cuticle development gene *svr* which are exclusively observed in the *A. fraterculus* overexpressed gene cluster. Serine protease inhibitors have been associated with several different processes in which the prevention of proteolysis in the body is required, ranging from digestive and development to immune responses, and most of them may have important impacts on species differences (36). Serine protease inhibitors (SPIs) are in general a large family of genes in insects and most of its members are specific to one or few targets, making it difficult to infer the role for a specific SPI. The combination of proteases and protease inhibitors have been shown to have an important impact in species differences for different taxa (37)

To evaluate if there were some clusters of genes or a cascade of genes that were consistently over- or under-expressed in certain species we built a generic GO term list for the genes inferred from each contig for each library and compared that to the full annotation of genes inferred from the *Anastrepha* transcriptome, considering three great classes of ontology terms, biological process (Figure 4), molecular function (Figure 5) and cellular component (Figure 6). The investigation for terms that were enriched in over- or under- expressed genes when compared to the overall transcriptome, performed in GOrilla, produced very interesting results, suggesting some classes of genes that may be preferentially involved with species differences, which are discussed below. We should point out that GOrilla associated GO terms only to 75 out of 123 over-expressed genes in *A. obliqua* and 167 out of 237 that are over-expressed in *A. fraterculus*. This occurs either because the gene is repeated, and therefore removed from the list, or because there is no GO term associated with the inferred gene. Repeated genes in the list of inferred genes may be derived from alternative splices but also to recent duplicated gene copies, which is not uncommon in large gene families. We should also point out that even though we obtained several significant results for GO terms, no terms have achieved significance after using FDr correction for multiple testes, possibly because of the small number of genes involved.

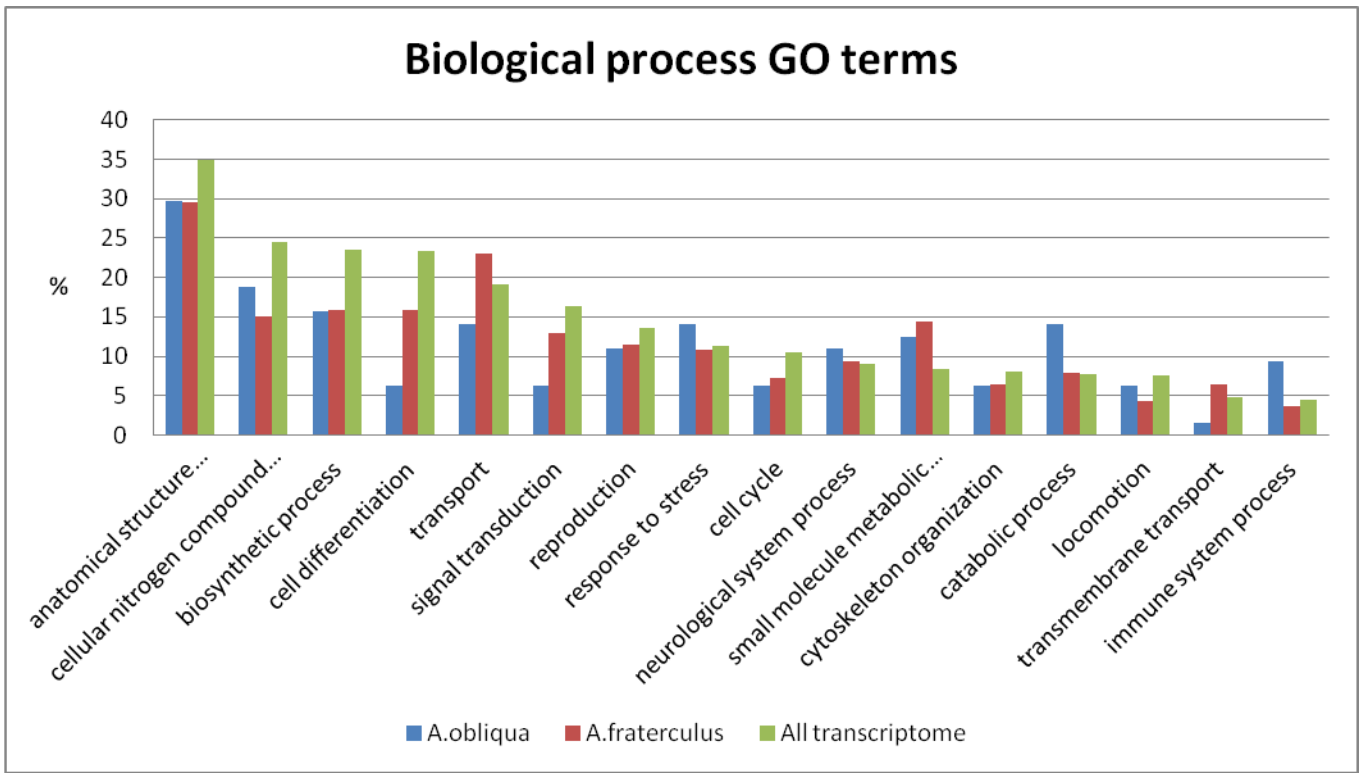


Figure 4. Distribution on percentage of biological process class of the GO slim terms. Comparison of the 123 genes with the pattern of high expression in *A. obliqua* (Blue pillar) against the 237 genes with the pattern of high expression in *A. fraterculus* (red pillar), against the full annotation of transcriptome result (green pillar) and against genome frequency of use made by GO consortium's based on FlyBase Generic GO slim.

Most of the differences in classes of GO terms was observed in the analysis of "biological processes". We have identified 10 GO terms that were significantly overrepresented among over-expressed genes in *A. fraterculus* (P-value < 10^{-3}), but only two that were overrepresented in *A. obliqua*, which may be due to the overall number of genes, since there are more than twice as many genes in the former than in the latter. To that effect, the most significant over representation of GO terms in *A. fraterculus* was "multicellular organism reproduction", which was one of the two significant term found in *A. obliqua*, with an enrichment of 7.35 and 9.2 fold, respectively. A subset of this category was also significant in *A. fraterculus* and contains basically the same sets of genes in this higher level category, which encompass several very interesting genes that were discussed before, such as "*vsg*", "*Spn43Ab*", and *OBP56e*. The presence of genes related to reproduction when investigating genes that are differentially expressed in different species comes as no surprise since the rapid evolution of reproductive genes has been extensively described in the literature. The class with the greatest enrichment, though, "immune response-regulating cell surface receptor signaling pathway

involved in phagocytosis", has immune genes that may be related to response to parasites, which has recently been associated to speciation events (for instance, see (38)). One category of Biological process that was only found enriched in *A. obliqua* was Locomotory Behavior which is associated with several genes, among which *cryptochrome*, an important gene involved with the circadian rhythm that has been linked to speciation in closely related species of *Bactrocera*, another Tephritidae (39). Behavior, such as feeding or reproductive behavior, has been consistently considered as one of the most important aspect of species differences (40), and some of the behavioral differences have been associated with the circadian cycles and their genes (41).

Three of the other significantly enriched GO terms in biological process involve over-expression of Glutatione S-transferase, which are important enzymes for several processes, chief among them the cellular detoxification of xenobiotic and endobiotic substances (42). The other three significantly enhanced GO terms involve Transferrins and Ferritins, which, like some Glutatione Transferases, are important antioxidants that protect against dietary and endogenous oxidants. These enzymes, as well as a few other, are probably responsible for the significantly enriched terms in different molecular functions (Table 2). Toxins and oxidants are some of the most common chemical defenses of host plants and fruits against insects (43) hence, the adaptation to different hosts, which is a common theme in tephritids and the case for the two species here studied, may be directly related to effective mechanisms for detoxification against plant secondary products.

The several next-generation libraries of cephalic tissues from two species of tephritids, *A. fraterculus* and *A. obliqua* here described provide a first glance of the myriad of genes expressed in this model organism about which very little is known. The initial analyses here performed we hope will be followed by a plethora of new studies that should help better the understanding on these important economic pests. These studies have indicated several genes that are potentially involved with species maturation and differentiation which should be further investigated to investigate their role not only in the differentiation of this species pair, but of other *Anastrepha* as well.

Table 2. Result of Enriched GO terms made by Gorilla

Term Description	*A. fraterculus*			*A. obliqua*		
	P-value	FDR q-value	Enrichment	P-value	FDR q-value	Enrichment
Biological Process						
multicellular organism reproduction	4.0E-05	2.2E-01	7.35	8.9E-04	1.0E+00	9.2
Reproduction	7.3E-05	2.0E-01	6.7		-	
cation transport	1.0E-04	1.8E-01	3.66		-	
glutathione metabolic process	1.1E-04	1.5E-01	10.11		-	
chemical homeostasis	3.0E-04	3.2E-01	5.38		-	
immune response-regulating cell*	3.3E-04	3.0E-01	54.59		-	
peptide metabolic process	4.6E-04	3.5E-01	7.58		-	
metal ion transport	5.3E-04	3.6E-01	4.28		-	
cation homeostasis	7.5E-04	4.5E-01	6.82		-	
sulfur compound metabolic process	9.0E-04	4.8E-01	5.28		-	
locomotory behavior		-		6.9E-04	1.0E+00	5.56
Molecular Function						
transferase activity**	9.0E-05	2.2E-01	7.99		-	
oxidoreductase activity	9.4E-05	1.1E-01	2.55		-	
glutathione transferase activity	5.1E-04	4.0E-01	10.4		-	
inorg. cation transm. transp. act.***	9.4E-04	5.6E-01	3.05		-	
Cellular Component						
oxidoreductase complex	6.9E-04	6.2E-01	5.55			
Envelope		-		2.0E-04	1.8E-01	9.2
organelle envelope		-		2.0E-04	8.8E-02	9.2

P-value' is the enrichment p-value computed according to the mHG or HG model; '**FDR q-value**' is the correction of the above p-value for multiple testing using the Benjamini and Hochberg (1995) method. Namely, for the i^{th} term (ranked according to p-value) the FDR q-value is (p-value * number of GO terms) / i. '**Enrichment**' is defined as follows by the formula (Enrichment = (b/n) / (B/N)), being N - the total number of genes, B - the total number of genes associated with a specific GO term, n - the number of genes in the top of the user's input list or in the target set when appropriate and b - the number of genes in the intersection. * abreviation of "immune response-regulating cell surface receptor signaling pathway involved in phagocytosis"; ** abreviation of "transferase activity, transferring alkyl or aryl (other than methyl) groups"; *** abreviation of "inorganic cation transmembrane transporter activity".

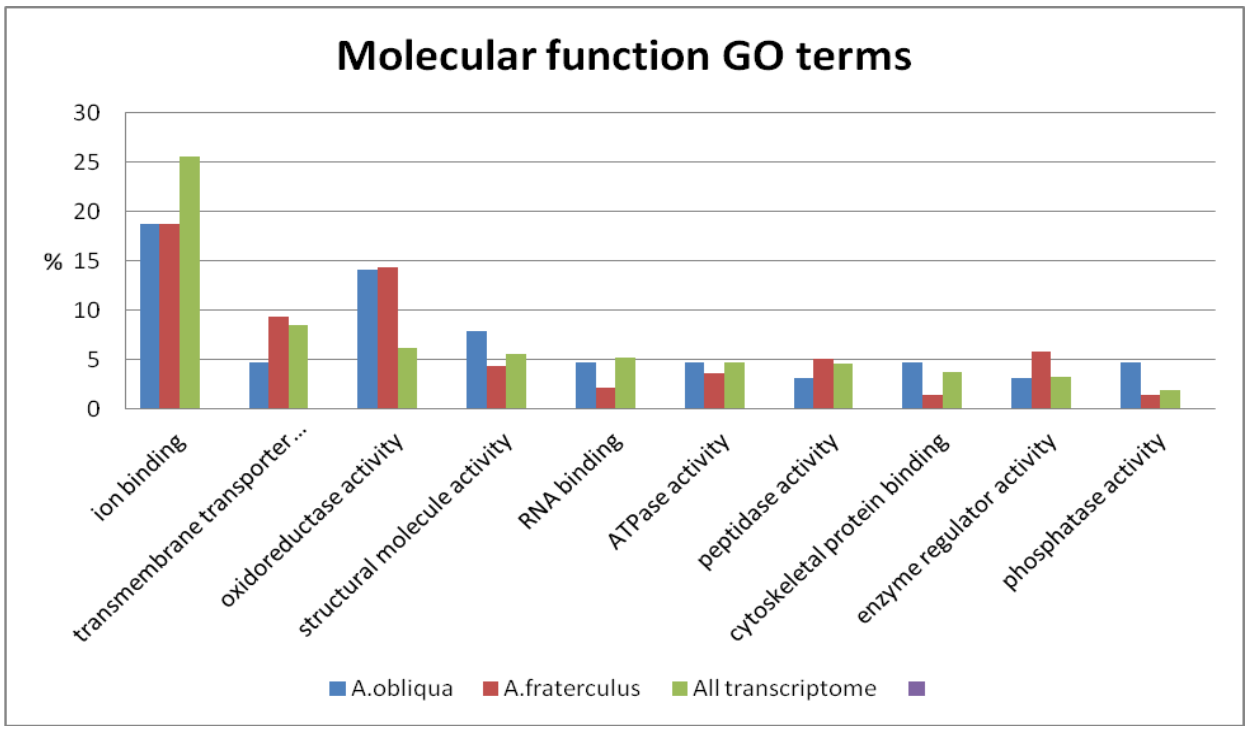


Figure 5. Distribution on percentage of molecular function class of the GO slim terms. Comparison of the 123 genes with the pattern of high expression in *A. obliqua* (Blue pillar) against the 237 genes with the pattern of high expression in *A. fraterculus* (red pillar), against the full annotation of transcriptome result (green pillar) and against genome frequency of use made by GO consortium's based on FlyBase Generic GO slim.

