Natalia de Souza Araujo

Expressão de genes envolvidos no comportamento social em abelhas que apresentam diferentes níveis de eussocialidade

Expression of genes involved in the social behaviour of bees with different levels of eusociality

São Paulo

2017

Natalia de Souza Araujo

Expressão de genes envolvidos no comportamento social em abelhas que apresentam diferentes níveis de eussocialidade

Expression of genes involved in the social behaviour of bees with different levels of eusociality

Tese apresentada ao Instituto de Biociências da Universidade de São Paulo, para a obtenção de Título de Doutor em Ciências, na Área de Genética e Biologia Evolutiva.

Orientador(a): Maria Cristina Arias
Co-orientador(a): Tatiana Teixeira Torres

São Paulo

2017

Araujo, Natalia de Souza
Expressão de genes envolvidos no comportamento social em abelhas que apresentam diferentes níveis de eussocialidade / Natalia de Souza Araujo; orientadora Maria Cristina Arias. -- São Paulo, 2017.

116 f.

Tese (Doutorado) - Instituto de Biociências da Universidade de São Paulo, Departamento de Genética e Biologia Evolutiva.

1. Eussocialidade. 2. Transcriptomica. 3. Next generation sequencing . I. Arias, Maria Cristina, orient. II. Título.

Catalogação da Publicação
Serviço de Biblioteca do Instituto de Biociências

Comissão Julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

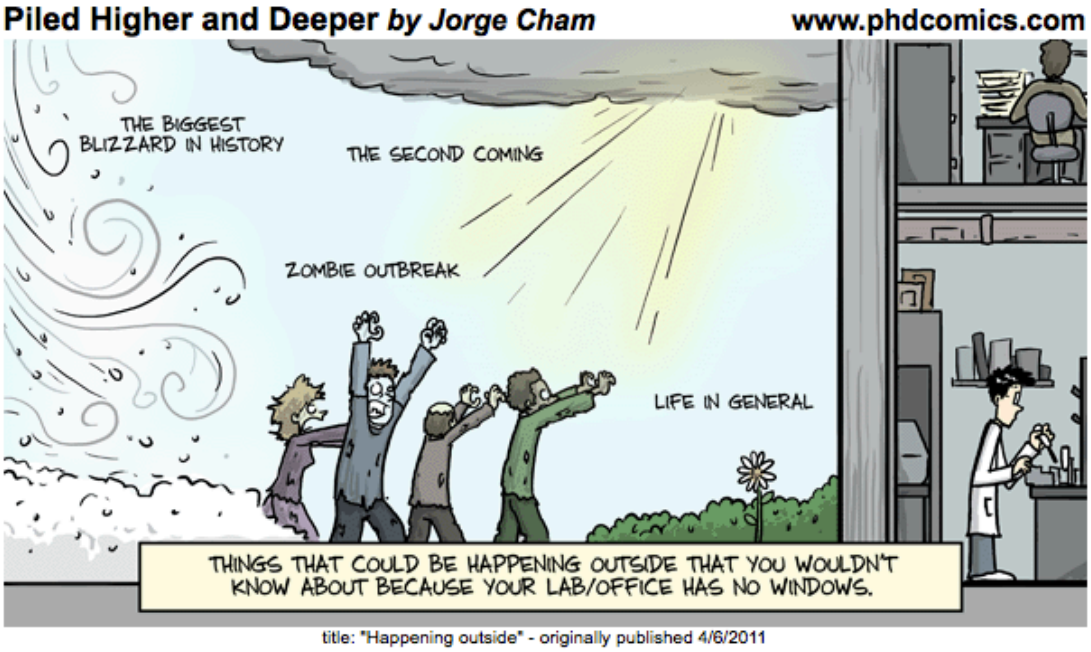
Profa. Dra. Maria Cristina Arias
Orientadora

Dedicatória

À minha família,

que fez tudo isso valer muito mais a pena

Epígrafe



title: "Happening outside" - originally published 4/6/2011

"Natura nusquam magis est tota quam in minimis"

"Em parte alguma encontramos a natureza na sua totalidade como nas sua menores criaturas"

Plinio – naturalista que morreu ao estudar a erupção do Versúvio, em 79 d. C.

Agradecimentos

Essa tese, e os trabalhos aqui apresentados, não teriam sido possíveis sem o auxílio de um grande número de pessoas, as quais sou imensamente grata. A começar por ela, minha grande e querida orientadora, que foi "*The best* advisor in the world*" e aceitou trabalhar comigo nesta arriscada empreitada. Mais do que isso, ela sempre esteve por perto pra me apoiar, arrumar dinheiro, estudar comigo e se irritar comigo. Podia escrever pelo menos uma página inteira sobre como ela foi importante nesse período, mas vou me limitar a um parágrafo. Muito obrigada mesmo, Maria Cristina Arias!

Claro, que por trás de uma grande orientadora, sempre tem um grupo maior ainda, e eu não poderia deixar de agradecer a todas as pessoas que fazem e fizeram parte do LGEA enquanto eu estive aqui. Mas principalmente, meu muito obrigada aos amigos Alexandre (anexo do LGEA), Elaine, Paulo, Priscila e Vanessa (também anexa). Sério, não sei como este trabalho seria possível sem vocês. Vocês ouviram, discutiram, apoiaram e xingaram todas as minhas ideias científicas nos últimos anos e com isso contribuíram muito para a finalização deste doutorado. E isso tudo além das contribuições diretas: me ensinando a usar a formatação automática do Word e o itálico do Mendley (Alexandre); lendo exaustivamente a minha tese até o ultimo segundo (Priscila); rindo da minha cara e dizendo "Você sabe que você não vai conseguir fazer tudo isso né?!" toda vez que eu dizia que queria incluir uma nova análise ou um novo dado no trabalho (Elaine); e deixando eu atrapalhar sua concentração, arruinando seu dia de trabalho, só porque eu queria conversar (Paulo). Vocês são tão queridos que é difícil de acreditar que são reais! E Susy, não me esqueci de você! Muito obrigada por todo o apoio durante esses anos, as conversas, fofocas, chás, cafés, etc, etc... Você é uma pisciana muito gente boa que faz toda diferença no nosso dia a dia do lab.

O segundo laboratório mais legal dessa Universidade, certamente, é o Laboratório de Abelhas. Isso porque as abelhas de lá são o máximo... Brincadeira! Isso é verdade também, mas o que tornou esse lugar especial para mim foram as pessoas nele e o quanto elas sempre estiveram dispostas a me ajudar! Em especial a Dra Isabel Alves dos Santos, que me abriu as portas desse lugar mágico e sempre esteve disponível pra conversar sobre as abelhas e compartilhar todo o seu

conhecimento; o Guara, que não só me forneceu informações privilegiadas sobre a *Tetrapedia* mas teve paciência de me mostrar como lidar com elas na prática; a Priscila, que deixou eu coletar do seu único ninho de *Euglossa*; e a Sheina. Ah a Sheina... como agradecer esse amor de pessoa que me ajudou a coletar mais da metade dos dados e ainda fingia que eu não estava atrapalhando/ incomodando?!

Aliás, isso foi algo que todas as pessoas que me ajudaram nas coletas fizeram. Muito obrigada Dra Denise Alves, Dra Solange Augusto, Larissa e Eiko por me ajudarem nessa etapa. Principalmente à De e à Sol que compartilharam sua abelhas especiais comigo de muito bom grado e ainda me ensinaram muito sobre elas e sobre o que é fazer parte desse grupo incrível de pesquisadores de abelhas no Brasil. Outra pessoa desse grupo de pesquisadores que também foi bastante importante para este trabalho, e talvez nem saiba disso, foi o Dr Klaus Hartfelder. Em todos os congressos que estive, ou sempre que surgia a oportunidade, ele me presenteou com seu conhecimento sobre o tema e levantou aspectos importantes que foram sempre considerados no meu trabalho.

Durante este estudo também recebi ajuda de pesquisadores de outras áreas, como o Luciano e o Lucas (do lab Rita), que dividiram materiais de laboratório e artigos sobre pipeline de análises. O pessoal do laboratório do prof Dr Luis Netto, especialmente o Thiago Alegria, pra quem eu corria sempre que precisava de algum reagente ou equipamento. A prof Dra Tatiana Teixeira Torres, sem a qual esse projeto provavelmente não teria começado, já que as contribuições dela no primeiro delineamento foram muito importantes. Mas não pesquisadores também contribuíram muito para este trabalho de diversas forma. O pessoal do Estúdio Multimeios da STI (antigo CCE) e o Ricardo Nonaka do IB, por exemplo, me ajudaram muito com a parte técnica de informática, especialmente com o serviço de Cloud USP. Meliponicultores também foram importantes para este estudo, como o Gerson Pinheiro do SOS Abelha sem Ferrão e o Antonio. A todo esse pessoal, obrigada!

Contudo, os não pesquisadores que mais colaboram com os trabalhos apresentados aqui foram minha filha Julia e meu marido Ricardo. Ambos me ajudaram nas coletas e em todas as vezes que eu parava pra observar uma abelha. Mas acima de tudo, sempre me deram todo o suporte para que eu pudesse fazer esse trabalho, tanto emocional/ psicológico quanto prático. Sem vocês ao meu lado durante todos esses anos eu, sinceramente, eu não conseguiria.

I also had important collaborators from other countries to whom I would like to thank. Dr Yannick Wurm, welcomed me in his lab for my sandwich PhD during one year, and during this period he contributed significantly to my career. With meaningful discussions and new perspectives, he and his group improved the studies reported here. The Wurm lab group were imperative for this PhD thesis. Thank you so much guys!!

Still, for the time I spent in London, I must thank Dr. Lars Chittka and Dr. Stephan Wolf, from Bee Sensory and Behavioural Ecology Lab (Queen Mary University of London) and Dr. Andres Arce and Dr. Richard Gill (Imperial College London – Silwood Park), for all support during bumblebee sampling. They gave me more than their bees to work with, they shared their knowledge with me too. Additionally, I would like to thank Bob Schmitz (University of Georgia) who accepted to collaborate with a random researcher from a completely different field after a conference. He opened my mind to the importance of DNA methylation.

E finalmente, eu gostaria de agradecer imensamente a Fundação de Amparo à Pesquisa do Estado de São Paulo, a FAPESP, que foi a grande financiadora deste projeto e viabilizou os estudos apresentados aqui, não só custeando a pesquisa (processo No 2013/12530-4), mas também as minhas bolsas de doutorado no país (processo No 2012/18531-0) e no exterior (processo No 2014/04943). E ao Centro de Biodiversidade e Computação (BioComp) da Universidade de São Paulo, que também auxiliou financeiramente em parte das análises.

Índice

Resumo

O comportamento social pode ser descrito como qualquer atividade de interação intraespecífica incluindo a escolha entre parceiros reprodutivos, reconhecimento da espécie, comportamento altruísta e organização da sociedade animal. Entre as espécies de animais mais sintonizadas com seu ambiente social estão os insetos que, como por exemplo nas espécies de abelhas das tribos Apini e Meliponini, apresentam um padrão complexo de socialidade conhecido como comportamento altamente eussocial. As abelhas constituem um grupo ideal para o estudo das bases da evolução deste comportamento, pois apresentam uma grande diversidade de organização social, desde espécies solitárias até altamente eussociais. Embora a evolução da eussocialidade tenha sido motivo de muitos estudos, as mudanças genéticas envolvidas nesse processo não são completamente conhecidas. Dados da literatura fornecem um ponto de partida para o entendimento da relação entre alterações gênicas específicas e a eussocialidade, mas questões fundamentais na evolução do comportamento social ainda precisam ser respondidas. Recentemente, novas tecnologias de sequenciamento têm permitido o estudo de organismos modelo e não modelo de forma mais detalhada e não direcional. Análises deste tipo são promissoras para o estudo evolutivo de características complexas como o comportamento. Neste contexto, realizamos um amplo estudo sobre as bases moleculares envolvidas em diferentes características comportamentais relacionadas à evolução da socialidade em abelhas. Para tanto, o padrão global de expressão de genes, em espécies e fases do desenvolvimento distintas, foram analisados comparativamente através de múltiplas abordagens. No Capítulo 1, utilizamos contaminantes do transcriptoma da abelha solitária *Tetrapedia diversipes* para analisar os recursos florais utilizados por esta espécie em suas duas gerações reprodutivas. Neste estudo concluímos que a riqueza de espécies visitadas durante a primeira geração é muito maior do que durante a segunda geração, o que está provavelmente relacionado à floração de primavera durante o primeiro período reprodutivo. No Capítulo 2, verificamos que o padrão de expressão dos genes das fêmeas fundadoras possivelmente afeta o desenvolvimento larval em *T. diversipes*. O padrão bivoltino de reprodução desta espécie, com diapausa em uma das gerações, pode ser importante para a evolução do comportamento social.

Além disso, entre os genes possivelmente envolvidos nessa característica, podemos encontrar genes mitocondriais e lncRNAs. Os resultados obtidos no Capítulo 3 sugerem que a especialização em subcastas de operárias ocorreu posteriormente nas diferentes linhagens de abelhas, envolvendo genes específicos. No entanto, esses genes afetam processos biológicos comuns nas diferentes espécies. Por sua vez, o Capítulo 4 apresenta um método promissor para a identificação de genes comportamentais em diferentes espécies de abelhas, através de uma análise de expressão comparativa. Com base nessas análises, 787 genes comportamentais, que possivelmente fazem parte de um *toolkit* eussocial em abelhas, foram encontrados. O padrão de metilação desses genes, em espécies com diferentes níveis sociais, indicou ainda que o contexto genômico da metilação pode ser relevante para eussocialidade. Os resultados obtidos nesses estudos apresentam novas perspectivas metodológicas e evolutivas para o estudo da evolução do comportamento social em abelhas.

Abstract

The social behaviour can be widely described as any intraspecific interaction in the animal life, including but not restricted to, female choice, species recognition, altruistic behaviour and the organization of animal society. Among the animal species most attuned to their social environment are the insects that, for example, in the Apini and Meliponini tribes, present a complex behaviour known as highly eusocial. Bees are an ideal group to study the evolution of the social behaviour because they have a great diversity of social life styles that evolved independently. The tribes Apini and Meliponini comprise only highly eusocial species whereas various levels of sociality can be detected in other tribes, being most bees indeed solitary. Although the evolution of eusociality has been the subject of many studies, the genetic changes involved in the process have not been completely understood. Results from studies conducted so far provide a starting point for the connection between specific genetic alterations and the evolution of eusocial behaviour. However fundamental questions about this process are still open. Recently, new sequencing technologies have allowed genetic studies of model and non-model organisms in a deep and non-directional way, which is promising for the study of complex characteristics. Herein, we present a broad analysis of the molecular bases of different behavioural characteristics related to the evolution of sociality in bees. To that end, the global expression pattern of genes involved in different behavioural features, in a number of bee species and distinct developmental stages, was comparatively studied using multiple approaches.

Through these approaches different results were obtained. In Chapter 1, we used contaminant transcripts from the solitary bee *Tetrapedia diversipes* to identify the plants visited by this bee, during its two reproductive generations. These contaminant transcripts revealed that the richness of plant species visited during the first reproductive generation was considerably greater than during the second generation. Which is probably related to the floral boom occurring in spring during the first reproductive period. In Chapter 2, data suggests that the expression pattern in foundresses affect larval development in *T. diversipes*. The bivoltinism presented by this species, with diapause in one generation, might be an important feature for the evolution of sociality. Our results suggest that mitochondrial genes and lncRNAs are involved in this reproductive pattern. Results described in Chapter 3 indicate that

specialization in worker subcastes occurred posteriorly in distinct bee lineages, driven by specific genes. However, these genes affected common biological processes in the different species. In Chapter 4 is described a promising analyses method to identify, comparatively, genes involved in bee social behaviour. Using this approach, we identified 787 behavioural genes that might be involved in social behaviour of different species. The methylation pattern of these genes suggests that the DNA context in which methylation marks occur, might be especially relevant to bee sociality. Results obtained here presents new methodological and evolutionary approaches to the study of social behaviour in bees.