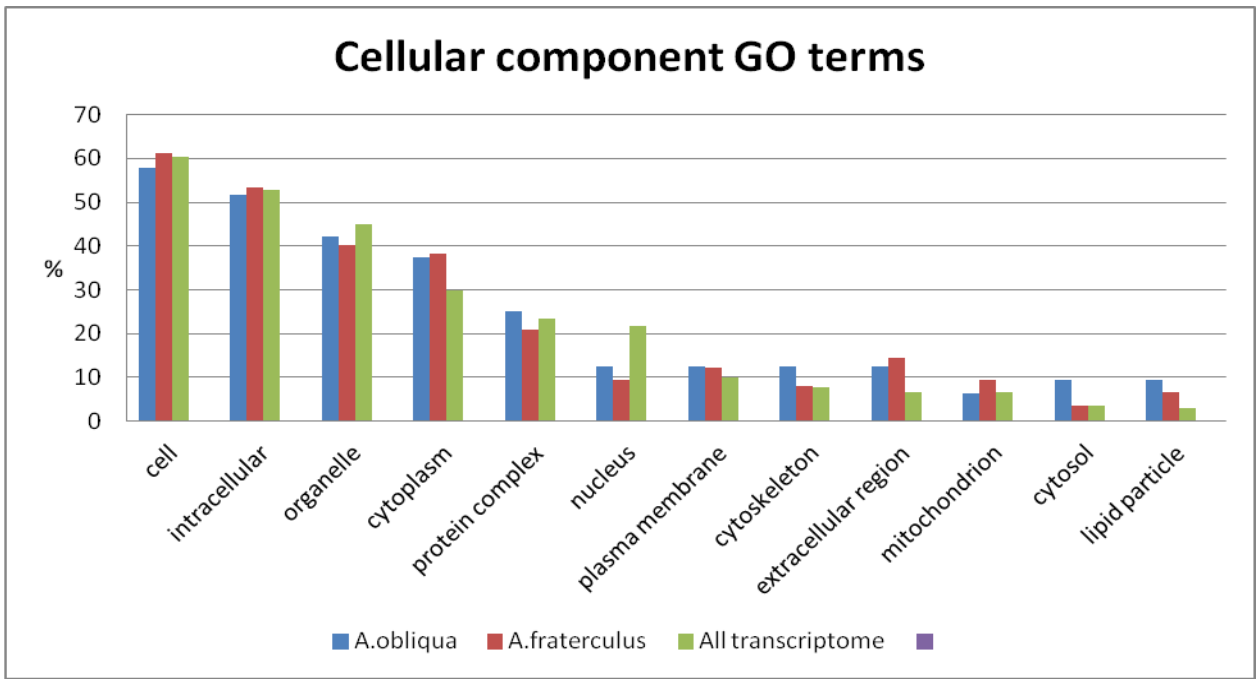


Figure 6. Distribution on percentage of cellular component class of the GO slim terms. Comparison of the 123 genes with the pattern of high expression in *A. obliqua* (Blue pillar) against the 237 genes with the pattern of high expression in *A. fraterculus* (red pillar), against the full annotation of transcriptome result (green pillar) and against genome frequency of use made by GO consortium's based on FlyBase Generic GO slim. The asterisks represent significant for 3 chi-square analysis with error probability of 5% values.

Authors' contributions

Rezende and de Brito conceived and designed the research. Rezende and de Brito wrote the manuscript. Rezende, Congrains, Lima, Campanini, Nakamura, Oliveira, Chahad-Ehlers, Sobrinho and de Brito worked in the establishment and maintenance of the flies populations in the laboratory and were also involved in tissue collect and RNA extraction. Rezende performed the bioinformatics analysis of assembly *de novo*, mapping, differential gene expression, annotation and statistical. Congrains, Sobrinho and Chahad-Ehlers contributed to the development of the project and writing of the manuscript. All authors read and approved the final manuscript.

References

1. Ben-Dor A, Shamir R, Yakhini Z (1999) Clustering Gene Expression Patterns. J Comput Biol 6: 281–297. doi:10.1089/106652799318274.

2. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution: Review of gene expression variation. Mol Ecol 15: 1197–1211. doi:10.1111/j.1365-294X.2006.02868.x.

3. Dos Santos P, Uramoto K, Matioli SR (2001) Experimental Hybridization Among *Anastrepha*; Species (Diptera: Tephritidae): Production and Morphological Characterization of F1 Hybrids. Ann Entomol Soc Am 94: 717–725. doi:10.1603/0013-8746(2001)094(0717:EHAASD)2.0.CO;2.

4. Selivon D, Perondini ALP, Morgante JS (2005) A Genetic–Morphological Characterization of Two Cryptic Species of the *Anastrepha fraterculus* Complex (Diptera: Tephritidae). Ann Entomol Soc Am 98: 367–381. doi:10.1603/0013-8746(2005)098(0367:AGCOTC)2.0.CO;2.

5. Henning F, Matioli SR (2006) Mating time of the West Indian fruit fly *Anastrepha obliqua* (Macquart)(Diptera: Tephritidae) under laboratory conditions. Neotrop Entomol 35: 145–148.

6. Robacker DC, Heath RR (1997) Decreased attraction of *Anastrepha ludens* to combinations of two types of synthetic lures in a citrus orchard. J Chem Ecol 23: 1253–1262.

7. Nunes MZ, Santos RS, Boff MIC, Rosa JM (2013) Avaliacão de atrativos alimentares na captura de *Anastrepha fraterculus* (Wiedemann, 1830)(Diptera: Tephritidae) em pomar de macieira. Rev Fac Agron Plata 112: 91–96.

8. Lawniczak MK, Begun DJ (2004) A genome-wide analysis of courting and mating responses in *Drosophila melanogaster* females. Genome 47: 900–910. doi:10.1139/g04-050.

9. Goldman TD, Arbeitman MN (2007) Genomic and functional studies of Drosophila sex hierarchy regulated gene expression in adult head and nervous system tissues. PLoS Genet 3: e216.

10. Hughes ME, Grant GR, Paquin C, Qian J, Nitabach MN (2012) Deep sequencing the circadian and diurnal transcriptome of *Drosophila* brain. Genome Res 22: 1266–1281. doi:10.1101/gr.128876.111.

11. Galindo K, Smith DP (2001) A large family of divergent *Drosophila* odorant-binding proteins expressed in gustatory and olfactory sensilla. Genetics 159: 1059–1072.

12. Mitchell RF, Hughes DT, Luetje CW, Millar JG, Soriano-Agatón F, et al. (2012) Sequencing and characterizing odorant receptors of the cerambycid beetle *Megacyllene caryae*. Insect Biochem Mol Biol 42: 499–505. doi:10.1016/j.ibmb.2012.03.007.

13. Zhu J-Y, Zhao N, Yang B (2012) Global Transcriptional Analysis of Olfactory Genes in the Head of Pine Shoot Beetle, *Tomicus yunnanensis*. Comp Funct Genomics 2012: 1–10. doi:10.1155/2012/491748.

14. DeLay B, Mamidala P, Wijeratne A, Wijeratne S, Mittapalli O, et al. (2012) Transcriptome analysis of the salivary glands of potato leafhopper, *Empoasca fabae*. J Insect Physiol 58: 1626–1634. doi:10.1016/j.jinsphys.2012.10.002.

15. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509–1517. doi:10.1101/gr.079558.108.

16. Wang L, Feng Z, Wang X, Wang X, Zhang X (2009) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26: 136–138. doi:10.1093/bioinformatics/btp612.

17. Riesgo A, Andrade SCS, Sharma PP, Novo M, Pérez-Porro AR, et al. (2012) Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Front Zool 9: 33. doi:10.1186/1742-9994-9-33.

18. Vijay N, Poelstra JW, Künstner A, Wolf JBW (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Ecol 22: 620–634. doi:10.1111/mec.12014.

19. Xiao M, Zhang Y, Chen X, Lee E-J, Barber CJS, et al. (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J Biotechnol 166: 122–134. doi:10.1016/j.jbiotec.2013.04.004.

20. Chomczynski P, Mackey K (1995) Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide- and proteoglycan-rich sources. BioTechniques 19: 942–945.

21. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652. doi:10.1038/nbt.1883.

22. Borodina T, Adjaye J, Sultan M (2011) A Strand-Specific Library Preparation Protocol for RNA Sequencing. Methods in Enzymology. Elsevier, Vol. 500. pp. 79–98.

23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

24. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinforma Oxf Engl 20: 3710–3715. doi:10.1093/bioinformatics/bth456.

25. Henschel R, Nista PM, Lieber M, Haas BJ, Wu L-S, et al. (2012) Trinity RNA-Seq assembler performance optimization ACM Press. p. 1. Available: http://dl.acm.org/citation.cfm?doid=2335755.2335842. Accessed 15 August 2013.

26. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

27. Li B, Dewey C (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

28. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi:10.1093/bioinformatics/btp616.

29. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.

30. Robinson MD, Smyth GK (2007) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9: 321–332. doi:10.1093/biostatistics/kxm030.

31. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48. doi:10.1186/1471-2105-10-48.

32. Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ravi Ram K, et al. (2007) Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila. Genetics 177: 1321–1335. doi:10.1534/genetics.107.078865.

33. Wang P, Lyman RF, Mackay TFC, Anholt RRH (2009) Natural Variation in Odorant Recognition Among Odorant-Binding Proteins in Drosophila melanogaster. Genetics 184: 759–767. doi:10.1534/genetics.109.113340.

34. Laughlin JD, Ha TS, Jones DNM, Smith DP (2008) Activation of Pheromone-Sensitive Neurons Is Mediated by Conformational Activation of Pheromone-Binding Protein. Cell 133: 1255–1265. doi:10.1016/j.cell.2008.04.046.

35. Sun Y-L, Huang L-Q, Pelosi P, Wang C-Z (2012) Expression in Antennae and Reproductive Organs Suggests a Dual Role of an Odorant-Binding Protein in Two Sibling Helicoverpa Species. PLoS ONE 7: e30040. doi:10.1371/journal.pone.0030040.

36. Cerenius L, Kawabata S, Lee BL, Nonaka M, Söderhäll K (2010) Proteolytic cascades and their involvement in invertebrate immunity. Trends Biochem Sci 35: 575–583. doi:10.1016/j.tibs.2010.04.006.

37. Van Hoef V, Breugelmans B, Spit J, Simonet G, Zels S, et al. (2013) Phylogenetic distribution of protease inhibitors of the Kazal-family within the Arthropoda. Peptides 41: 59–65. doi:10.1016/j.peptides.2012.10.015.

38. Karvonen A, Seehausen O (2012) The Role of Parasitism in Adaptive Radiations—When Might Parasites Promote and When Might They Constrain Ecological Speciation? Int J Ecol 2012: 1–20. doi:10.1155/2012/280169.

39. An X (2004) The cryptochrome (cry) Gene and a Mating Isolation Mechanism in Tephritid Fruit Flies. Genetics 168: 2025–2036. doi:10.1534/genetics.104.028399.

40. Mullen SP, Shaw KL (2014) Insect Speciation Rules: Unifying Concepts in Speciation Research. Annu Rev Entomol 59: 339–361. doi:10.1146/annurev-ento-120710-100621.

41. Peixoto AJ, White WB (2007) Circadian blood pressure: Clinical implications based on the pathophysiology of its variability. Kidney Int 71: 855–860. doi:10.1038/sj.ki.5002130.

42. Felton GW, Summers CB (1995) Antioxidant systems in insects. Arch Insect Biochem Physiol 29: 187–197. doi:10.1002/arch.940290208.

43. Sheehan D, Meade G, Foley VM, Dowd CA (2001) Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. Biochem J 360: 1–16.

Supplementary Material

Table S1. Sequencing effort per library

Libraries Profile		Replicates	Reads	
			A. fraterculus	*A. obliqua*
Female	Virgin	A	6,500,046	7,497,373
		B	6,822,474	6,891,525
	Post mating	A	7,183,605	8,335,243
		B	9,149,410	6,710,548
	Post oviposition	A	6,531,891	8,208,959
		B	9,420,206	8,649,653
Male	Virgin	A	7,736,532	6,195,410
		B	8,683,757	7,021,347
	Post mating	A	11,364,858	7,840,733
		B	8,388,907	6,808,349
Total by species			81,781,686	74,159,140
Total			155,940,826	

Table S2. Total number of filtered reads, percentage of reads retained and total number of bases produced per library and per species.

Libraries		Rep.	*A. fraterculus*						*A.obliqua*					
			Reads	**%**	**Bases**	**%**	**PE1***	**PE2***	**Reads**	**%**	**Bases**	**%**	**PE1***	**PE2***
Female	Virgin	A	5,652,273	86.95	1,086,767,944	83.59	96	96	6,684,039	89.15	1,278,769,665	85.28	95	95
		B	6,280,265	92.05	1,205,934,440	88.37	95	96	6,171,699	89.55	1,178,119,819	85.47	95	95
	Post mating	A	6,186,816	86.12	1,173,833,135	81.70	97	92	7,520,056	90.22	1,436,357,133	86.16	95	95
		B	8,519,269	93.11	1,647,889,520	90.05	96	96	6,066,139	90.39	1,152,296,056	85.85	94	95
	Post oviposition	A	5,492,619	84.08	1,034,443,269	79.18	97	91	7,345,410	89.48	1,406,454,916	85.66	95	95
		B	8,773,354	93.13	1,697,983,441	90.12	96	96	7,811,619	90.31	1,492,531,036	86.27	95	95
Male	Virgin	A	7,241,354	93.59	1,401,642,013	90.58	96	96	5,661,286	91.37	1,090,347,302	87.99	96	96
		B	8,021,534	92.37	1,548,755,751	89.17	96	96	6,422,330	91.46	1,236,636,692	88.06	96	96
	Post mating	A	9,663,362	85.02	1,827,937,120	80.42	97	92	7,004,874	89.33	1,340,493,523	85.48	95	95
		B	7,806,393	93.05	1,510,592,958	90.03	96	96	6,168,962	90.60	1,186,715,711	87.15	96	96
Total by species			73,637,239		14,135,779,591				66,856,414		12,798,721,853			

* Average size by sequence reading of paired-end (PE).