Introdução Geral

Atividades de interações intraespecíficas como a escolha da fêmea, reconhecimento da espécie, comportamento altruísta e organização da sociedade animal, são denominadas de comportamento social (ROBINSON et al., 1997). Existem espécies que interagem com coespecíficos somente no momento da cópula e espécies que vivem altamente estruturadas em sociedades, com vida social complexa, nas quais praticamente todas as atividades são influenciadas pelas interações entre os indivíduos (ROBINSON & BEN-SHAHAR, 2002). Este tipo de socialidade é altamente derivada e evoluiu de forma independente em diferentes linhagens de animais (WILSON, 2000), gerando sistemas sociais diversos controlados por múltiplos genes (ROBINSON & BEN-SHAHAR, 2002; TOTH et al., 2007; BERENS et al., 2014; TOTH & REHAN, 2017).

Entre as espécies de animais mais sintonizadas com seu ambiente social estão os insetos, nos quais o comportamento social surgiu independentemente ao menos 12 vezes (FISCHMAN et al., 2011). Apesar desta convergência evolutiva, inúmeras diferenças nos níveis de complexidade social são conhecidas (ROBINSON et al., 1997). Há desde espécies que formam pequenas colônias (com conflitos evidentes em relação à reprodução) a espécies nas quais as colônias apresentam centenas de milhares de operárias estéreis altamente especializadas, geradas por uma ou poucas rainhas (WILSON, 1971; WOODARD et al., 2011). Dentre os insetos, as abelhas destacam-se como modelo para o estudo das bases da evolução do comportamento social, pois apresentam uma grande diversidade de estilos de vida social (de solitárias a altamente eussociais) em um único clado (TOTH & REHAN, 2017). Adicionalmente, a eussocialidade parece ter surgido de forma independente em diferentes tribos de abelhas. Essa peculiaridade torna possível comparar as múltiplas e independentes origens das distintas organizações sociais no grupo (WOODARD et al., 2011).

As espécies altamente eussociais (tribos Apini e Meliponini) apresentam grandes colônias, com ciclo colonial perene e castas de rainha e operária altamente especializadas (MICHENER, 2007; GRÜTER et al., 2017). Em contraste, algumas espécies, como a maioria das abelhas na tribo Bombini, são denominadas primitivamente eussociais. Estas, formam pequenas colônias, com ciclo colonial anual

e castas menos especializadas (THOMPSON & OLDROYD, 2004; WOODARD et al., 2014). A maioria das espécies de abelhas, no entanto, não é social e as fêmeas são solitárias (BATRA, 1984). Nas espécies solitárias, uma única fêmea realiza todas as tarefas como postura de ovos, construção de ninhos e forrageamento (ALVES-DOS-SANTOS et al., 2002; THOMPSON & OLDROYD, 2004; WOODARD et al., 2011). A **Tabela I** apresenta com mais detalhes dados sobre o comportamento e a biologia das diferentes abelhas de especial interesse no presente trabalho.

Tabela I Detalhes sobre as principais características das espécies de abelha de interesse neste trabalho.

Bombus terrestris (Linnaeus, 1758)	Popularmente conhecida como mamangava da cauda amarela, esta abelha é nativa de regiões temperadas da Europa (RASMONT et al., 2008). No entanto, devido à sua grande capacidade adaptativa e ao seu uso como inseto polinizador em plantações comerciais, atualmente sua distribuição é quase cosmopolita (RASMONT et al., 2008; SARAIVA et al., 2012). Nas áreas invadidas, *B. terrestris* contribui para declínio de espécies de abelhas nativas (SARAIVA et al., 2012). Esta espécie tem ciclo de vida anual e forma colônias **primitivamente eussociais** (RIBEIRO, 1997; GOULSON, 2010).
Euglossa annectans Dressier, 1982	Espécie pertencente à tribo Euglossini, única tribo de abelhas corbiculadas cujas espécies não são eussociais (MICHENER, 2007). Abelhas desse gênero constroem seus ninhos geralmente em células de resina aglomerada em cavidades pré-existentes, e a maioria das espécies é solitária (SILVEIRA et al., 2002). Algumas espécies podem dividir cooperativamente o mesmo ninho, como é o caso de *E. annectans*, em que as fêmeas podem reutilizar e compartilhar ninhos já existentes, mas cada uma aprovisiona sua própria célula independentemente (GAROFALO et al., 1998). O que a torna uma espécie **comunal** (MICHENER, 1969). Sua ocorrência abrange o Brasil, Argentina e Paraguai (MOURE et al., 2012).
Friesella schrottkyi (Friese, 1900)	Única espécie descrita no gênero *Friesella* da tribo Meliponini de abelhas **altamente eussociais** (SILVEIRA et al., 2002). Produzem colônias de cerca de 300 operárias construídas em cavidades de madeira pré-existentes (IMPERATRIZ-FONSECA & KLEINERT, 1998). Estas abelhas fecham a entrada do ninho durante a noite e constroem favos de cria irregulares sem invólucro protetor, também há a construção de células de cria para a produção de rainhas e câmeras de aprisionamento para rainhas virgens (NUNES et al., 2010). As operárias de *F. schrottkyi*, diferentemente de muitos meliponíneos, não produzem ovos tróficos, mas durante um curto período de tempo podem produzir ovos reprodutivos (IMPERATRIZ-FONSECA & KLEINERT, 1998).
Frieseomelitta varia (Lepeletier, 1836)	Espécie **altamente eussocial** da tribo Meliponini que pode ser encontra em algumas regiões do Brasil e na Bolívia (MOURE et al., 2012). Os ninhos são formados em cavidades ocas pré-existentes onde as células de cria são construídas em formato de cacho (MICHENER, 2007). Segundo Michener (2007), é provável que na maioria dos meliponíneos células de cria em formato de cacho seja um comportamento ancestral. No entanto, em *F. varia*, essa característica parece ser derivada, o que é corroborado pelo mecanismo alternativo de criação de rainhas, que utiliza células de operárias comuns suplementadas com células auxiliares de alimentação (FAUSTINO et al., 2002). Outra característica marcante desta espécie é a ausência de ovários desenvolvidos nas operárias, nem mesmo para a produção de ovos tróficos (BOLELI et al., 2000).
Melipona bicolor Lepeletier, 1836	O gênero *Melipona* é o mais distinto da tribo Meliponini. Nas espécies desse gênero a produção de rainhas é constante, sem a construção de células reprodutivas maiores (MICHENER, 2007). *M. bicolor*, por sua vez, é a única espécie de abelha conhecida que apresenta poliginia (presença de múltiplas rainhas) permanente e facultativa (CEPEDA, 2006). Esta espécie, **altamente eussocial,** é endêmica de regiões de Mata Atlântica, e está presente em fragmentos remanescentes de mata desde o Estado da Bahia até o Rio Grande do Sul (SILVEIRA et al., 2002).
Nannotrigona testaceicornis Lepeletier, 1836	Espécie **altamente eussocial** encontrada no Brasil, Argentina e Paraguai (MOURE et al., 2012) é representante de um gênero pequeno com maior diversidade na área da bacia amazônica (CARDOSO & SILVEIRA, 2012). São abelhas mansas que constroem ninhos em ocos de árvores, mas não ocupam todo o

	espaço, que portanto, pode ser habitado por outras espécies (NOGUEIRA-NETO, 1997). As colônias podem ter de 2.000 a 3.000 abelhas e os favos de cria são construídos em formato helicoidal (LINDAUER & KERR, 1960). Existem células reprodutivas para a produção de rainhas no ninho, no entanto, operárias também podem se tornar rainhas anãs (IMPERATRIZ-FONSECA et al., 1997)
Neocorynura sp.	Este é um dos maiores gêneros da tribo Augochlorini (Halictidae), com aproximadamente 60 a 65 espécies de abelhas descritas (SMITH-PARDO, 2005), que se distribuem do norte Argentino à região central do México (MOURE et al., 2012). Abelhas da subfamília Halictinae são importantes no estudo do comportamento social pois exibem uma grande variabilidade de comportamentos, e a eussocialidade evoluiu recentemente neste grupo (BRADY et al., 2006). O gênero *Neocorynura* é grupo irmão dos gêneros sociais *Augochlorella* e *Augochlora* (BRADY et al., 2006). No entanto, abelhas do gênero *Neocorynura* são geralmente **solitárias** com algumas raras espécies que apresentam comportamento pré-social (SMITH-PARDO, 2005).
Plebeia remota (Holmberg, 1903)	O gênero **altamente eussocial** *Plebeia* é especialmente importante para o estudo da socialidade por ser considerado um dos grupos com morfologia e comportamento mais primitivos de Meliponini (VAN BENTHEM et al., 1995). Esta espécie Neotropical (MOURE et al., 2012) nidifica em diferentes cavidades onde suas células de cria são orientadas horizontalmente (MICHENER, 2007). A produção dos favos é sincronizada, cerca de 50 células de crias são construídas de uma única vez, e o invólucro do ninho é restrito a períodos de inverno, quando também ocorre diapausa reprodutiva da rainha (VAN BENTHEM et al., 1995).
Scaptotrigona aff. depilis (Moure, 1942)	As espécies do gênero *Scaptotrigona* são abelhas **altamente eussociais** bastante robustas (MICHENER, 2007). Distribuem-se por toda a região Neotropical e apresentam uma grande diversidade de formas, sendo portanto, um grupo que apresenta muitas incertezas taxonômicas (SILVEIRA et al., 2002; MOURE et al., 2012). Operárias desta espécie podem participar da produção de machos na colônia de forma bastante frequente e constante ao longo do ano (PAXTON et al., 2003).
Tetragonisca angustula (Latreille, 1811)	As espécies do gênero *Tetragonisca* são caracterizadas pela reduzida corbícula (MICHENER, 2007). Apenas três espécies são descritas nesse gênero, incluindo *T. angustula*, popularmente conhecida como Jataí, espécie do gênero com maior área de ocorrência no Brasil (SILVEIRA et al., 2002; MOURE et al., 2012). Esta abelha **altamente eussocial** é bem adaptada à vida urbana devido a sua plasticidade na escolha dos locais de nidificação (CORTOPASSI-LAURINO & NOGUEIRA-NETO, 2003). A Jataí também foi a primeira abelha sem ferrão em que subcastas morfológicas de operárias foram descritas (WITTMANN, 1985; GRÜTER et al., 2012, 2017).
Tetrapedia diversipes Klug, 1810	São abelhas coletoras de óleo que ocorrem na região Neotropical (MOURE et al., 2012). A tribo Tetrapedini inclui apenas dois gêneros, *Tetrapedia* e *Coelioxoides*, sendo este último parasita dos ninhos de *Tetrapedia* (MICHENER, 2007). Na cidade de São Paulo, *T. diversipes* apresenta duas principais gerações reprodutivas durante o ano, uma no período de seca e outra no período chuvoso (ALVES-DOS-SANTOS et al., 2002). O tempo de desenvolvimento dos indivíduos em cada geração reprodutiva pode variar de 21 a 366 dias (CORDEIRO, 2009), pois as larvas da época de seca entram em diapausa (CAMILLO, 2005; ALVES-DOS-SANTOS et al., 2006). Essa espécie **solitária** pode ser facilmente coletada em ninhos armadilhas (CAMILLO, 2005; ALVES-DOS-SANTOS et al., 2006; MENEZES et al., 2012; NEVES et al., 2012; ROCHA-FILHO & GARÓFALO, 2015).
Xylocopa frontalis Olivier, 1789	*X. frontalis* é uma abelha grande e robusta da tribo Xilocopini, que nidifica em madeira durante todo o ano (SILVEIRA et al., 2002). Os ninhos possuem poucos indivíduos e nenhuma rainha. No entanto, as fêmeas fundadoras permanecem no ninho alimentando a prole com néctar por cerca de 30 dias após a emergência. Depois deste período, as filhas podem permanecer no ninho de origem, geralmente expulsando a mãe, mantendo certo tipo de interação social entre fêmeas de uma mesma geração (CAMILLO & GAROFALO, 1989; SILVA et al., 2014). Sendo assim, esta espécie, que pode ser encontrada na região Neotropical e no México (MOURE et al., 2012; SILVA et al., 2014), apresenta um comportamento conhecido como **subsocial** (MICHENER, 2007).

Essa diversidade comportamental pode ser encontrada em diferentes espécies de uma única subfamília de abelhas, como por exemplo, a subfamília Apinae.

Segundo MICHENER (2007), a subfamília Apinae compreende dezenove tribos das quais dezesseis estão presentes na região Neotropical. Destas, quinze tribos são nativas e apenas uma é introduzida (Apini - representada pela espécie *Apis mellifera*). Quatro tribos de Apinae (Apini, Bombini, Euglossini e Meliponini) compreendem as chamadas "Apidae corbiculadas", pois possuem uma estrutura especializada, a corbícula, no último par de pernas que é utilizada para o transporte de pólen e resinas (MICHENER, 2007). Apini e Meliponini são as únicas tribos, dentro da subfamília, que apresentam o grau de comportamento social mais avançado, o comportamento altamente eussocial (MICHENER, 2007). Estudos filogenéticos, em especial com base em dados moleculares, sugerem que o comportamento eussocial pode ter evoluído independente duas vezes na subfamília Apinae, uma origem na tribo Apini e a outra nas Meliponini e Bombini (WOODARD et al., 2011) (**Figura 1**). No entanto, este assunto é bastante discutido na literatura e estudos recentes sugerem que a eussocialidade teve uma única origem evolutiva nas corbiculadas (ROMIGUIER et al., 2015; BOSSERT et al., 2017).

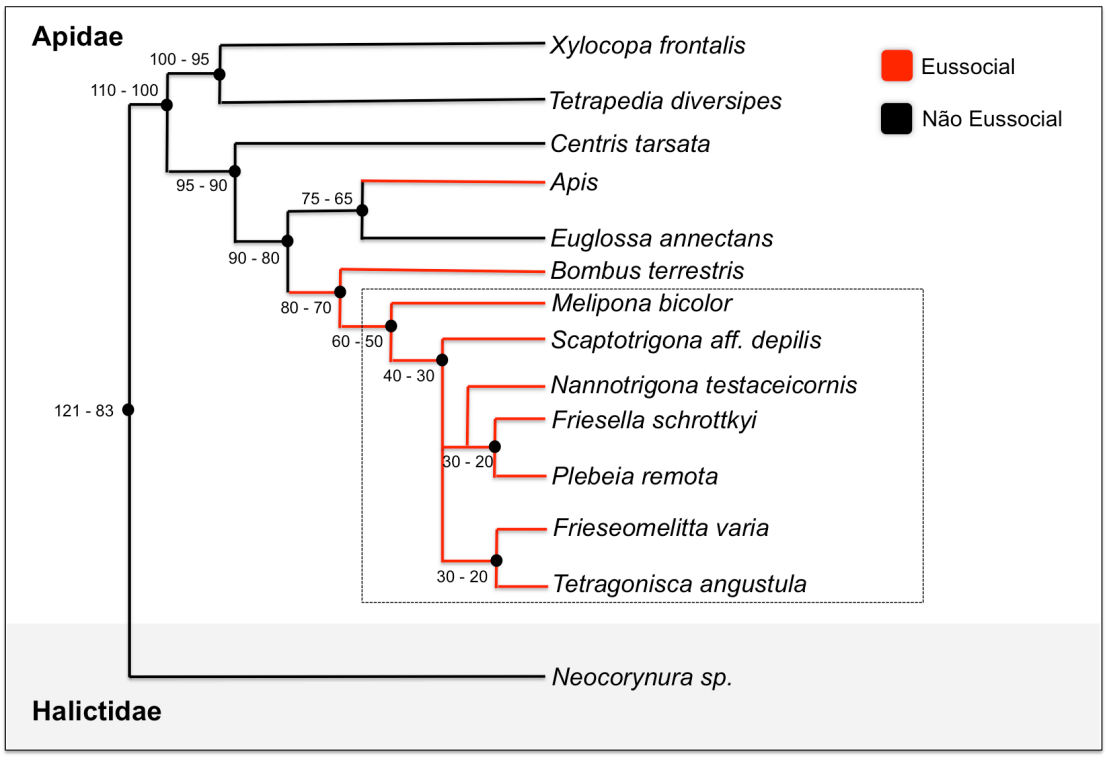


Figura 1. Cladograma das espécies de interesse neste trabalho ilustrando a possível dupla origem da eussocialidade na família Apidae. No quadrado pontilhado encontram-se as espécies da tribo Meliponini. Datação próxima ao nó de cada clado em milhões de anos. Figura e datação baseadas em filogenias previamente publicadas por outros autores (BRADY et al., 2009; CARDINAL et al., 2010; HEDTKE et al., 2013; GRÜTER et al., 2017).