CAPÍTULO 4 – MANUSCRITO 2

Candidate genes involved in differentiation between two non-model species of fruit flies (*Anastrepha, Tephritidae*) screened from head transcriptomes

Victor Borges Rezende[1]
Email: victorrez85@yahoo.com.br

Carlos Congrains Castillo[1]
Email: carlos_congrains@outlook.com

André Luís A. Lima[1]
Email: andreluisbio@gmail.com

Emeline Boni Campanini[1]
Email: emelinebc@gmail.com

Aline Minali Nakamura[1]
Email: alinemnk@gmail.com

Janaína Lima Oliveira[1]
Email: janajpb@hotmail.com

Samira Chahad-Ehlers[1]
Email: schahad@gmail.com

Iderval Sobrinho Junior[1]
Email: iderval_sobrinho_jr@yahoo.com

Reinaldo Alves de Brito[1*]
[*]Corresponding author
Email: brito@power.ufscar.br

[1]Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, Brazil

Abstract

Background: Several species of the *Anastrepha* fruit flies are of great economic importance for causing damage to a variety of fleshy fruits. Some of the most important species in the genus belong to the closely related and recently diverged "*fraterculus* group". Some species in this group have similar morphological attributes that render their identification difficult, reinforcing the relevance of identifying new genetic/molecular markers that may differentiate species. In this work, we investigate patterns of expression in head tissues from two closely related *Anastrepha* species (*Anastrepha obliqua* and *A. fraterculus*) to identify single nucleotide polymorphisms (SNPs) and transcripts under positive selection that may help us discriminate these species. To do so, we built multiple Next Generation Sequencing (NGS) libraries from head tissues of these species, at different reproductive stages, for both sexes.

Results: cDNA libraries were generated from total RNA extracted from fly heads pooled according to species, gender and reproductive stages, totaling 10 libraries per species. Almost 156 million reads of 100 paired-end sequences were generated by runs on an Illumina HiSeq 2000. The Illumina reads were trimmed and filtered by quality and about 10% of all reads were discarded. The remaining 140 million reads were assembled using the Trinity short read assembler, which resulted in 154,787 contigs with N50 of 2012 and an average length of 1,027 bp. Over 45,000 contigs had more than 1,000 bp and almost 23,000 had more than 2,000 bp, with the longest one being 25,704 bp. The use of SAMtools and Bowtie in the alignments of each species' reads mapped against total assembly identified 62,973 intraspecific SNPs that were polymorphic only in *A. fraterculus*, 32,616 in *A. obliqua* and 6,698 SNPs in both. We found 2,822 contigs with at least one SNP present in both species, and 175 of these had fixed interspecific differences between species. A positive selection analysis on these transcripts resulted in 12 interesting genes that may be involved with group differentiation, among them a serine protease (*Ser6*), an odorant-binding protein (*Obp99a*) and Transferrins (*Tfr*).

Conclusions: These results generate a set of candidate genes that are potentially important to help us understand the evolution and differentiation of *A. obliqua* and *A. fraterculus* which might help us identify genetic markers that would be relevant to study the evolution and identification of species in the *fraterculus* group.

Keywords: Transcriptome, RNA-Seq, *De novo* assembly, Next generation sequencing, SNP calling, Tephritidae, *Anastrepha*, *Fraterculus* group, Head tissue.

Background

New technologies, such as Next Generation Sequencing (NGS), are revolutionizing the ability to produce large scale genetic data (1) in short time and at a low price, enabling the *de novo* sequencing of entire eukaryotic genomes and transcriptomes (2). Furthermore, studies on transcriptomes from non-model organisms are becoming more achievable and computationally tractable, even more so than genome projects (3), although a high computational effort is still required (4).

RNA-seq strategies have provided a powerful tool for identifying new sequences which may be useful for genetic as well as evolutionary studies, such as *de novo* transcript determination (2, 5–7), and the identification of coding SNPs (8–10) of non-model organisms, such as *Anastrepha* fruit flies (Diptera: Tephritidae). These flies are economically important pests because of the serious damages they inflict to a wide variety of fruits caused by oviposition and larval growth. *Anastrepha* is a Neotropical genus distributed from southern regions of United States to South America, except in Chile (11, 12). The genus is comprised of 237 species which are morphologically divided in 17 species groups (13, 14). The "fraterculus group" consists of 29 species, some of them cryptic owing to their morphological and genetic similarities (14, 15). *Anastrepha obliqua* and *A. fraterculus* are closely related species of the "fraterculus group" that because of their broad distribution and wide host range are two of the most economically important species in the genus (16, 17). In spite of their relevance, there is a lack of evolutionary and genetic studies not only on these species, but in other species of the *fraterculus* group as well. Traditional taxonomy is complicated by an overlap of certain morphological attributes (16), even though there is some host preference and some species show specific timing of mating (14), which may indicate early stages of speciation in phytophagous insects (18, 19). Molecular studies performed to date have also failed to identify species specific markers and monophyletic assemblages separating species, using mtDNA markers (15, 20) or nuclear markers (21–24).

We seek to understand the genes involved with the speciation process in the "*fraterculus* group", aiming to identify candidate genes and evolutionary forces involved with their differentiation. Since the species under study are recently diverged, we looked for rapidly evolving genes that would have a better chance of tracking the groups' differentiation. There is very limited genetic data available for the species here studied, so we generated Next-Generation transcriptome

libraries of different tissues in order to identify rapidly evolving genes and potential species-specific SNPs. We chose the cephalic region of *A. obliqua* and *A. fraterculus* because of the presence of the brain and several sensorial and visual organs. Pheromones and their receptors, as well as other olfactory receptors in general, may be produced and stored in the salivary glands of the head as well as other parts of the body (25, 26). In addition, tephritids exhibit a peculiarity that distinguishes some members of the "*fraterculus* group" from most other insects. While in the majority of insects, it is common for females to produce sex pheromone to attract males, in these fruit flies the males are the ones who produce sex pheromone to attract females instead (25). It is possible then that some of the genes involved in the determination of interspecific differences between *A. fraterculus* and *A. obliqua* are expressed in the cephalic region, not only because of ecological but also because of reproductive choices. The transcriptomes studies allow us to identify markers distinguishing these species, but also to make evolutionary analyses based on the direction and magnitude of natural selection within and between species.

Methods

Samples, RNA extraction and cDNA library

We extracted RNA from head tissues of from *Anastrepha fraterculus* and *A. obliqua* at different reproductive stages. Virgin and postmating males and females were sampled as well as female post oviposition. The description of populations, samples, and molecular procedures are presented elsewhere (Rezende,VB; Congrains, CC; Lima, ALA; Campanini, EB; Nakamura, AM; Oliveira, JL; Chahad-Ehlers, S; Sobrinho Jr, I; de Brito, RA; unpublished data). All libraries were produced with two replicates totaling 20 cDNA libraries.

Sequence cleaning and assembly

The cleaning of the reads for quality on SeqyClean software, assembly with Trinity and functional annotation of the generated contigs are detailed elsewhere (Rezende,VB; Congrains, CC; Lima, ALA; Campanini, EB; Nakamura, AM; Oliveira, JL; Chahad-Ehlers, S; Sobrinho Jr, I; de Brito, RA; unpublished data). We assembled the reads per species, combining the 10 libraries produced into a single pooled assembly, which was used as a reference transcriptome for orthology determination and SNP calling for each species, which were independently screened for selection as described below. We also combined all 20 libraries into a single pooled assembly to investigate for interspecific SNPs.

Mapping reads against a reference transcriptome and SNP discovery

The Bowtie software (version 1.0.0) (27) was used to align the Illumina reads back to the reference transcripts assembled by Trinity. The default parameters were used as implemented in the Trinity software package (28). All reads were mapped against the reference transcriptome generating a general reference map to call SNPs for both species in conjunction. We also independently mapped species-specific reads against the same reference transcriptome to call SNPs from each species. The resulting files were screened for SNPs with SAMtools mpileup (29) with the option that includes a per-sample read depth, and SNPs were recovered with Bayesian inference with bcftools view (29). The SNPs were selected using custom scripts and only those with minimum Phred-scaled read quality of 30 and minimum read coverage of 100 were retained.

Differentiation level between the transcripts of *A. fraterculus* and *A. obliqua*

In order to select a group of candidate transcripts showing higher level of divergence between *A. fraterculus* and *A. obliqua*, we estimated the interspecific differentiation index (D and $\overline{D}$) (30, 31) using Python scripts. D is defined as the absolute value of the difference among the allelic frequencies of a SNP variant of *A. fraterculus* and *A. obliqua* ($D = |F_{Af} - F_{Ao}|$). $\overline{D}$ is the average D value for SNPs from a particular transcript. We plotted the distribution of $\overline{D}$ and used $\overline{D} \geq 0.94$ as the threshold value to separate a group of transcripts, which will be evaluated the evolutionary rates of *p*N/*p*S.

Ka/Ks calculation

We used Trinotate to generate likely CDSs of the transcripts in the assembly of both species with $\overline{D} > 0.94$ and each individual assembly, for *A. fraterculus* and *A. obliqua*. tBLASTx algorithm was used to find potentially orthologous sequences for the previously selected transcripts in CDSs of transcripts from *Ceratitis capitata* RefSeq project #PRJNA201381 and the *A. fraterculus* and *A. obliqua* assemblies. Orthologous CDSs were selected with the highest bit score of the hits with an e-value < 0.005. Each sequence set was translated to aminoacid, aligned using the Muscle algorithm (32), then back-translated to the original nucleotide sequence implemented by TranslatorX (33). Pairwise Ka/Ks was estimated for each species pair and

orthologous set using the MS model (34) by the KaKs calculator program (35). Genes showing Ka/Ks rates higher than 0.5 were considered as potentially evolving under positive selection.

We show a schematic workflow for our protocol to select candidate genes that show alleles likely involved in the differentiation process of two closely related non-model species using RNA-seq data (Figure 1).

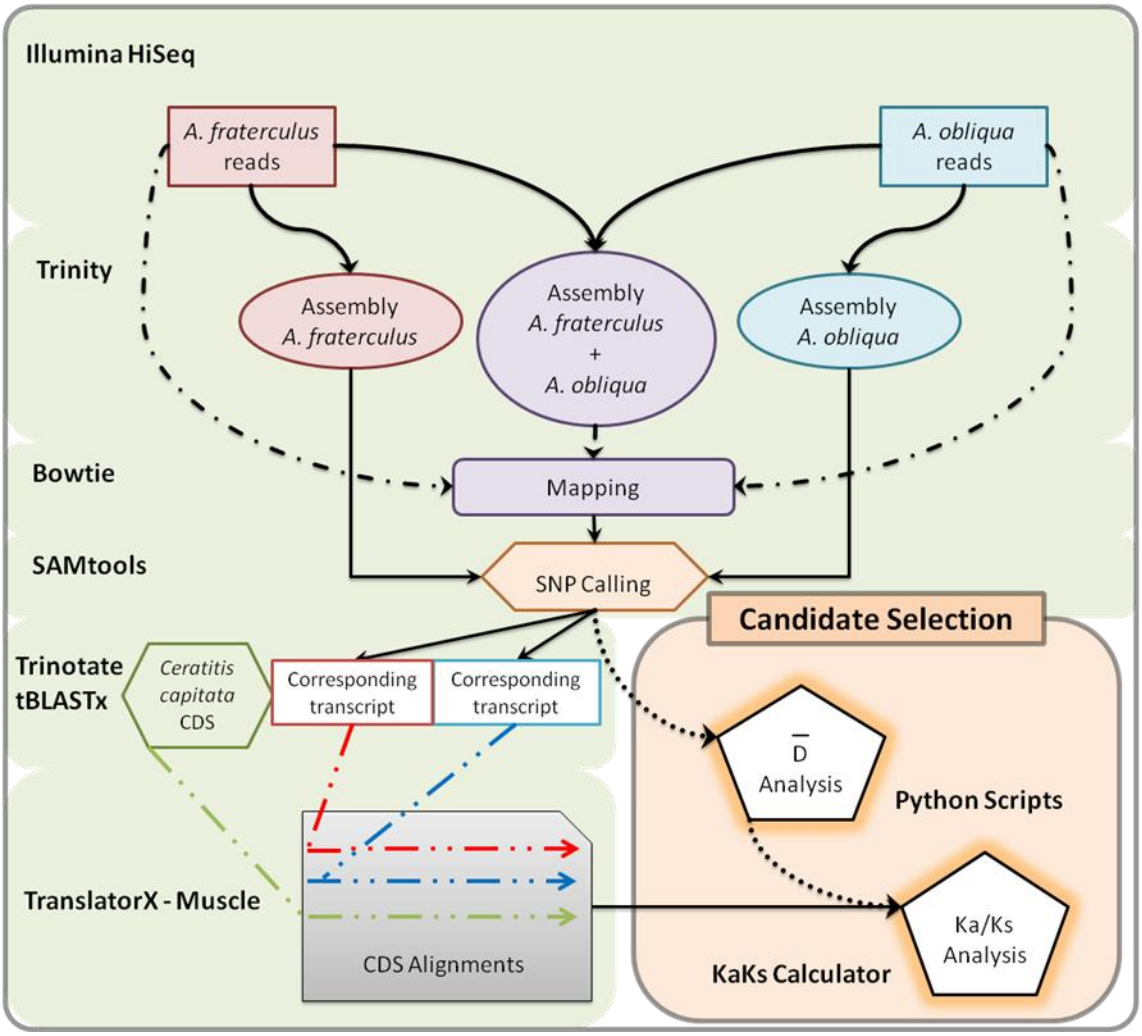


Figure 1 Framework for screen transcriptome libraries and estimate Ka/Ks in differentiation candidate genes.

Results

Transcriptome analyses

The 20 cDNA libraries (5 profiles per species with replicates) generated 155,940,826 raw paired-end reads, with length of 100bp, which were then filtered by quality, resulting in 140,493,653 paired-end reads and over 28 giga base pair. The libraries produced more than 7 million reads per replicate for *A. fraterculus* and 6,68 million for *A. obliqua* with an average of 95 base pair per read for both species. We processed this data in Trinity to create a pooled assembly which had a total of 154,787 contigs. We also assembled the different libraries per species producing separate assemblies that resulted in 112,862 and 98,549 contigs for *A. fraterculus* and *A.obliqua*, respectively. A summary of the transcriptome data, including the results of Illumina sequencing and assemblies is shown on Table 1. A detailed characterization of cDNA libraries of *A. fraterculus* and *A. obliqua* and the full assembly information that was basis to this work are present elsewhere (Rezende,VB; Congrains, CC; Lima, ALA; Campanini, EB; Nakamura, AM; Oliveira, JL; Chahad-Ehlers, S; Sobrinho Jr, I; de Brito, RA; unpublished data).