Em abelhas eussociais, as castas de rainha e operária estão diretamente relacionadas à função intracolonial desses indivíduos, sendo a rainha a responsável pela reprodução, enquanto as operárias atuam em tarefas relacionadas ao crescimento e manutenção da colônia (ROBINSON et al., 1997; GROZINGER et al., 2007). A especialização de indivíduos em castas é um exemplo clássico de polifenismo, onde o mesmo genoma pode gerar múltiplos fenótipos (morfológicos e comportamentais) (GROZINGER et al., 2007).

A presença de operárias que não se reproduzem e cuidam de seus irmãos é característico da eussocialidade e um grande desafio na teoria evolutiva (TOTH et al., 2007; NOWAK et al., 2010). Por isso, algumas das principais hipóteses sobre a evolução do comportamento social destacam a divisão de castas como um fator essencial no surgimento da eussocialidade (REHAN & TOTH, 2015; TOTH & REHAN, 2017). A hipótese do "plano de fundo ovariano" (*ovarian ground-plan hypothesis*) estabelece que o comportamento social evoluiu pela dissociação do comportamento reprodutivo e de forrageamento (REHAN & TOTH, 2015). Assim, a interação entre genes reprodutivos e de forrageio deu origem a indivíduos com perfis "mais reprodutivos" e outros "mais operários", o que posteriormente originou as castas de rainha e operária respectivamente (AMDAM et al., 2006). Já na hipótese de "heterocronia materna" (*maternal heterochrony hypothesis*), a divisão reprodutiva do trabalho ocorreu pela dissociação temporal do comportamento reprodutivo e de cuidado materno apresentado pelas fêmeas, sem necessariamente haver a separação de vias regulatórias de forrageamento e reprodução (LINKSVAYER & WADE, 2005; HUNT, 2012). Esta hipótese fundamenta-se especialmente em resultados de expressão gênica, que indicam que a expressão de genes relacionados ao comportamento de cuidado materno é similar à expressão de genes envolvidos no cuidado de imaturos realizado pelas operárias (TOTH et al., 2007; WOODARD et al., 2014).

Uma hipótese mais geral sobre a evolução do comportamento social e que, portanto, pode englobar as ideias anteriores é a hipótese do *toolkit* gênico (ROBINSON et al., 1997, 2008). Essa hipótese, surgiu com base em ideais eco-evolutivas de que mudanças regulatórias em genes com papeis conservados entre as espécies seriam importantes para a evolução de novos fenótipos (REHAN & TOTH, 2015). Assim, em insetos sociais, um grupo de genes específicos estaria envolvido na

evolução do comportamento social em diferentes espécies, formando um *toolkit* gênico social (ROBINSON & BEN-SHAHAR, 2002; ROBINSON et al., 2008; REHAN & TOTH, 2015). Baseado nessa hipótese, inúmeros estudos sobre o envolvimento de genes específicos como: *foraging*, *malvolio*, *period*, *ace*, *royal jelly protein*, *protein kinase C*, *vitellogenin*, *telomerase*, *chico*, *tor*, *juvenile hormone* e *octopamine* (SCHULZ & ROBINSON, 2001; ROBINSON et al., 2005; GROZINGER et al., 2007) e análises em grande escala genômica (CHANDRASEKARAN et al., 2011; WOODARD et al., 2011; FISCHMAN et al., 2011), foram realizados para entender os mecanismos moleculares envolvidos no comportamento social em abelhas. No entanto, a maioria dos trabalhos publicados focou essencialmente no comportamento das espécies altamente sociais do gênero *Apis*, com poucas exceções (KAPHEIM, 2016; TOTH & REHAN, 2017).

Um dos trabalhos realizados com várias espécies de abelhas foi o de WOODARD e colaboradores (2011). Neste estudo, os autores analisaram taxas de evolução sinônima e não sinônima em genes de nove espécies de abelhas, com níveis de comportamento social diversos. Após as análises comparativas de seleção, os autores encontraram genes com altas taxas evolutivas compartilhados entre espécies com o mesmo comportamento, e genes especificamente sob seleção em uma única linhagem. Estes dados sugerem que a convergência evolutiva deste comportamento pode ser decorrente de um padrão mosaico de mudanças compartilhadas e específicas para cada linhagem. Ao comparar dados de expressão entre abelhas, formigas e vespas, BERENS et al. (2014) também verificaram certa sobreposição entre o padrão de expressão de genes nas diferentes espécies, mas apenas quando consideradas as funções desses genes e não os genes propriamente. Assim, com o aumento de dados genômicos de larga escala para espécies não modelo, a ideia de *toolkit* gênico foi ampliada para envolver um *toolkit* de redes gênicas que estariam envolvidas em certas funções metabólicas conservadas (REHAN & TOTH, 2015).

Recentemente, com a diminuição dos custos de sequenciamento em larga escala, outro fator foi adicionado às discussões sobre a evolução do comportamento social: a importância dos mecanismos de regulação da expressão gênica (KAPHEIM, 2016). Estes dados deram origem à hipótese da "regulação conservada" (*conserved regulation hypothesis*) (REHAN & TOTH, 2015; KAPHEIM, 2016). Segundo esta, mudanças nos mecanismos de regulação seriam essenciais para o surgimento da

eussocialidade, e alterações na composição nucleotídica dos genomas estariam envolvidas apenas em adaptações linhagem-específicas posteriores (KAPHEIM, 2016). A importância dos mecanismos de regulação da expressão para o comportamento social em abelhas foi demonstrada primeiramente em 2015, após a conclusão do sequenciamento de 10 genomas completos de abelhas (KAPHEIM et al., 2015). Com a análise comparativa desses genomas os autores não encontraram similaridades claras entre sequências de espécies com comportamentos similares. Em contrapartida, descobriram evidências de que espécies sociais apresentam um controle de expressão (incluindo a metilação) maior do que espécies solitárias. Ainda em 2015, SØVIK et al. demonstraram também que outros mecanismos regulatórios como microRNAs, podem ter contribuído para a evolução de características importantes para eussocialidade.

Logo, apesar de bastante estudado, ainda existem diversas lacunas sobre os mecanismos e genes envolvidos na evolução molecular do comportamento social (SØVIK et al., 2015; KAPHEIM, 2016; TOTH & REHAN, 2017). O que torna imperativo o desenvolvimento de pesquisas que utilizem novas ferramentas de sequenciamento na análise das diversas características relacionadas à socialidade, em diferentes espécies. Além disso, genes que controlam características complexas, como comportamento, apresentam variações dinâmicas e os níveis de expressão desses genes podem ser mais informativos do que análises de mutações em suas sequências (SIMOLA et al., 2013; ZHANG et al., 2014). Especialmente porque ao analisar a expressão dos genes podemos verificar os efeitos combinados dos diferentes mecanismos de controle de expressão e das alterações nucloetídicas. Neste contexto, o presente trabalho busca expandir o conhecimento sobre a evolução molecular do comportamento social através da análise do perfil global da expressão de genes em diferentes espécies de abelhas e características comportamentais.

Referências

ALVES-DOS-SANTOS, I.; MELO, G. A. R.; ROZEN JR, J. G. Biology and Immature Stages of the Bee Tribe Tetrapediini (Hymenoptera : Apidae). **American Museum of Natural History**, v. 3377, p. 1–45, 2002.

ALVES-DOS-SANTOS, I.; NAXARA, S. R. C.; PATRÍCIO, E. F. L. R. A. Notes on the Morphology of *Tetrapedia diversipes* KLUG 1810 (Tetrapedini, Apidae), an oil-collecting bee. **Braz. J. Morphol. Sci.**, v. 23, n. 3–4, p. 425–430, 2006.

AMDAM, G. V et al. Complex social behaviour derived from maternal reproductive traits. **Nature**, v. 439, n. 7072, p. 76–78, 2006.

BATRA, S. W. T. Solitary Bees. **Scientific American**, v. 250, n. 2, p. 120–127, fev. 1984.

BERENS, A. J.; HUNT, J. H.; TOTH, A. L. Comparative transcriptomics of convergent evolution: Different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. **Molecular Biology and Evolution**, p. 1–14, 2014.

BOLELI, I. C.; PAULINO-SIMÕES, Z. L.; BITONDI, M. M. G. Regression of the lateral oviducts during the larval-adult transformation of the reproductive system of *Melipona quadrifasciata* and *Frieseomelitta varia*. **Journal of Morphology**, v. 243, n. 2, p. 141–151, 2000.

BOSSERT, S. et al. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. **Molecular Phylogenetics and Evolution**, v. 111, p. 149–157, 2017.

BRADY, S. G. et al. Recent and Simultaneous Origins of Eusociality in Halictid Bees. **Proceedings of the Royal Society B**, v. 273, p. 1643–1649, 7 jul. 2006.

BRADY, S. G.; LARKIN, L.; DANFORTH, B. N. Bees, ants, and stinging wasps (Aculeata). **The Timetree of Life**, p. 264–269, 2009.

CAMILLO, E. Nesting biology of four *Tetrapedia* species in trap nests (Hymenoptera: Apidae: Tetrapediini). **Revista de Biologia Tropical**, v. 53, p. 175–186, 2005.

CAMILLO, E.; GAROFALO, C. A. Social Organization in reactivated nests of three species of *Xylocopa* (Hymenoptera, Anthophoridae) in southeastern Brasil. **Insectes Sociaux**, v. 36, n. 2, p. 92–105, 1989.

CARDINAL, S.; STRAKA, J.; DANFORTH, B. N. Comprehensive phylogeny of apid bees reveals the evolutionary origins and antiquity of cleptoparasitism. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 37, p. 16207–11, 2010.

CARDOSO, C. F.; SILVEIRA, F. A. Nesting biology of two species of *Megachile* (*Moureapis*) (Hymenoptera: Megachilidae) in a semideciduous forest reserve in southeastern Brazil. **Apidologie**, v. 43, p. 71–81, 2012.

CEPEDA, O. I. Division of labor during brood production in stingless bees with special reference to individual participation. **Apidologie**, v. 37, p. 175–190, 2006.

CHANDRASEKARAN, S. et al. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 44, p. 18020–18025, 2011.

CORDEIRO, G. D. **Abelhas solitárias nidificantes em ninhos-armadilha em quatro áreas de Mata Atlântica do Estado de São Paulo**. [s.l.] Universidade de São Paulo, 2009.

CORTOPASSI-LAURINO, M.; NOGUEIRA-NETO, P. Notas sobre a bionomia de *Tetragonisca weyrauchi* Schwarz (Apidae, Meliponini). **Acta Amazonica**, v. 33, n. 4, p. 643–650, 2003.

FAUSTINO, C. D. et al. First record of emergency queen rearing in stingless bees (Hymenoptera, Apinae, Meliponini). **Insectes soc.**, v. 49, p. 111–113, 2002.

FISCHMAN, B. J.; WOODARD, S. H.; ROBINSON, G. E. Molecular evolutionary analyses of insect societies. **Proceedings of the National Academy of Sciences of the United States of America**, p. 10847–54, 28 jun. 2011.

GAROFALO, C. A. et al. Nest Structure and Communal Nesting in Euglossa (Glossura) annectans dressler (Hymenoptera, Apidae, Euglossini). **Revta bras. Zool. 15**, v. 15, n. 3, p. 589–596, 1998.

GOULSON, D. **Bumblebees : behaviour, ecology, and conservation**. [s.l.] Oxford University Press, 2010.

GROZINGER, C. M. et al. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (Apis mellifera). **Molecular Ecology**, v. 16, n. 22, p. 4837–4848, 2007.

GRÜTER, C. et al. A morphologically specialized soldier caste improves colony defense in a neotropical eusocial bee. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 4, p. 1182118–6, 2012.

GRÜTER, C. et al. Repeated evolution of soldier sub-castes suggests parasitism drives social

complexity in stingless bees. **Nature Communications**, v. 8, n. 1, p. e4, 2017.
HEDTKE, S. M.; PATINY, S.; DANFORTH, B. N. The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. **BMC Evolutionary Biology**, v. 13, p. 138, 2013.
HUNT, J. H. A conceptual model for the origin of worker behaviour and adaptation of eusociality. **Journal of Evolutionary Biology**, v. 25, n. 1, p. 1–19, jan. 2012.
IMPERATRIZ-FONSECA, V.; CRUZ-LANDIM, C.; MORAES, R. S. DE. Dwarf gynes in *Nannotrigona testaceicornis* (Apidae, Meliponinae, Trigonini). Behaviour, exocrine gland morphology and reproductive status. **Apidologie**, v. 28, p. 113–122, 1997.
IMPERATRIZ-FONSECA, V. L.; KLEINERT, A. M. P. Worker Reproduction in the Stingless Bee Species *Friesella schrottkyi* (Hymenoptera: Apidae: Meliponinae). **Entomologia Generalis**, v. 23, n. 3, p. 169–175, 1 out. 1998.
KAPHEIM, K. M. et al. Genomic signatures of evolutionary transitions from solitary to group living. **Science**, v. 348, n. 6239, p. 1139–1143, 2015.
KAPHEIM, K. M. Genomic sources of phenotypic novelty in the evolution of eusociality in insects. **Current Opinion in Insect Science**, v. 13, p. 24–32, 2016.
LINDAUER, M.; KERR, W. E. Communication between the Workers of Stingless Bees. **Bee World**, v. 41, n. 3, p. 65–71, 31 mar. 1960.
LINKSVAYER, T. A; WADE, M. J. The evolutionary origin and elaboration of sociality in the aculeate Hymenoptera: maternal effects, sib-social effects, and heterochrony. **The Quarterly review of biology**, v. 80, n. 3, p. 317–336, 2005.
MENEZES, G. B. et al. Nesting and use of pollen resources by *Tetrapedia diversipes* Klug (Apidae) in Atlantic Forest areas (Rio de Janeiro, Brazil) in different stages of regeneration. **Revista Brasileira de Entomologia**, v. 56, p. 86–94, 2012.
MICHENER, C. D. Comparative social behavior of bees. **Annual Review of Entomology**, v. 14, p. 299–342, 1969.
MICHENER, C. D. **The Bees of the World**. second ed. [s.l.] JHU Press, 2007.
MOURE, J. S.; MELO, G. A. R.; VIVALLO, F. **Catalogue of Bees (Hymenoptera, Apoidea) in the Neotropical Region - online version.** Disponível em: <http://www.moure.cria.org.br/catalogue>.
NEVES, C. M. DE L. et al. Morphometric Characterization of a Population of *Tetrapedia diversipes* in Restricted Areas in Bahia, Brazil (Hymenoptera: Apidae). **Sociobiology**, v. 59, n. 3, p. 767–782, 2012.
NOGUEIRA-NETO, P. **A vida e criação de abelhas indígenas sem ferrão**. [s.l.] Nogueirapis, 1997.
NOWAK, M. A.; TARNITA, C. E.; WILSON, E. O. The evolution of eusociality. **Nature**, v. 466, n. 7310, p. 1057–1062, 2010.
NUNES, T. M. et al. Caste-specific cuticular lipids in the stingless bee *Friesella schrottkyi*. **Apidologie**, v. 41, n. 5, p. 579–588, 29 set. 2010.
PAXTON, R. J. et al. Low mating frequency of queens in the stingless bee *Scaptotrigona postica* and worker maternity of males. **Behavioral Ecology and Sociobiology**, v. 53, n. 3, p. 174–181, 2003.
RASMONT, P. et al. An overview of the Bombus terrestris (L. 1758) subspecies (Hymenoptera: Apidae). **Annales de la Société Entomologique de France (N.S.)**, v. 44, n. 2, p. 243–250, 2008.
REHAN, S. M.; TOTH, A. L. Climbing the social ladder: The molecular evolution of sociality. **Trends in Ecology and Evolution**, v. 30, n. 7, p. 426–433, 2015.
RIBEIRO, M. DE F. **Larval nutrition in the bumble bee *Bombus terrestris*, and its influence in caste differentiation**. [s.l.] Universiteit Utrecht, 1997.
ROBINSON, G. E.; BEN-SHAHAR, Y. Social behavior and comparative genomics: new genes or new gene regulation? **Genes, brain, and behavior**, v. 1, n. 4, p. 197–203, nov. 2002.
ROBINSON, G. E.; FAHRBACH, S. E.; WINSTON, M. L. W. Insect societies and the molecular biology of social behavior. **Bioessays**, v. 19, n. 12, p. 1099–1108, 1997.
ROBINSON, G. E.; FERNALD, R. D.; CLAYTON, D. F. Genes and social behavior. **Science**, v. 322, n. 5903, p. 896–900, 7 nov. 2008.
ROBINSON, G. E.; GROZINGER, C. M.; WHITFIELD, C. W. Sociogenomics: social life in molecular terms. **Nature Reviews Genetics**, v. 6, n. 4, p. 257–70, 2005.
ROCHA-FILHO, L. C.; GARÓFALO, C. A. Natural History of *Tetrapedia diversipes* (Hymenoptera: Apidae) in an Atlantic Semideciduous Forest Remnant Surrounded by Coffee Crops, *Coffea arabica* (Rubiaceae). **Annals of the Entomological Society of America**, v. 0, n. 0, p. 1–15, 2015.
ROMIGUIER, J. et al. Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. **Molecular Biology and Evolution**, v. 33, n. 3, 2015.
SARAIVA, A. M. et al. *Bombus terrestris* na América do Sul: Possíveis rotas de invasão deste polinizador exótico até o Brasil. In: IMPERATRIZ-FONSECA, V. L. et al. (Eds.). **Polinizadores no Brasil**. [s.l.] Edusp, 2012. p. 315–334.