Table 1. Summary of transcriptome head libraries and assemblies

Result of Illumina sequencing			
Total of Paired-end reads	168,716,434		
Result of Trinity assembly			
	A. frat + A. obli	*A. fraterculus*	*A. obliqua*
Filtered reads*	140,493,653	73,637,239	66,856,414
Assembly Length Distribution			
Total number of contigs	154,787	112,862	98,549
N50	2,012	2,504	2,637
Contigs longer than 1000 bp	45,602	38,105	34,982
Contigs longer than 2000 bp	22,297	21,054	19,964
Contigs longer than 10000 bp	213	329	269
Assembly Length Statistics			
Average contig length (bp)	1,027	1,196	1,254
Median contig length (bp)	494	535	563
Shortest contig (bp)	201	201	201
Longest contig (bp)	25,704	27,513	22,394

* Reads were filtered based on the quality (maximum average error of 0.01, maximum error at the ends of 0.05) and length larger than 50bp.

SNP discovery and interspecific allele differentiation

We used SAMtools and Bowtie to map reads from each species against the overall assembly and then filtered for a coverage higher than 100 and a Phred score of 30. We identified 62,973 SNPs in 2,773 contigs that were exclusive to *A. fraterculus*, 32,616 SNPs in 1,840 contigs exclusive to *A. obliqua* and 6,698 SNPs present in 2,822 contigs common to both species. As a measure of differentiation between *A. fraterculus* and *A. obliqua*, we searched for SNPs with highly differentiated frequencies, indicated by differentiation indexes (D). We estimated a mean differentiation index ($\overline{D}$) for each contig, which considered the contribution of all SNPs' D values per contig and identified 175 highly differentiated contigs using such statistics The distribution of $\overline{D}$ frequencies had a bell-shaped distribution, with the exception of the most differentiated transcripts, which apparently showed an outlier pattern (Figure 2).

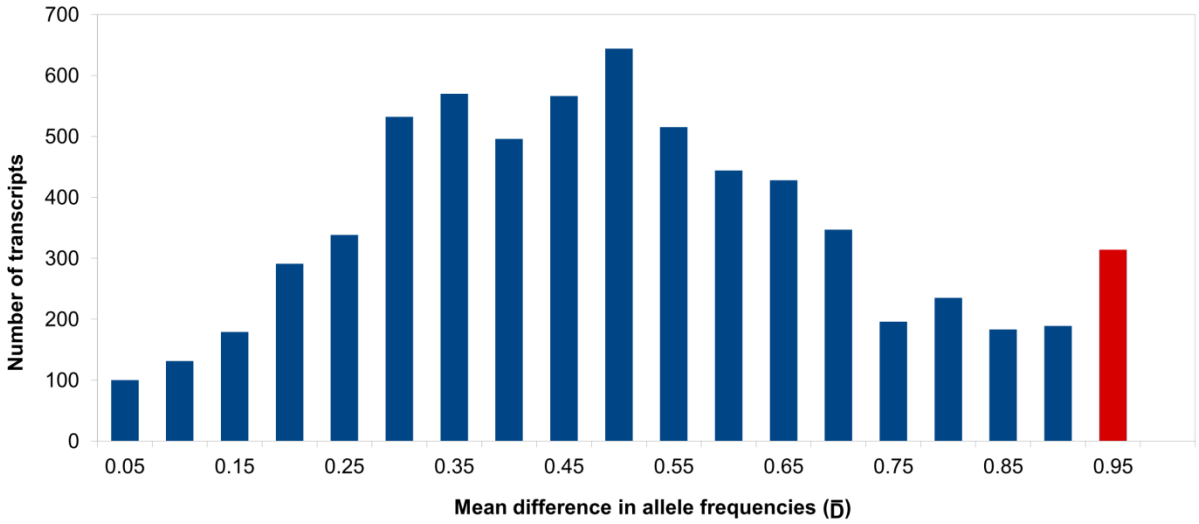


Figure 2 Frequency distribution of $\overline{D}$. Distribution of the average allelic frequency differences between *A. fraterculus* and *A. obliqua* per transcript ($\overline{D}$) in 6,698 SNPs across 2,822 transcripts. The red bar represents a deviation in a bell-shape distribution and the data chosen for positive selection analysis.

Selective pressure on highly differentiated transcripts

The highly differentiated contigs were isolated and submitted to a Ka/Ks analysis by identifying each corresponding transcript and extracting the main coding DNA sequence (CDS). Each species' CDS was used to align to the other species' as well as to the corresponding CDS from *Ceratitis capitata*, and these alignments provided three pairwise rates of Ka/Ks per transcript as result, among each pair of species (Table 2). Since our main goal was to identify genes that might be involved with the differentiation between the two *Anastrepha* species, we looked for

genes that would show low values of Ka/Ks in the branch that connects these species to *Ceratitis* (Ka/Ks < 0.5), but that would show higher rates (Ka/Ks ≥ 0.5) in the branch between *A. fraterculus* and *A. obliqua*, which might suggest that these genes were being positively selected only between the closely related species pair under study.

Table 2 Pairwise Ka/Ks analysis.

Transcript	Gene Ontology Annotation	Pairwise Ka/Ks *		
		A.frat x A.obli	***A.frat x C.capi***	***A.obli x C.capi***
comp115107_c0_seq1	*Serine protease 6*	1.2575(0.024/0.019)	0.1254(0.305/2.429)	0.0975(0.297/3.047)
comp112616_c0_seq1	*MTERF domain containing 1*	0.7382(0.007/0.01)	0.1067(0.167/1.564)	0.1051(0.164/1.564)
comp118949_c0_seq1	*Extra macrochaetae*	0.7369(0.079/0.107)	0.1489(0.17/1.143)	0.1083(0.102/0.945)
comp110683_c0_seq1	*CG16817*	0.7325(0.007/0.009)	0.0747(0.155/2.079)	0.0819(0.156/1.908)
comp111330_c0_seq1	*Rab GTPase*	0.7167(0.002/0.003)	0.0826(0.168/2.03)	0.0814(0.169/2.069)
comp122022_c1_seq1	*Cytop. linker prot 190*	0.6996(0.068/0.097)	0.1266(0.142/1.121)	0.1488(0.18/1.208)
comp120945_c0_seq1	*Transferrin 3*	0.6494(0.007/0.01)	0.0528(0.088/1.674)	0.0522(0.088/1.684)
comp122327_c0_seq1	*GPI-anchored cell glycoprotein*	0.6476(0.012/0.018)	0.1342(0.243/1.806)	0.1243(0.242/1.947)
comp114185_c0_seq1	*Odorant-binding protein 99a*	0.6038(0.03/0.049)	0.1280(0.122/0.953)	0.1401(0.124/0.887)
comp123860_c0_seq1	*Microtubule-associated prot 205*	0.5924(0.018/0.03)	0.2877(0.328/1.139)	0.2762(0.322/1.166)
comp122776_c0_seq3	*Legless*	0.5727(0.02/0.035)	0.0823(0.166/2.019)	0.1068(0.201/1.886)
comp114253_c0_seq1	*Mitochondrial RPS2*	0.5462(0.01/0.017)	0.0343(0.083/2.425)	0.0311(0.083/2.675)

A. frat - *A. fraterculus*; A. obli - *A. obliqua*; C. capi - *C. capitata*.
* Ka/Ks rate are present out of parenthesis and within parenthesis are the values of Ka and Ks.

Using a Model Selection framework (34), we identified 12 transcripts with these attributes (Figure 3). An analysis of the sequences of these 12 transcripts revealed that five of them had a synonymous substitution on the position of the analyzed SNP that was initially associated with the $\overline{D} > 0.94$ and the remaining seven transcripts, with non synonymous substitution, became final candidates for being involved in the differentiation between the two species of *Anastrepha*. The candidate transcripts, their respective annotation and D values and the substitution analysis are presented on Table 3.

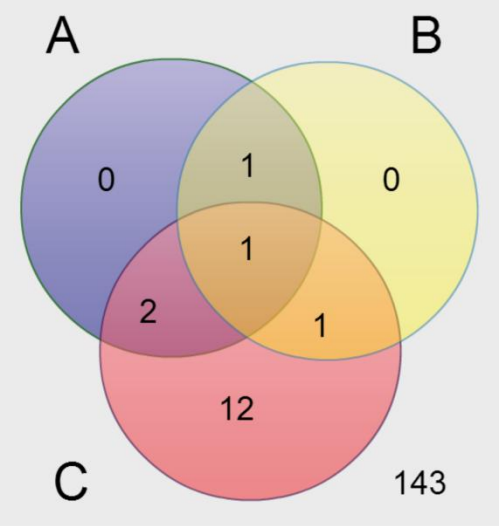


Figure 3. Venn diagram summarizing the pairwise Ka/Ks results for candidate genes with $\overline{D}$ > 0.94. A) Pairwise comparison between *A. fraterculus* and *C. capitata*. B) Pairwise comparison between *A. obliqua* and *C. capitata*. C) Pairwise comparison between *A. fraterculus* and *A. obliqua*.

Table 3 Annotation of the transcripts with $\overline{D}$ > 0.94 analysis and change in aminoacid substitution of associated SNP. All transcripts have one SNP that belong to CDS region.

Transcript	Gene Ontology Annotation	$\overline{D}$	Aminoacid Substitution		Substitution Type
comp115107_c0_seq1	*Serine protease 6*	0.950	Gln	Gln	S
comp112616_c0_seq1	*MTERF domain containing 1*	0.950	Arg	Ser	N
comp118949_c0_seq1	*Extra macrochaetae*	0.950	Glu	Glu	S
comp110683_c0_seq1	*CG16817*	0.949	Ser	Arg	N
comp111330_c0_seq1	*Rab GTPase domain-containing protein*	0.949	Ala	Pro	N
comp122022_c1_seq1	*Cytop. linker prot 190*	0.950	Ile	Leu	N
comp120945_c0_seq1	*Transferrin 3*	0.943	Leu	Ser	N
comp122327_c0_seq1	*GPI-anchored cell glycoprotein*	0.950	Leu	Leu	S
comp114185_c0_seq1	*Odorant-binding protein 99a*	0.950	Thr	Ile	N
comp123860_c0_seq1	*Microtubule-associated protein 205*	0.949	Glu	Glu	S
comp122776_c0_seq3	*Legless*	0.949	Cys	Ser	N
comp114253_c0_seq1	*Mitochondrial RPS2*	0.950	Thr	Thr	S

*N: Nonsynonymous substitution and S: Synonymous substitution

Functional Annotation

The 175 transcripts with $\overline{D}$ > 0.94 were blasted against the gene ontology database and distributed in 40 different classes among the 3 great gene ontology classes by GOTermMapper tool. These three main classes cover the domains that belong to the parts of a cell or its extracellular environment (cellular component), the activity of a gene product at the molecular level (molecular function) and the function of certain level of living units (biological process). These distributions are shown in Figure 4. For the three main classes: biological process, molecular function and cellular component, the most frequent terms are "Anatomical structure development" (28,37%), "Ion binding" (23,4%) and "Cell" (63,12%), respectively. Even though, cellular component has five other terms with a larger than 20% representation, only in the Biological process class there another category at that percentage (Figure 4). This possibly reflects the smaller number of classes found in cellular component category The best inferred potential ORFs for the 175 transcripts were also blasted against a database of ESTs from *Drosophila melanogaster* (155 transcripts or 88,5% of homologous hits) and *Ceratitis capitata* (159 transcripts or 90,8% of homologous hits).

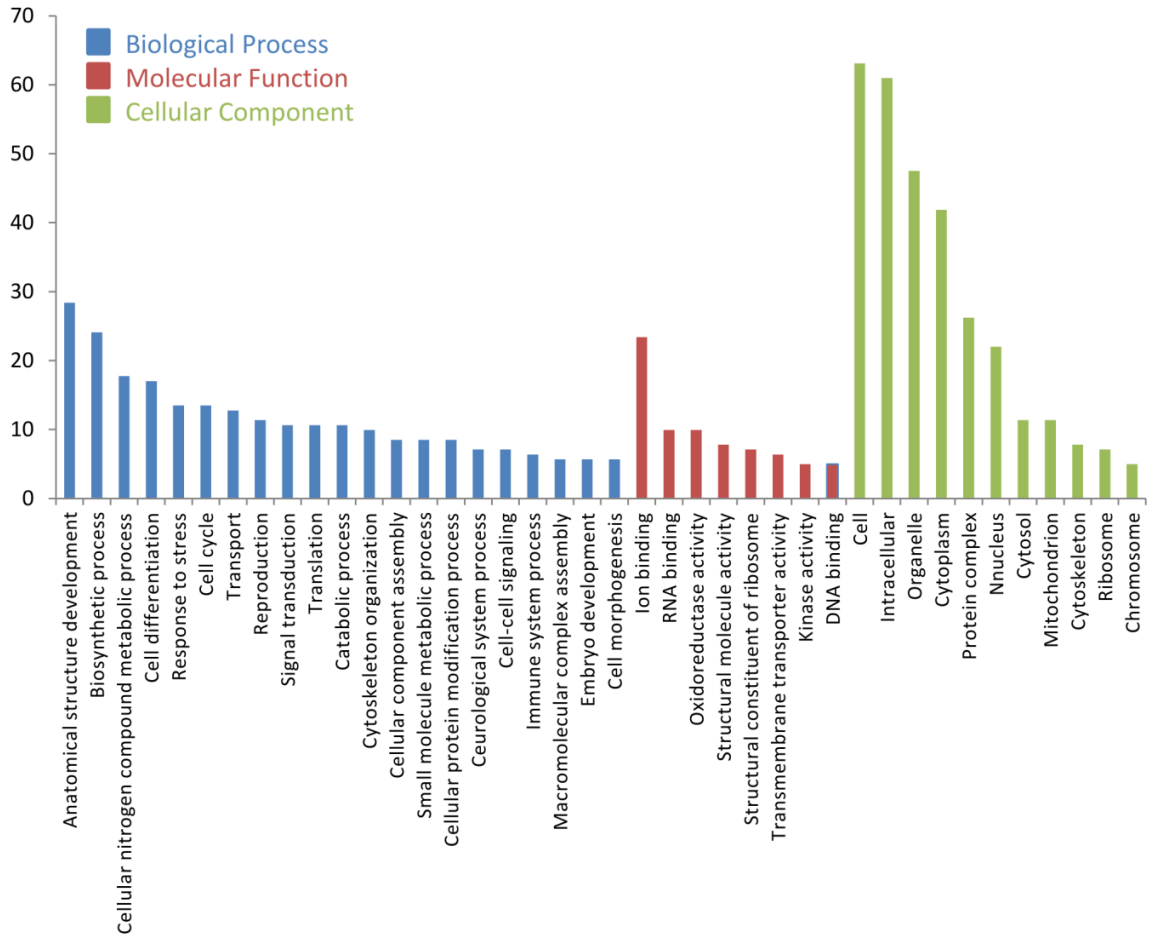


Figure 4 Gene onthology (GO) classification of the 175 transcripts with differentiation parameter ($\overline{D}$)> 0.94. Transcripts divided in GO main categories. 112 transcripts annotated in Biological Process, 98 transcripts annotated in Molecular function and 92 transcripts in Cellular Components.

Discussion

To investigate for candidate genes involved in differentiation between two closely related non-model species we generated a large amount of next-generation transcriptomes and finished with seven contigs with high level of divergence between species that bear fixed non synonymous SNPs.

The large raw data here generated was systematically checked for contaminations, adaptors and quality of the reads, so what is left was reliable data as input for two strategies to do the assembly. In spite of the smaller number of contigs generated in individual species assembly,

these produced better results than the full data assembly, combining both species, as evaluated by the average, median and N50 size of contigs (Table 1). Even though this could be caused by erros associated with the combination of data from different, divergent, species, it may also have been caused by limitations caused by the analysis of a large amount of data. The N50 index is considered to be a good parameter to evaluate the quality of assembly when the reference genome are made from closely related species under the justified assumption that gene length is usually well conserved (36) and such as this, our full *de novo* assembly N50 are better than other similar transcriptomes (37–40). Another parameter which tends to confirm the quality of the assembly is the percentage of contigs annotated in closely related species like *Drosophila melanogaster* (88,5%) and *Ceratitis capitata* (90,8%), which we achieve for the database of ESTs generated from the 175 contigs selected by the evolutionary analysis.

However, we analyzed large scale data from non-model organisms that may have some potential errors associated. The assembly is the critical phase that will provide the raw materials for the following stages. We performed the assembly using Trinity software that showed to be a powerful tool to deal with redundant data with alternative splicing typically found in transcriptome data (41, 42). Nevertheless, subtle errors in the assembly process may cause erroneous contig inferences. A lot of these assembly problems would cease to exist if we had a reference genome of *Anastrepha*, which would also help to evaluate the real genetic variability for the widely distributed species here studied. Because we used a single population per species, there may be some biases in the $\overline{D}$ estimates, but our main objective here was to provide highly differentiated candidate genes, not an overall estimate of variability for the species that would probably require the use of fine scale strategies including geographically widespread samples to corroborate the role of these genes in speciation process.

We provide a pipeline to perform a screening of transcriptomes data from *A. fraterculus* and *A. obliqua* that allowed for the identification of 12 genes showing highly differentiated alleles that have potentially evolved under positive selection between two closely related fruit flies species. We used D and $\overline{D}$ as parameters that evaluate the allelic frequency to estimate the genetic differentiation between the investigated species. We choose a robust threshold of $\overline{D}$ (0.94) to select only the genes with variants practically fixed among *A. fraterculus* and *A. obliqua* (30, 31). Because we only considered a single population for our analysis, as previously discussion, it

is possible that these estimates fail to consider within species variation, but other studies have reported low levels of fixed intraspecific variation in species of the fraterculus group (43).

The non-synonymous or missense variant is a single base change in a coding region that causes an amino acid change in a corresponding protein. If this non-synonymous SNP alters protein function, the change can have drastic phenotypic consequences (44). When these alterations produce more advantageous copies in a specific environment, the SNP will be favorably selected and increase its frequency in the population. We only considered here non-synonymous changes, even though some synonymous changes have been found to be differentially selected (45, 46), because we were dealing with coding regions, and because these changes are more likely to promote radical changes in the protein. This constraint further reduced our list of 12 genes highly differentiated genes to the seven annotated genes with non-synonymous substitution, which are the final candidates to be involved to the differentiation between *A. fraterculus* and *A. obliqua*.

Out of the seven annotated candidates potentially involved with species differences, *MTERF domain containing 1* (*CG15390*), *CG16817, Rab GTPase domain-containing protein* and *Cytoplasmic linker protein-190* (*CLIP-190*) will not be extensively discussed here due to the nonexistence of sufficient knowledge in the literature that would allow us to associate them to the likely causes of differentiation between species.

Based on the BLOSUM62 table (47), the most radical amino acid substitution among the seven annotated sequences occurs in *Transferrins* (*Tsf*), in which the change between a leucine (Leu), one of the most aliphatic amino acids, and a serine (Ser), the simplest hydroxyl amino acid, had a -2 score, probably causing an important change in protein properties. *Tsf* is one of the classes of proteins involved in iron transport and storage. They are found in multicellular animals and are characterized by a high capacity for iron storage (48). Based on studies of *Tsf* from several insect species, the following roles have been proposed for insect *Tsf*: iron transport, antibiotic agent, vitellogenic protein, and juvenile hormone regulated protein (49). In particular, *Transferrin 3* (*Tsf3*) has not been well studied yet, but it has a putative function on *Drosophila melanogaster* circadian biology (50). Considering that there is evidence for differences in reproductive activity between *A. fraterculus* and *A. obliqua*, it is possible that genes involved with the circadian rhythm may be involved with the species differentiation. Overall, *Tsf3* has reduced expression

during several developmental stages and tissues in *Drosophila melanogaster*, but it reaches a high expression peak in white prepupae salivary gland (51).

There is also evidence for positive selection strongly acting on an *odorant-binding protein* (*OBP*). Proteins of the *OBP* gene family are the first component of the insect olfactory system, solubilizing and carrying the chemical signals from the environment to the odorant receptors (52). *OBP*s play a crucial role in the survival and reproduction of individuals, and they are involved in the adaptation of insects to a wide variety of environments and lifestyles (53). There has been evidence that positive selection has been involved in the evolution of the *OBP* gene family in insects (54–57), suggesting that changes in *OBP*s may result in changes in olfactory behaviours.

We found positive selection in the gene encoding the *OBP99a* in the comparison between *A. fraterculus* and *A. obliqua*, but not between any *Anastrepha* and *C. capitata*. This *OBP* (as well as *OBP99c* and *OBP99d*) is responsible to recognize benzaldehyde (58), an aromatic compound that is known as an insect repellent (59). However, little is known about how the molecular variation that occurs in the *OBP* genes affects the individual variation in olfactory behavior (60). One possible cause of this nonneutral evolution observed can be an ecological adaptation: the increase in *Ka*/*Ks* invoke changes in the selective environment experienced by species (55). *OBPs* might evolve rapidly to adapt to changes in the chemical environment, and the interaction between the chemical environment and a population is not constant (58). Maybe *A. fraterculus* and *A. obliqua* have experienced different ecological constraints, since the separation of these two lineages. However, we might have expected that the same thing would have happened between these species and *C. capitata*, which was not revealed by our results, but it is possible that the changes in these genes have been so extense that they have eliminated some of the signal of positive selection, due to homoplasy. The percentage of divergence between the two *Anastrepha* species for this *OBP* is 3,2%. However, between these species and *C. capitata* the average divergence is approximately 25%, which may not be sufficient to lead to overall homoplasy, but it may lead to homoplasy in the regions that evolving more rapidly, which could explain our results.

Another cause can be because positive selection frequently is involved in the functional differentiation of copies in gene duplication events in a gene family (52). After the speciation of

A. fraterculus and *A. obliqua*, the *OBP99a* gene may have experienced different functional divergence in different species, like subfunctionalization or neofunctionalization. Positive selection generally acts to promote functional divergence modifying the binding specificities, or altering the conformational changes involved in the *OBP* functional mechanism, such as shown in some pheromone-binding proteins (61–63).

Legless (*Lgs*) (*Drosophila* ortholog of *BCL9*) is a nuclear factor that shows a moderate expression in adult brain of *D. melanogaster* (51). The gene contains a functional nuclear localization signal (NLS) and interacts with other nuclear components (64) for targeting regulatory region of Wingless (Wg)/WNT target genes (65). Essential for such interactions are two important conserved *Lgs* domains: HD2 (homology domain 2) and HD1 (64). Mutations on Lgs lead to lethal or severe effects on embryonic development. Depending on the locus, adult mutant flies may have one or both metathoracic legs absent (the legless phenotype) and CNS defects (66). It is being suggested that, like the *runt* gene, *leg* may also exhibit an expected rate of interspecific divergence with low intraspecific polymorphism (67, 68).

Considering the potential candidate genes here identified, we not only identified genes that showed patterns of molecular changes that are consistent with changes driven by positive selection, but we also identified the portion of such genes that are differentially expressed between two closely species, *A. fraterculus* and *A. obliqua*. Therefore, we have moved one step closer to our quest of identifying genes that are involved with species differences in this important group of fruit flies and we should now confirm the potential of these candidate genes not only as species-specific markers when considering other populations, but also the potential for these genes to help us understand processes affecting other species in the fraterculus group, as well as other *Anastrepha*.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Rezende, Congrains, Sobrinho and de Brito conceived and designed the research project. Rezende, and de Brito wrote the manuscript. Rezende, Congrains, Lima, Campanini, Nakamura, Oliveira, Chahad-Ehlers, Sobrinho and Brito worked in the establishment and maintenance of the flies populations in the laboratory. Rezende, Congrains, Sobrinho and de Brito performed the bioinformatics and evolutionary analysis containing assembly *de novo*, mapping, annotation, SNP calling and Ka/Ks calculator. Congrains, Campanini, Oliveira, Chahad-Ehlers, Sobrinho contributed to the development of the project and writing of the manuscript. All authors read and approved the final manuscript.

References

1. Martinez DA, Nelson MA: **The next generation becomes the now generation**. *PLoS Genet* 2010, **6**:e1000906.

2. Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Alex Buerkle C: **Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies**. *Mol Ecol* 2010:no–no.

3. Kumar S, Blaxter ML: **Comparing de novo assemblers for 454 transcriptome data**. *BMC Genomics* 2010, **11**:571.

4. Schwartz DC, Waterman MS: **New generations: Sequencing machines and their computational challenges**. *J Comput Sci Technol* 2010, **25**:3–9.

5. Davey JW, Blaxter ML: **RADSeq: next-generation population genetics**. *Brief Funct Genomics* 2010, **9**:416–423.

6. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I: **De novo assembly and analysis of RNA-seq data**. *Nat Methods* 2010, **7**:909–912.

7. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57–63.

8. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nat Rev Genet* 2011, **12**:443–451.

9. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease-causing potential of sequence alterations**. *Nat Methods* 2010, **7**:575–576.

10. Cánovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF: **SNP discovery in the bovine milk transcriptome using RNA-Seq technology**. *Mamm Genome* 2010, **21**:592–598.

11. Malavasi A, Zucchi RA: *Mosca-Das-Frutas de Importância Econômica No Brasil: Conhecimento Básico E Aplicado*. Ribeirão Preto: Holos; 2000.

12. Fu L, Li Z-H, Huang G-S, Wu X-X, Ni W-L, Qü W-W: **The current and future potential geographic range of West Indian fruit fly, *Anastrepha obliqua* (Diptera: Tephritidae)**. *Insect Sci* 2013:n/a–n/a.

13. Norrbom AL, Korytkowski CA: *A Revision of the* Anastrepha Robusta *Species Group (Diptera: Tephritidae)*. Auckland, N.Z.: Magnolia Press; 2009.

14. Norrbom AL, Zucchi RA, Hernández-Ortiz V: **Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotripanini) based on morphology**. In *Fruit Flies Tephritidae Phylogeny Evol Behav*. Boca Raton: CRC Press; 2000:944.

15. Smith-Caldas MRB, Mcpheron BA, Silva JG, Zucchi RA: **Phylogenetic Relationships Among Species of the *fraterculus* Group (*Anastrepha*: Diptera: Tephritidae) Inferred from DNA Sequences of Mitochondrial Cytochrome Oxidase I**. *Neotrop Entomol* 2001, **30**:565–573.

16. Zucchi RA: **Espécies de *Anastrepha*, Sinonímias, Plantas Hospedeiras e Parasitóides**. In *Mosca--Frutas Importância Econômica No Bras*. Ribeirão Preto: Holos; 2000:41–48. (vol. 4)

17. Solferini VN, Morgante JS: **Karyotype study of eight species of *Anastrepha* (Diptera: Tephritidae)**. *Caryologia* 1987, **40**:229–241.

18. Malavasi A, Morgante JS: **Genetic Variation in Natural Populatios of *Anastrepha* (Diptera Tephritidae)**. *Rev Bras Genet* 1982, **2**:263–278.

19. Feder JL, Berlocher SH, Opp SB: **Sympatric host race formation and speciation in *Rhagoletis* (Diptera: Tephritidae): a tale of two species for Charles D.** In *Genet Struct Nat Insect Popul Eff Host Plants Life Hist*. New York: Chapman & Hall; 1998:408–441.

20. Aluja M: **Bionomics and management of *Anastrepha***. *Annu Rev Entomol* 1994, **39**:155–178.

21. Ruiz MF, Milano A, Salvemini M, Eirín-López JM, Perondini ALP, Selivon D, Polito C, Saccone G, Sánchez L: **The Gene Transformer of *Anastrepha* Fruit Flies (Diptera, Tephritidae) and Its Evolution in Insects**. *PLoS ONE* 2007, **2**:e1239.