SCHULZ, D. J.; ROBINSON, G. E. Octopamine influences division of labor in honey bee colonies. **Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology**, v. 187, n. 1, p. 53–61, 2001.

SILVA, C. I. DA et al. **Manejo dos Polinazadores e Polinização de Flores do Maracujazeiro**. 1. ed. Fortaleza - CE: Editora Fundação Brasil Cidadão, 2014.

SILVEIRA, F. A.; MELO, G. A. R.; ALMEIDA, E. A. B. **Abelhas Brasileiras: Sistemática e Identificação**. First edit ed. Belo Horizonte: [s.n.].

SIMOLA, D. F. et al. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. **Genome Research**, v. 23, n. 8, p. 1235–1247, 2013.

SMITH-PARDO, A. H. The Bees of the Genus *Neocorynura* of Mexico (Hymenoptera: Halictidae: Augochlorini). **Folia Entomol. Mex**, v. 44, n. 2, p. 165–193, 2005.

SØVIK, E.; BLOCH, G.; BEN-SHAHAR, Y. Function and evolution of microRNAs in eusocial Hymenoptera. **Frontiers in Genetics**, v. 6, n. MAY, p. 1–11, 2015.

THOMPSON, G. J.; OLDROYD, B. P. Eavaluating alternative hypotheses for the origin of eusociality in corbiculate bees. **Molecular Phylogenetics and Evolution 33**, v. 33, p. 452–456, 2004.

TOTH, A. L. et al. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. **Science**, v. 318, n. 5849, p. 441–4, 19 out. 2007.

TOTH, A. L.; REHAN, S. M. Molecular Evolution in Insect Societies: An Eco-Evo-Devo Synthesis. **Annual Review of Entomology**, v. 62, n. 1, p. 419–442, 7 jan. 2017.

VAN BENTHEM, F. D. J.; IMPERATRIZ-FONSECA, V. L.; VELTHUIS, H. H. W. Biology of the stingless bee *Plebeia remota* (Holmberg): observations and evolutionary implications. **Insectes Sociaux**, v. 42, n. 1, p. 71–87, 1995.

WILSON, E. O. **The Insect Societies**. First ed. [s.l: s.n.].

WILSON, E. O. **Sociobiology : the new synthesis**. [s.l.] Belknap Press of Harvard University Press, 2000.

WITTMANN, D. Aerial defense of the nest by workers of the stingless bee *Trigona* (*Tetragonisca*) *angustula* (Latreille) (Hymenoptera: Apidae). **Behavioral Ecology and Sociobiology**, v. 16, n. 2, p. 111–114, 1985.

WOODARD, S. H. et al. Genes involved in convergent evolution of eusociality in bees. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 18, p. 7472–7477, 2011.

WOODARD, S. H. et al. Molecular heterochrony and the evolution of sociality in bumblebees (*Bombus terrestris*). **Proceedings of the Royal Society B: Biological Sciences**, v. 281, n. 1780, 19 fev. 2014.

ZHANG, Z. H. et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. **PLoS ONE**, p. 0–35, 2014.

Objetivos e Organização da Tese

O objetivo geral dos estudos apresentados nesta tese foi investigar a evolução do comportamento eusocial em abelhas pela análise comparativa dos padrões de expressão de genes em espécies com diferentes níveis de organização social e em diferentes estágios de desenvolvimento. Para tanto utilizamos como principal ferramenta a técnica de RNA-Seq, o que possibilitou a análise do transcriptoma de muitas espécies não-modelo em diferentes abordagens comparativas (Capítulos de 2 à 4).

Para aumentar a eficiência dessas análises tivemos uma preocupação especial com a qualidade dos transcritos montados e, em uma das etapas de tratamento dos dados, verificamos que informações muitas vezes negligenciadas em certas análises, podem ser centrais a outras abordagens. O que nos motivou a incluir o **Capítulo 1** nesta tese. A presença de duas gerações reprodutivas em *T. diversipes*, também possibilitou o estudo de genes envolvidos no bivoltinismo, e apesar deste não ser um comportamento social, é possível que este tipo de estratégia reprodutiva esteja envolvida com o surgimento do comportamento social. Por isso, incluímos essas análises no **Capítulo 2**. Contudo, as principais perguntas que motivaram esta tese foram: **I-** Quais genes, redes regulatórias e características ontológicas estão envolvidas nas diferentes facetas do comportamento social? **II-** Existe alguma conservação entre esses genes, redes regulatórias e características ontológicas em diferentes espécies com o mesmo comportamento, ou seja, existe um *toolkit* genômico comportamental? E essas questões são abordadas nos **Capítulos 3 e 4**.

Assim, a tese está dividida em 4 capítulos de análises redigidos no formato de manuscritos, no idioma e formatação de publicação. Devido aos direitos de *copyright* de algumas revistas, a versão disponível na tese não representa a versão final do artigo, ou seja, a versão apresentada aqui pode ser diferente da versão disponibilizada pela revista. A seguir são descritos, resumidamente, os objetivos gerais de cada capítulo.

Chapter 1

Nesta seção é descrito como dados de transcriptomas podem ser usados em diferentes abordagens de estudo. Em especial, como transcritos considerados contaminantes podem se tornar dados importantes para análises.

Chapter 2

O objetivo deste capítulo foi identificar genes envolvidos no bivoltinismo da abelha solitária *Tetrapedia diversipes*, um dos principais modelos utilizados nessa tese. Este tipo de padrão reprodutivo pode levar a implicações evolutivas importantes na estrutura populacional e social da espécie, como discutido.