22. Sarno F, Ruiz MF, Eirín-López JM, Perondini AL, Selivon D, Sánchez L: **The gene transformer-2 of *Anastrepha* fruit flies (Diptera, Tephritidae) and its evolution in insects**. *BMC Evol Biol* 2010, **10**:140.

23. Sobrinho IS, de Brito RA: **Evidence for positive selection in the gene fruitless in *Anastrepha* fruit flies**. *BMC Evol Biol* 2010, **10**:293.

24. Sobrinho IS, de Brito RA: **Positive and Purifying Selection Influence the Evolution of Doublesex in the *Anastrepha fraterculus* Species Group**. *PLoS ONE* 2012, **7**:e33446.

25. Gonçalves GB, Silva CE, Dos Santos JC., Dos Santos ES, Do Nascimento RR, Da Silva EL, De Lima Mendonça A, Do Rosário Tenório De Freitas M, Sant'Ana AE.: **Comparison of the volatile components released by calling males of *Ceratitis capitata* (Diptera: Tephritidae) with those extractable from the salivary glands**. *Fla Entomol* 2006, **89**:375–379.

26. Lima IS, House PE, Nascimento RR: **Volatile substances from male *Anastrepha fraterculus* wied.(Diptera: Tephritidae): identification and behavioural activity**. *J Braz Chem Soc* 2001, **12**:196–201.

27. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**:R25.

28. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc* 2013, **8**:1494–1512.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools**. *Bioinforma Oxf Engl* 2009, **25**:2078–2079.

30. Renaut S, Nolte AW, Bernatchez L: **Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae)**. *Mol Ecol* 2010, **19**:115–131.

31. Andres JA, Larson EL, Bogdanowicz SM, Harrison RG: **Patterns of Transcriptome Divergence in the Male Accessory Gland of Two Closely Related Species of Field Crickets**. *Genetics* 2012, **193**:501–513.

32. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792–1797.

33. Abascal F, Zardoya R, Telford MJ: **TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations**. *Nucleic Acids Res* 2010, **38**(Web Server):W7–W13.

34. Posada D: **Using MODELTEST and PAUP* to Select a Model of Nucleotide Substitution**. In *Curr Protoc Bioinforma*. Edited by Baxevanis AD, Davison DB, Page RDM, Petsko GA, Stein LD, Stormo GD. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2003.

35. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J: **KaKs_Calculator: calculating Ka and Ks through model selection and model averaging**. *Genomics Proteomics Bioinformatics* 2006, **4**:259–263.

36. Xu L: **Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms**. *Mol Biol Evol* 2006, **23**:1107–1108.

37. Crawford JE, Guelbeogo WM, Sanou A, Traoré A, Vernick KD, Sagnon N, Lazzaro BP: **De Novo Transcriptome Sequencing in Anopheles funestus Using Illumina RNA-Seq Technology**. *PLoS ONE* 2010, **5**:e14202.

38. Garg R, Patel RK, Tyagi AK, Jain M: **De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification**. *DNA Res* 2011, **18**:53–63.

39. Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F: **De novo Transcriptome Assembly and SNP Discovery in the Wing Polymorphic Salt Marsh Beetle Pogonus chalceus (Coleoptera, Carabidae)**. *PLoS ONE* 2012, **7**:e42605.

40. Xiao M, Zhang Y, Chen X, Lee E-J, Barber CJS, Chakrabarty R, Desgagné-Penix I, Haslam TM, Kim Y-B, Liu E, MacNevin G, Masada-Atsumi S, Reed DW, Stout JM, Zerbe P, Zhang Y, Bohlmann J, Covello PS, De Luca V, Page JE, Ro D-K, Martin VJJ, Facchini PJ, Sensen CW: **Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest**. *J Biotechnol* 2013, **166**:122–134.

41. Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y: **De Novo Sequencing and Transcriptome Analysis of the Central Nervous System of Mollusc Lymnaea stagnalis by Deep RNA Sequencing**. *PLoS ONE* 2012, **7**:e42546.

42. Liang C, Liu X, Yiu S-M, Lim BL: **De novo assembly and characterization of Camelina sativa transcriptome by paired-end sequencing**. *BMC Genomics* 2013, **14**:146.

43. Fernandes F: **Análise multilocus de parâmetros populacionais, evolução molecular e diferenciação em espécies de moscas-das-frutas do grupo *fraterculus* (Diptera, Tephritidae)**. *Tese de Mestrado em Genética e Evolução*. Universidade Federal de São Carlos; 2010.

44. Ng PC, Henikoff S: **Predicting the Effects of Amino Acid Substitutions on Protein Function**. *Annu Rev Genomics Hum Genet* 2006, **7**:61–80.

45. Andolfatto P: **Adaptive evolution of non-coding DNA in Drosophila**. *Nature* 2005, **437**:1149–1152.

46. Suzuki Y, Gojobori T: **A method for detecting positive selection at single amino acid sites.** *Mol Biol Evol* 1999, **16**:1315–1328.

47. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci* 1992, **89**:10915–10919.

48. Dunkov B, Georgieva T: **Insect iron binding proteins: Insights from the genomes**. *Insect Biochem Mol Biol* 2006, **36**:300–309.

49. Nichol H, Law JH, Winzerling JJ: **I RON M ETABOLISM IN I NSECTS**. *Annu Rev Entomol* 2002, **47**:535–559.

50. Mandilaras K, Missirlis F: **Genes for iron metabolism influence circadian rhythms in Drosophila melanogaster**. *Metallomics* 2012, **4**:928.

51. Gelbart WM, Emmert DB: **FlyBase High Throughput Expression Pattern Data**. 2013.

52. Sánchez-Gracia A, Rozas J: **Divergent evolution and molecular adaptation in the Drosophila odorant-binding protein family: inferences from sequence variation at the OS-E and OS-F genes**. *BMC Evol Biol* 2008, **8**:323.

53. Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y: **Odorant-Binding Proteins OBP57d and OBP57e Affect Taste Perception and Host-Plant Preference in Drosophila sechellia**. *PLoS Biol* 2007, **5**:e118.

54. Foret S, Maleszka R: **Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (Apis mellifera)**. *Genome Res* 2006, **16**:1404–1413.

55. McBride CS, Arguello JR: **Five Drosophila Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily**. *Genetics* 2007, **177**:1395–1416.

56. Vieira FG, Sánchez-Gracia A, Rozas J: **Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution**. *Genome Biol* 2007, **8**:R235.

57. Gotzek D, Robertson HM, Wurm Y, Shoemaker D: **Odorant Binding Proteins of the Red Imported Fire Ant, Solenopsis invicta: An Example of the Problems Facing the Analysis of Widely Divergent Proteins**. *PLoS ONE* 2011, **6**:e16289.

58. Wang P, Lyman RF, Shabalina SA, Mackay TFC, Anholt RRH: **Association of Polymorphisms in Odorant-Binding Protein Genes With Variation in Olfactory Response to Benzaldehyde in Drosophila**. *Genetics* 2007, **177**:1655–1665.

59. Devaud J-M: **Experimental studies of adult Drosophila chemosensory behaviour**. *Behav Processes* 2003, **64**:177–196.

60. Wang P, Lyman RF, Mackay TFC, Anholt RRH: **Natural Variation in Odorant Recognition Among Odorant-Binding Proteins in Drosophila melanogaster**. *Genetics* 2009, **184**:759–767.

61. Horst R, Damberger F, Luginbuhl P, Guntert P, Peng G, Nikonova L, Leal WS, Wuthrich K: **NMR structure reveals intramolecular regulation mechanism for pheromone binding and release**. *Proc Natl Acad Sci* 2001, **98**:14374–14379.

62. Leal WS, Chen A, Ishida Y, Chiang V, Erickson M, Morgan T, Tsuruda J: **Kinetics and molecular properties of pheromone binding and release**. *Proc Natl Acad Sci* 2005, **102**:5386–5391.

63. Laughlin JD, Ha TS, Jones DNM, Smith DP: **Activation of Pheromone-Sensitive Neurons Is Mediated by Conformational Activation of Pheromone-Binding Protein**. *Cell* 2008, **133**:1255–1265.

64. Kramps T, Peter O, Brunner E, Nellen D, Froesch B, Chatterjee S, Murone M, Züllig S, Basler K: **Wnt/Wingless Signaling Requires BCL9/Legless-Mediated Recruitment of Pygopus to the Nuclear β-Catenin-TCF Complex**. *Cell* 2002, **109**:47–60.

65. Städeli R, Basler K: **Dissecting nuclear Wingless signalling: Recruitment of the transcriptional co-activator Pygopus by a chain of adaptor proteins**. *Mech Dev* 2005, **122**:1171–1182.

66. Perrimon N, Smouse D, Miklos G: **Developmental genetics of loci at the base of the X chromosome of *Drosophila melanogaster*.** *Genetics* 1989, **121**:313–331.

67. Labate JA, Biermann CH, Eanes WF: **Nucleotide variation at the runt locus in *Drosophila melanogaster* and *Drosophila simulans*.** *Mol Biol Evol* 1999, **16**:724–731.

68. Zurovcova M, Ayala FJ: **Polymorphism patterns in two tightly linked developmental genes, *Idgf1* and *Idgf3*, of *Drosophila melanogaster***. *Genetics* 2002, **162**:177–188.

CAPÍTULO 5 – CONSIDERAÇÕES FINAIS

Neste trabalho a base dos dados foi obtida através da metodologia de sequenciamento de mRNA (RNA-Seq) a partir de 10 perfis de cDNA obtidos de diferentes fases do ciclo reprodutivo do tecido de cabeça de *A. fraterculus* e *A. obliqua* com réplicas. A partir deste conjunto de dados dois manuscritos, a serem publicados em periódicos científicos, foram redigidos e estão descritos nos capítulos acima, no entanto, seus principais resultados e conclusões serão reapresentados e discutidos de forma a sobrepor objetivos comuns aos artigos e consequentemente ao trabalho como um todo.

- No capítulo 2 apresentamos o artigo de caracterização do transcriptoma de cabeça que possui o resultado do sequenciamento Illumina que gerou um total de 155,940,826 reads paired-end distribuídos nas 20 bibliotecas (5 perfis para cada espécie com réplica), com uma média de pouco mais de 8 milhões de reads para *A. fraterculus* e 7,4 milhões de reads para *A. obliqua* (ver tabela S1 do material suplementar na página 40).
- O resultado da filtragem dos reads tem o intuito de eliminar possíveis erros de sequenciamento e homogeneizar a qualidade dos reads que serão utilizados na montagem do transcriptoma. O número total e a porcentagem de reads e bases que permaneceu na análise separados por biblioteca e espécie são apresentados na tabela S2 (página 40).
- Um total de 140,493,653 reads foram utilizados para a montagem do transcriptoma comum das duas espécies pelo software Trinity (Tabela 1, página 27). Outros dois transcriptomas também foram individualmente montados por espécie (Tabela 1, página 47) e todos apresentaram ótimos resultados quando comparados a diversos trabalhos de construção de transcriptomas *de novo* feitos por NGS em plataforma da Illumina. Para esta avaliação consideramos parâmetros como quantidade de contigs, N50, média e o tamanho máximo dos contigs (Anexo A).
- Foram encontrados 1991 genes com expressão gênica diferencial entre *A. fraterculus* e *A. obliqua* em pelo menos uma etapa da vida reprodutiva. Dois conjuntos de genes, um com 123 e outro com 237, parecem separar as espécies ao se analisar todas as bibliotecas em conjunto (Figura 3, página 30). O resultado da expressão gênica diferencial entre as espécies e por etapa da vida reprodutiva está apresentado na Figura 2 (página 29) e mostra a comparação dos perfis de macho virgem com maior número de genes, 1560 e a

comparação entre fêmea pós cópula com o menor número de genes diferencialmente expressos entre as espécies, 25. A diferença observada entre o número de genes com expressão diferencial nas bibliotecas de macho é quase três vezes maior que nas bibliotecas de fêmeas, isto pode ser consequência da seleção positiva afetando de forma diferente os machos de cada espécie. Outro indício que fortalece esta hipótese é a presença em muito maior número de genes ligados a odores nas bibliotecas de machos do que nas de fêmeas.

- No capítulo 3 apresentamos uma análise evolutiva baseada na comparação dos transcriptomas das duas espécies de *Anastrepha*. Primeiramente a busca por SNPs resultou em 62,973 SNPs presentes em 2,773 contigs exclusivos de *A. fraterculus*, 32,616 SNPs em 1,840 contigs exclusivos de *A. obliqua* e 2,822 e 6,698 SNPs em 2,822 contigs comuns as duas espécies. O índice de diferenciação por SNP (D) entre as espécies foi calculado e os 175 contigs que possuíam frequências médias altamente diferenciadas entre as espécies ($\overline{D} > 0.94$) foram separados.
- Os 175 contigs resultantes da busca por SNPs tiveram suas regiões codificadoras (CDS) isoladas e alinhadas entre as espécies e das espécies com o grupo externo, *C. capitata*, e assim foram submetidas a análise evolutiva de Ka/Ks. Se o nosso objetivo principal é identificar genes que podem estar envolvidos no processo de diferenciação entre as espécies de *Anastrepha*, nós selecionamos como desejáveis genes com valor baixo de Ka/Ks das espécies contra *C. capitata* (Ka/Ks < 0.5) e com valores altos entre a *A. fraterculus* e *A. obliqua* (Ka/Ks ≥ 0.5) (Figura 3, página 50). Nestes parâmetros restaram 12 transcritos devidamente anotados e com os valores de Ka/Ks apresentados na tabela 2 (página 49).
- Uma última análise da sequência nucleotídica dos 12 genes candidatos até o momento mostrou que todos os SNPs estão presentes na região codificadora, porém 5 deles apresentam uma substituição sinônima para a mudança causada pelo SNP, o que em teoria, não causaria uma mudança na proteína final. Portanto os 7 os genes candidatos a estarem diretamente envolvidos no processo de diferenciação entre *A. fraterculus* e *A. obliqua* são anotados como: *MTERF domain containing 1*, CG16817, *Rab GTPase domain-containing protein*, *Cytoplasmic linker protein 190* (*CLIP-190*), *Transferrin 3*

(*Tsf3*), *Odorant-binding protein 99a* (*Obp99a*) e *Legless* (*Lgs*), sendo os três últimos mais amplamente discutidos no manuscrito.

Alguns resultados de tópicos comuns aos dois manuscritos serão comentados a seguir.

- O padrão de expressão gênica diferencial encontrado nos grupos de genes de *A. obliqua* e de *A. fraterculus* que parecem diferenciar as espécies presente no capítulo 3 foi contrastado com os 175 SNPs com $\overline{D} > 0.94$ mostrado no capítulo 4. Três dos 123 genes com o padrão de super expressão gênica em *A. obliqua* foram encontrados no resultado dos SNPs e outros três dos 237 genes super expressos em *A. fraterculus* possuem SNPs que distinguem as espécies (Anexo B). Nenhum destes seis genes está no resultado dos 12 selecionados após a análise de KaKs. Uma análise mais profunda desses 6 genes, com corroboração da expressão gênica em qPCR e uma pesquisa mais detalhada de suas sequências e funções pode incluí-los também na lista de genes candidatos a estarem envolvidos no processo de diferenciação dessas espécies. Eles foram anotados como sendo *CG42240*, *CG32425*, *LamC* "*Lamin C*", *Arl4* "ADP ribosylation factor-like 4", *Pgi* "*Phosphoglucose isomerase*" e um não apresentou resultado no blast como GO.
- A proporção de contigs que se encontra no resultado de expressão gênica diferencial (Manuscrito 1) e possuem SNPs (Manuscrito 2) é semelhante quando se observa as análises de cada biblioteca de estádio reprodutivo (Anexo C). Em torno de 14% a 18% dos contigs que possuem expressão gênica diferencial possuem ao menos um SNP, independentemente da quantidade de genes diferencialmente expressos entre as bibliotecas. Por exemplo, a biblioteca de macho virgem apresentou 1,560 genes diferencialmente expressos e desses 212 (14%) apresentaram SNPs, enquanto na biblioteca de fêmea pós cópula observamos 4 genes no resultado de expressão com SNPs do total de 24 (17%). É sugerido para trabalhos futuros um aprofundamento no resultado de expressão gênica diferencial de cada dos contrastes de bibliotecas.
- A expressão gênica diferencial observada da comparação das bibliotecas de fêmea pós oviposição foi a que apresentou o resultado mais discrepante, quando se analisa a proporção de genes diferencialmente expressos entre as espécies e também quanto aos SNPs (Anexo CC). Em *A. obliqua* foi observado um maior número de genes super expressos e SNPs em relação a *A. fraterculus*, porém a não existência de padrões

semelhantes nas outras 4 comparações de bibliotecas sugere que não haja um erro metodológico que tenha provocado tal desvio, nos levando a acreditar em algum fator biológico que ainda não foi encontrado. Portanto, a continuação e aprofundamento das pesquisas nestas bibliotecas de pós oviposição podem gerar resultados ainda mais interessantes, principalmente ligado a comportamento e preferências para oviposição nestas espécies.

- Ao vasculhar a expressão gênica para os 12 genes finais da análise evolutiva observa-se apenas 2 deles com expressão diferencial entre as espécies. O contig anotado como *Serino Protease 6* (*Ser6*) apresenta-se 3,5 vezes mais expresso em *A. obliqua* do que em *A. fraterculus* nas bibliotecas de fêmea virgem e o contig anotado como *Transferrina 3* (*Tsf3*), que possui uma substituição não sinônima radical, apresenta-se com uma expressão diferencial aproximadamente 10 vezes maior em *A. obliqua* do que em *A. fraterculus* na comparação das bibliotecas de macho virgem.
- Com estes dados, o transcrito identificado como "comp120945_c0_seq1" e anotado como *Transferrina 3* foi o que apresentou maior número de indícios de estar envolvido no processo de diferenciação entre *A. fraterculus* e *A. obliqua*. Outros 6 genes presentes no resultado principal da análise evolutiva também são considerados fortes candidatos a estarem relacionados com o processo de separação destas espécies, entre eles encontramos uma *OBP*, família gênica também bastante presente no resultado de expressão diferencial e que já está sendo profundamente investigada por uma doutoranda. O transcrito da *Ser6* apesar de possuir uma substituição sinônima ele faz parte dos 12 genes finais da análise evolutiva e também se encontra presente em resultados de expressão, o que o também deixa como candidato a estar envolvido no processo de diferenciação entre *A. fraterculus* e *A. obliqua*.
- Todos os resultados provenientes da análise de expressão gênica precisam ser corroborados por qPCR. Esta etapa já se iniciou no laboratório de Genética de Populações e Evolução e é fundamental para a publicação do manuscrito apresentado no capítulo 3.
- Os dados de sequenciamento de tecido cefálico das 20 bibliotecas adquiridas, processadas e apresentadas neste trabalho se encontram disponíveis no Lab. de Genética de Populações e Evolução e continuam a ser analisados de forma a gerar outras publicações.

REFERÊNCIAS

ALUJA, M. Bionomics and management of *Anastrepha*. *Annual Review of Entomology*, v. 39, n. 1, p.155–178, 1994.

ALUJA, M. ; NORRBOM, A.L. *Fruit flies (Tephritidae):* phylogeny and evolution of behavior. Boca Raton: CRC Press, 2000.

ARAUJO, E.L. ; ZUCCHI, R.A. Medidas do acúleo na caracterização de cinco espécies de *Anastrepha* do grupo fraterculus (Diptera: Tephritidae). *Neotropical Entomology*, v. 35, n. 3, p.29–337, 2006.

BAXTER, S.W. et al. Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism P. K. Ingvarsson, ed. *PLoS ONE*, v. 6, n. 4, p.e19315. 2011.

VAN BELLEGHEM, S.M. et al. De novo Transcriptome Assembly and SNP Discovery in the Wing Polymorphic Salt Marsh Beetle Pogonus chalceus (Coleoptera, Carabidae) Z. Liu, ed. *PLoS ONE*, v. 7, n. 8, p.e42605, 2012.

BERLOCHER, S.H. Radiation and divergence in the *Rhagoletis pomonella* Species Group: inferences from allozymes. *Evolution*, v. 54 n. 2, p.543–557, 2000.

BOKULICH, N.A. et al. Next-Generation Sequencing reveals significant bacterial diversity of botrytized wine M. Horn, ed. *PLoS ONE*, v. 7 n. 5, p.e36357, 2012.

BRAS, J. ; GUERREIRO, R. ; HARDY, J. Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease. *Nature Reviews Neuroscience*, v. 13 n. 7, p.453–464, 2012.

BUSH, G.L. Sympatric speciation in phytophagous parasitic insects. In P. W. Price, ed. *Evolutionary Strategies of Parasitic Insects and Mites*. Boston, MA: Springer US, p.187–206, 1975.

BUTLIN, R. ; RITCHIE, M.G. Evolutionary biology: searching for speciation genes. *Nature*, v. 412, n. 6842, p.31–33, 2001.

CARVALHO, M.C.C.G. ; SILVA, D.C.G. Next generation DNA sequencing and its applications in plant genomics. *Ciência Rural*, v. 40, n. 3, p.735–744, 2010.

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

CHEN, S. et al. De Novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits J. C. Zheng, ed. *PLoS ONE*, 5(12), p.e15633, 2010.

CLARIDGE-CHANG, A. et al. Circadian regulation of gene expression systems in the *Drosophila* head. *Neuron*, v. 32, n. 4, p.657–671, 2001.

CLARKE, A.R. et al. invasive phytophagous pests arising through a recent tropical evolutionary radiation: The *Bactrocera dorsalis* Complex of Fruit Flies. *Annual Review of Entomology*, v. 50 n. 1, p.293–319, 2005.

CONDON, M. et al. Uncovering tropical diversity: six sympatric cryptic species of *Blepharoneura* (Diptera: Tephritidae) in flowers of *Gurania spinulosa* (Cucurbitaceae) in eastern Ecuador: Six Sympatric Cryptic Species On A Single Host. *Biological Journal of the Linnean Society*, v. 93, n. 4, p.779–797, 2008.

CORTNER, J. ; WOUDE, G.F.V.. Essencials of molecular biology - cDNA libraries. In *Cancer: Principles and practice of oncology*. Philadelphia: Lippincott - Raven Publishers, p.545, 1997.

COYNE, J.A. *Speciation*, Sunderland, Mass: Sinauer Associates, 2004.

CRAWFORD, J.E. et al. De Novo transcriptome sequencing in *Anopheles funestus* using illumina RNA-Seq technology P. Awadalla, ed. *PLoS ONE*, v. 5, n. 12, p.e14202, 2010.

DAVEY, J.W. : BLAXTER, M.L. RADSeq: next-generation population genetics. *Briefings in functional genomics*, v. 9, n. 5-6, p.416–423, 2010.

DELAY, B. et al. Transcriptome analysis of the salivary glands of potato leafhopper, *Empoasca fabae*. *Journal of Insect Physiology*, v. 58, n. 12, p.1626–1634, 2012.

BEN-DOR, A., SHAMIR, R. ; YAKHINI, Z. Clustering Gene Expression Patterns. *Journal of Computational Biology*, v. 6, n. 3-4, p.281–297, 1999.

DUARTE, A.L. ; MALAVASI, A. Tratamentos Quarentenários. In *Mosca-das-Frutas de Importância Econômica no Brasil: Conhecimento Básico e Aplicado*. Ribeirão Preto: Holos, p.187–192, 2000.

EMERSON, K.J. et al. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, v. 107, n. 37, p.16196–16200, 2010.

FEDER, J.L., BERLOCHER, S.H. ; OPP, S.B. Sympatric host race formation and speciation in *Rhagoletis* (Diptera: Tephritidae): a tale of two species for Charles D. In *Genetic Structure in Natural Insect Populations: Effects of Host Plants and Life History*. New York: Chapman & Hall, p.408–441, 1998.

FEDER, J.L., CHILCOTE, C.A. ; BUSH, G.L. Genetic differentiation between sympatric host races of the apple maggot fly *Rhagoletis pomonella*. *Nature*, v. 336, n. 6194, p.61–64, 1988.

FERNANDES, F. *Análise multilocus de parâmetros populacionais, evolução molecular e diferenciação em espécies de moscas-das-frutas do grupo* fraterculus *(Diptera, Tephritidae)*. Tese de Mestrado em Genética e Evolução. Departamento de Genética e Evolução: Universidade Federal de São Carlos, 2010.

FUJII, S. ; AMREIN, H. Genes expressed in the *Drosophila* head reveal a role for fat cells in sex-specific physiology. *The EMBO journal*, v. 21, n. 20, p.5353–5363, 2002.

GALINDO, K. ; SMITH, D.P. A large family of divergent *Drosophila* odorant-binding proteins expressed in gustatory and olfactory sensilla. *Genetics*, v. 159, n. 3, p.1059–1072, 2001.

GOMPERT, Z. et al. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Molecular Ecology*, p.no–no, 2010.

GONÇALVES, G.B. et al. Comparison of the volatile components released by calling males of *Ceratitis capitata* (Diptera: Tephritidae) with those extractable from the salivary glands. *Florida Entomologist*, v. 89, n. 3, p.375–379, 2006.

GREENSPAN, R.J. ; FERVEUR, J.F. Courtship In *Drosophila*. *Annual Review of Genetics*, v. 34, n. 1, p.205–232, 2000.

HARISMENDY, O. et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, v. 10, n. 3, p.32, 2009.

HEATH, R.R. et al. Sexual pheromones of tephritid flies: clues to unravel phylogeny and behavior. In *Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior*. Boca Raton: CRC Press, p.793–809, 2000.

HENNING, F. ; MATIOLI, S.R. Mating time of the West Indian fruit fly *Anastrepha obliqua* (Macquart)(Diptera: Tephritidae) under laboratory conditions. *Neotropical Entomology*, v. 35, n. 1, p.145–148, 2006.

HOLSINGER, K.E. Next generation population genetics and phylogeography. *Molecular ecology*, v. 19, n. 12, p.2361–2363, 2010.

HUGHES, M.E. et al. Deep sequencing the circadian and diurnal transcriptome of *Drosophila* brain. *Genome Research*, v. 22, n. 7, p.1266–1281, 2012.

JACKSTADT, R. ; MENSSEN, A. ; HERMEKING, H. Genome-Wide Analysis of c-MYC-Regulated mRNAs and miRNAs, and c-MYC DNA Binding by Next-Generation Sequencing. In L. Soucek & N. M. Sodir, eds. *The Myc Gene*. Totowa, NJ: Humana Press, p.145–185, 2013.

JAGADEESHAN, S. Rapidly Evolving Genes of *Drosophila*: Differing Levels of Selective Pressure in Testis, Ovary, and Head Tissues Between Sibling Species. *Molecular Biology and Evolution*, v. 22, n. 9, p.1793–1801, 2005.

JEYASANKAR, A. Chemical ecology of fruit fly management. *J. Basic Appl. Biol*, v. 3, n. 1-2, p.1–5, 2009.

KOBOLDT, D.C. et al. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell*, v. 155, n. 1, p.27–38, 2013.

KU, C.-S. ; ROUKOS, D.H. From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. *Expert Review of Medical Devices*, v. 10, n. 1, p.1–6, 2013.

LIMA, I. ; HOWSE, P. ; STEVENS, I. Volatile components from the salivary glands of calling males of the south American fruit fly, *Anastrepha fraterculus*: partial identification and behavioural activity, 1996.

LIMA, I.S. ; HOUSE, P.E. ; NASCIMENTO, R.R. Volatile substances from male *Anastrepha fraterculus* wied.(Diptera: Tephritidae): identification and behavioural activity. *Journal of the Brazilian Chemical Society*, v. 12, n. 2, p.196–201, 2001.

LIOTTA, L. ; PETRICOIN, E. Molecular profiling of human cancer. *Nature Reviews Genetics*, v. 1, n. 1, p.48–56, 2000.

MALAVASI, A. ; MORGANTE, J.S. Genetic Variation in Natural Populatios of *Anastrepha* (Diptera Tephritidae). *Rev. Brasil. Genet.*, v. 2, p.263–278, 1982.

MALAVASI, A. ; ZUCCHI, R.A. *Mosca-das-Frutas de Importância Econômica no Brasil: Conhecimento Básico e Aplicado*, Ribeirão Preto: Holos, 2000.

MARIONI, J.C. et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, v. 18, n. 9, p.1509–1517, 2008.

MATSUMURA, H. et al. High-Throughput SuperSAGE for Digital Gene Expression Analysis of Multiple Samples Using Next Generation Sequencing J. Bähler, ed. *PLoS ONE*, v. 5, n. 8, p.e12010, 2010.

MAY, R.M. How many species are there on earth?. *Science(Washington)*, v. 241, n. 4872, p.1441–1449, 1988.

MCDONALD, M.J. ; ROSBASH, M. Microarray Analysis and Organization of Circadian Gene Expression in *Drosophila*. *Cell*, v. 107, n. 5, p.567–578, 2001.

MITCHELL, R.F. et al. Sequencing and characterizing odorant receptors of the cerambycid beetle *Megacyllene caryae*. *Insect Biochemistry and Molecular Biology*, v. 42, n. 7, p.499–505, 2012.

MORGANTE, J. ; MALAVASI, A. ; BUSH, G. Biochemical systematics and evolutionary relationships of neotropical *Anastrepha*. *Annals of the Entomological Society of America*, v. 73, n. 6, p.622–630, 1980.

NEVINS, J.R. ; POTTI, A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Reviews Genetics*, v. 8, n. 8, p.601–609, 2007.

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

NORRBOM, A.L. ; KORYTKOWSKI, C.A. New species of and taxonomic notes on *Anastrepha* (Diptera: Tephritidae). *Zootaxa*, v. 2740, p.1–23, 2011.

NORRBOM, A.L. ; ZUCCHI, R.A. ; HERNÁNDEZ-ORTIZ, V. Phylogeny of the genera *Anastrepha* and *Toxotrypana* (Trypetinae: Toxotripanini) based on morphology. In *Fruit Flies (Tephritidae): Phylogeny and Evolution of Behavior*. Boca Raton: CRC Press, p. 944, 2000.

NUNES, M.Z. et al. Avaliacão de atrativos alimentares na captura de *Anastrepha fraterculus* (Wiedemann, 1830)(Diptera: Tephritidae) em pomar de macieira. *Revista de la Facultad de Agronomía, La Plata*, v. 112, n. 2, p.91–96, 2013.

PAVEY, S.A. et al. The role of gene expression in ecological speciation: Gene expression and speciation. *Annals of the New York Academy of Sciences*, v. 1206, n. 1, p.110–129, 2010.

PLAUTZ, J.D. Independent Photoreceptive Circadian Clocks Throughout *Drosophila*. *Science*, v. 278, n. 5343, p.1632–1635, 1997.

RIESGO, A. et al. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Frontiers in Zoology*, v. 9, n. 1, p.33, 2012.

ROBACKER, D.C. ; HEATH, R.R. Decreased attraction of *Anastrepha ludens* to combinations of two types of synthetic lures in a citrus orchard. *Journal of chemical ecology*, v. 23, n. 5, p.1253–1262, 1997.

RUIZ, M.F. et al. The Gene Transformer of *Anastrepha* Fruit Flies (Diptera, Tephritidae) and Its Evolution in Insects J.-N. Volff, ed. *PLoS ONE*, v. 2, n. 11, p.e1239, 2007.

SADAMOTO, H. et al. De Novo Sequencing and Transcriptome Analysis of the Central Nervous System of Mollusc Lymnaea stagnalis by Deep RNA Sequencing M. Sakakibara, ed. *PLoS ONE*, v. 7, n. 8, p.e42546, 2012.

DOS SANTOS, P. ; URAMOTO, K. ; MATIOLI, S.R. Experimental Hybridization Among *Anastrepha*; Species (Diptera: Tephritidae): Production and Morphological Characterization of F1 Hybrids. *Annals of the Entomological Society of America*, v. 94, n. 5, p.717–725, 2001.

SARNO, F. et al. The gene transformer-2 of *Anastrepha* fruit flies (Diptera, Tephritidae) and its evolution in insects. *BMC Evolutionary Biology*, v. 10, n. 1, p.140, 2010.

SELIVON, D. Biologia e padrões de especiação. In *Mosca-das-Frutas de Importância Econômica no Brasil: Conhecimento Básico e Aplicado*. Ribeirão Preto: Holos, p.25–28, 2000.

SELIVON, D. ; PERONDINI, A.L.P. ; MORGANTE, J.S. A Genetic–Morphological Characterization of Two Cryptic Species of the *Anastrepha fraterculus* Complex (Diptera: Tephritidae). *Annals of the Entomological Society of America*, v. 98, n. 3, p.367–381, 2005.

SHENDURE, J. The beginning of the end for microarrays? *Nature Methods*, v. 5, n. 7, p.585–587, 2008.

SMITH-CALDAS, M.R.B. et al. Phylogenetic Relationships Among Species of the *fraterculus* Group (*Anastrepha*: Diptera: Tephritidae) Inferred from DNA Sequences of Mitochondrial Cytochrome Oxidase I. *Neotropical Entomology*, v. 30, n. 4, p.565–573, 2001.

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

SOBRINHO, I.S. ; BRITO, R.A. Positive and Purifying Selection Influence the Evolution of Doublesex in the *Anastrepha fraterculus* Species Group W. J. Etges, ed. *PLoS ONE*, v. 7, n. 3, p.e33446, 2012.

SOLFERINI, V.N. ; MORGANTE, J.S. Karyotype study of eight species of *Anastrepha* (Diptera: Tephritidae). *Caryologia*, v. 40, n. 3, p.229–241, 1987.

TEMPLETON, A.R. The Meaning of Species and Speciation: A Genetic Perspective. In *Speciation and its Consequences*. Sunderland, MA: Sinauer, p.3–27, 1989.

URAMOTO, K. ; WALDER, J.M. ; ZUCCHI, R.A. Biodiversidade de moscas-das-frutas do gênero *Anastrepha* (Diptera, Tephritidae) no campus da ESALQ-USP, Piracicaba, São Paulo. *Revista Brasileira de Entomologia*, v. 48, n. 3, p.409–414, 2004.

VIJAY, N. et al. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology*, v. 22, n. 3, p.620–634, 2013.

VIRGILIO, M. et al. Molecular evaluation of nominal species in the *Ceratitis fasciventris*, *C. anonae*, *C. rosa* complex (Diptera: Tephritidae). *Molecular Phylogenetics and Evolution*, v. 48, n. 1, p.270–280, 2008.

WANG, Z. ; GERSTEIN, M. ; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, v. 10, n. 1, p.57–63, 2009.

WHITEHEAD, A. ; CRAWFORD, D.L. Variation within and among species in gene expression: raw material for evolution: Review of gene expression variation. *Molecular Ecology*, v. 15, n. 5, p.1197–1211, 2006.

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

XIAO, M. et al. Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *Journal of Biotechnology*, v. 166, n. 3, p.122–134, 2013.

ZENG, Z.B. et al. Genetic Architecture of a Morphological Shape Difference Between Two Drosophila Species. *Genetics*, v. 154, n. 1, p.299 –310, 2000.

ZHU, J.Y. ; ZHAO, N. ; YANG, B. Global Transcriptional Analysis of Olfactory Genes in the Head of Pine Shoot Beetle, *Tomicus yunnanensis*. *Comparative and Functional Genomics*, v. 2012, p.1–10, 2012.

ZUCCHI, R.A. Espécies de *Anastrepha*, Sinonímias, Plantas Hospedeiras e Parasitóides. In *Mosca-das-Frutas de Importância Econômica no Brasil*. Ribeirão Preto: Holos, p.41–48, 2000.

ANEXOS

Anexo A. Comparação dos assemblies feitos pelo Trinity com o de outros Transcriptomas *de novo* feitos por RNA-seq da Illumina

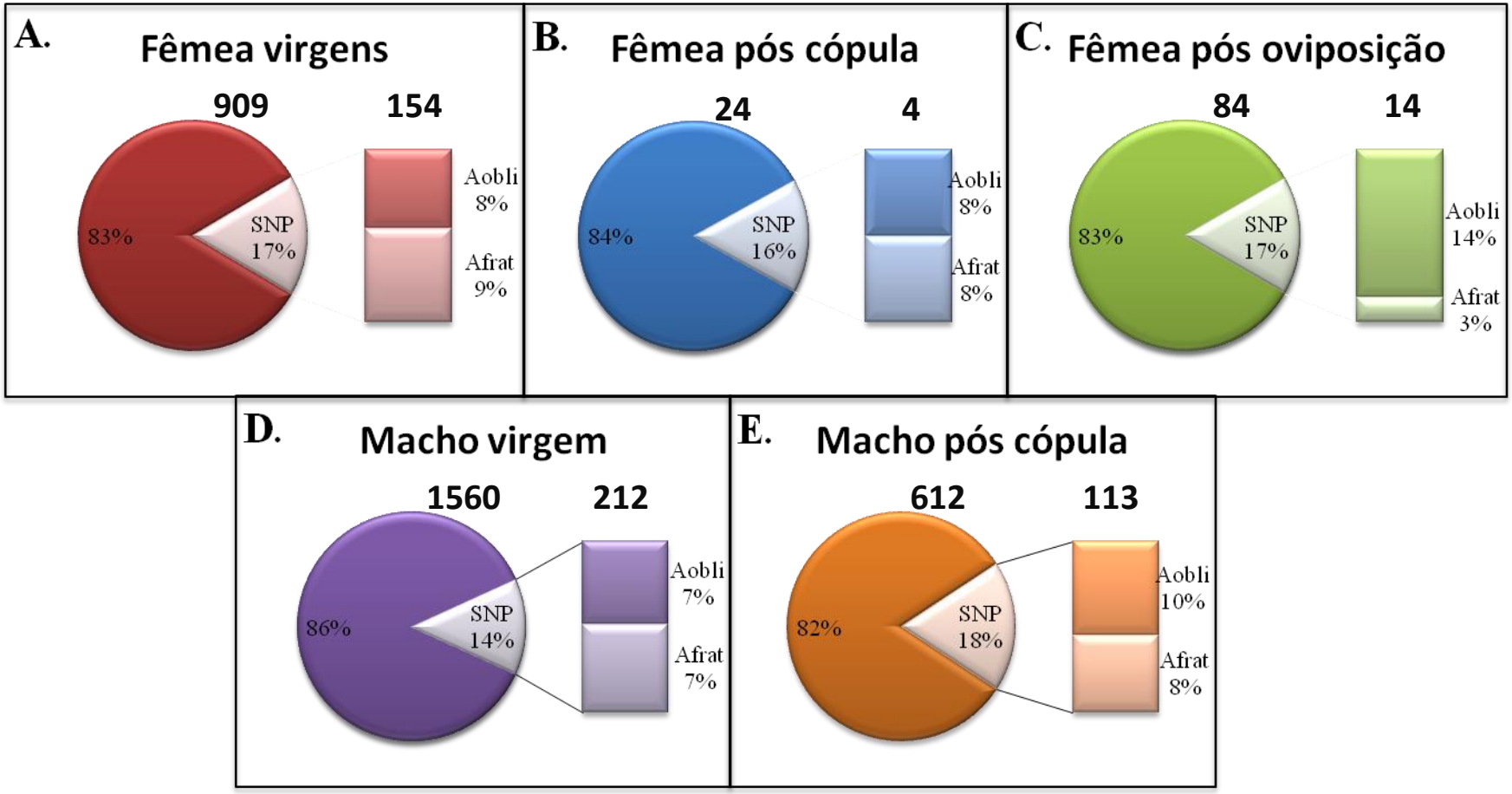
Parametros	**Rezende et al., 2014**			(Chen et al. 2010)	(Crawford et al. 2010)	(Van Belleghem et al. 2012)	(Sadamoto et al. 2012)
Espécie	*A. fraterculus*	*A. obliqua*	*A. frat + A. obli*	*P. chalceus*	*A. funestus*	*L. migratoria*	*L. stagnalis*
Software de Assembly	Trinity	Trinity	Trinity	Trinity	Velvet	SOAPdenovo**	Rnnotator
Total de reads	73,637,239	66,856,414	140,493,653	184,749,261	96,960,000*	376,000,000*	81,851,004
Número total de contigs	112,862	98,549	154,787	65,766	46,987	72,977	116,355
N50	2,504	2,637	2,012	1,904	1,140	2,275	1,438
Média (bp)	1,196	1,254	1,027	1,044	-	1,171	656
Maior contig (bp)	27,513	22,394	25,704	19,606	-	27,681	26,147

* = aproximadamente; ** = Pós assembly ainda houve uma montagem extra com CAP3; - = Informação não disponível; *L. migratoria* = *Locusta migratoria*; *A. funestus* = *Anopheles funestus*; *P. chalceus* = *Pogonus chalceus*; *L. stagnalis* = *Lymnaea stagnalis*.

Anexo B. Contigs anotados que apresentaram padrão de super expressão para alguma das espécies e sub expressão para a outra e também possuem SNPs com $\overline{D}$ > 0.94.

Espécie c/ super expressão	Contig	Anotação Gene Ontology	ID Flybase
A. fraterculus	comp110414_c0_seq1	Lamin C "LamC"	FBgn0010397
	comp115577_c0_seq1	CG42240	FBgn0250869
	comp120360_c0_seq1	-	-
A. obliqua	comp112125_c0_seq1	Phosphoglucose isomerase "Pgi"	FBgn0003074
	comp112327_c0_seq1	ADP ribosylation factor-like 4 "Arl4"	FBgn0039889
	comp112945_c0_seq1	CG32425	FBgn0052425

ANEXOS



Anexo C. Proporção de genes que, conjuntamente, possuem SNPs e são diferencialmente expressos. Os gráficos mostram todos os genes com expressão diferencial entre as espécies para cada perfil (número acima do círculo). A fatia menor são genes com SNPs, divididos por genes super expressos em *A. obliqua* ou *A. fraterculus* (número acima do retângulo).