Chapter 3

Neste capítulo o objetivo foi estudar outra faceta importante do comportamento social, a divisão de subcastas em abelhas eussociais. Neste estudo identificamos genes envolvidos no comportamento de nutrizes e forrageiras de diferentes espécies, que compartilham uma única origem evolutiva do comportamento eussocial. Desta forma buscamos verificar a conservação das vias gênicas envolvidas.

Chapter 4

Este capítulo representa o capítulo mais audacioso da tese. Uma análise comparativa abrangente entre diferentes espécies e níveis de comportamento em que buscamos: a- identificar genes envolvidos no comportamento social; b- verificar o padrão de metilação desses genes; c- verificar a ocorrência desses genes em diversas espécies. Para a identificação dos genes e análise de metilação utilizamos três espécies de abelhas como modelo e, para verificar a existência destes genes em outras abelhas, foram amostradas outras 9 espécies com comportamentos sociais distintos.

Chapter 1

Getting Useful Information from RNA-Seq Contaminants: A Case of Study in the Oil-Collecting Bee *Tetrapedia diversipes* Transcriptome

Natalia de Souza Araujo, Alexandre Rizzo Zuntini and Maria Cristina Arias

Publicado em: *OMICS A Journal of Integrative Biology* (2016). DOI: 10.1089/omi.2016.0054

Keywords: Transcriptomics / Next Generation Sequencing / Databases

To the Editor:

The RNA-Seq is a straightforward technique widely used in studies of gene expression, especially for non-model species. This approach results in a comprehensive data set of genes expressed and their frequency without the need of species-specific probes or a reference genome. Because plant and animal species are constantly interacting with each other and the RNA-Seq is not a species-specific approach, it is highly probably that one can find genes from alien species in a transcriptome dataset. Indeed, as reported herein, the data analyses of *Tetrapedia diversipes* transcriptome, revealed contaminant transcripts from plants and parasites. The deep exploitation of this plant contaminant dataset proved to be a useful source of information concerning the biology and behaviour of this bee.

The oil collecting bee *Tetrapedia diversipes* is a solitary species native of the Neotropical region. This species is bivoltine, i.e. presents two main reproductive generations during the year, one in the hot and wet season (generation one – G1) and the second during the cold and dry season (generation two – G2). The developmental cycle since egg till adult varies significantly between the two generations because the pre-pupal larvae from G2 enter in diapause (Alves-dos-Santos *et al*., 2006). To understand the differences between each reproductive generation we have used the

RNA-Seq technique to sequence the transcriptome from female foundresses from G1 and G2.

Nine *T. diversipes* from each reproductive generation were collected in front of trap nests at the city of São Paulo (Brazil) in an area close to a small secondary semi-deciduous forest containing many native and ornamental plants (Alves-dos-Santos *et al.*, 2006). To identify the contaminant transcripts, complete assembled transcriptome from G1 and G2 of *T. diversipes* foundresses were blasted against the UniRef database (August, 2015) using the Annocript program (v1.2 – Musacchia *et al.*, 2015). Scripts in R (v3.1.3), bash, Python (v2.7.9) and manual checking were then used to identify and select contaminant transcripts from plants (pipeline and scripts available at https://github.com/nat2bee/trans-contamination/tree/master).

From the transcriptomes of G1 and G2 female foundresses, respectively, 857 and 538 transcripts were identified as plant contaminants. Contaminant transcripts from G1 blasted against 28 plant families and almost 50% of them (13) were found exclusively in G1. While in G2, 19 different families were identified and four were only found in G2 (Table I). These results indicates that the richness of plants visited by females from G1 is greater than from plants visited during G2, which may be related to the floral bloom during spring. Our data corroborate early study on pollen diversity storage in *T. diversipes* nests (Menezes *et al.*, 2012).

Table I presents all plant families identified among the contaminant transcripts in each generation and their proportion in the dataset. These findings are in agreement with previous ecological studies (Alves-dos-Santos *et al.*, 2006, Menezes *et al.*, 2012) specially regarding the use of the Euphorbiaceae as the main pollen source in larval feeding. Furthermore, because Amaranthacea and Euphorbiaceae are the two main families visited during G1 and Euphorbiaceae the main source during G2, the hypothesis that *T. diversipes* is not a truly polilectic species but have preferences for specific families is supported. As oil source, it is known that this bee uses plants from the Malpighiacea family but it is not clear if other families are also visited (Alves-dos-Santos *et al.*, 2006). In the present dataset transcripts from the Cucurbitaceae and Solanaceae families were found in both generations, which suggests that these families may be also visited for oil collection.

Therefore, as described here, the use of contaminant transcripts might be a useful source of information not only for the study of insect-plant interactions but also

for analyses of other associations such as parasitism and symbioses. This data is usually neglected in transcriptomic studies, but the present results indicate that contaminant transcripts from any transcriptomic dataset can be extremely valuable to answer different biological questions.

Nevertheless during this type of analyses it is important to keep in mind that the public databases used for identification through blast are incomplete and transcripts identification may be deficient in some cases. Also when the transcripts are from highly conserved genes the identification of a taxonomic group may be compromised. Thus, it is recommended the use of the reported approach associated with ecological observations or as a general and comparative tool.

Acknowledgments

The authors would like to thank Isabel Alves-dos-Santos for the support during bee collection, to Susy Coelho for technical assistance, to FAPESP (São Paulo Research Foundation, processes number 2013/12530-4 and 2012/18531-0) for financial support and to the reviewers for suggestions. This work was developed in the Research Center on Biodiversity and Computing (BioComp) of the Universidade de São Paulo (USP), supported by the USP Provost's Office for Research.

Author Disclosure Statement

The authors have no competing financial interests to declare.

Table I. Classification, numbers and proportion of the contaminant transcripts from plants found in the transcriptome of *T. diversipes* foundresses from generation one (G1) and two (G2).

Plant Family	No. of G1 contaminant transcripts	% of G1 contaminant transcripts	No. of G2 contaminant transcripts	% of G2 contaminant transcripts
Aizoaceae	1	0.13	-	-
Amaranthaceae	255	31.95	1	0.20
Arecaceae	6	0.75	-	-
Asteraceae	2	0.25	-	-
Brassicaceae	11	1.38	6	1.19
Cactaceae	1	0.13	-	-
Caryophyllaceae	2	0.25	-	-
Chenopodiaceae	3	0.38	-	-
Cleomaceae	2	0.25	5	0.99
Curcubitaceae	8	1.00	4	0.79
Euphorbiaceae	329	41.23	331	65.67
Fabaceae	16	2.00	10	1.98
Lamiaceae	1	0.13	-	-
Lentibulariaceae	2	0.25	-	-
Lythraceae	-	-	1	0.20
Malvaceae	19	2.38	25	4.96
Moraceae	-	-	6	1.19
Musaceae	3	0.38	-	-
Myrtaceae	4	0.50	44	8.73
Nelumbonaceae	2	0.25	6	1.19
Oleaceae	2	0.25	-	-
Onagraceae	-	-	3	0.60
Pedaliaceae	30	3.76	-	-
Phrymaceae	33	4.14	-	-
Poaceae	1	0.13	1	0.20
Rhizophoraceae	-	-	1	0.20
Rosaceae	14	1.75	17	3.37
Rubiaceae	6	0.75	-	-
Rutaceae	5	0.63	10	1.98
Salicaceae	18	2.26	22	4.37
Solanaceae	11	1.38	8	1.59
Vitaceae	11	1.38	3	0.60

References

Alves-dos-Santos I, Naxara SRC, Patrício EFLRA (2006). Notes on the Morphology of *Tetrapedia diversipes* Klug 1810 (Tetrapediini, Apidae), an Oil-collecting bee. Braz. J. Morphol. Sci. 23(3-4), 425-430.

Menezes GB, Gonçalves-Esteves V, Bastos EMAF, Augusto SC, Gaglianone MC (2012). Nesting and use of pollen resources by *Tetrapedia diversipes* Klug (Apidae) in Atlantic Forest areas (Rio de Janeiro, Brazil) in different stages of regeneration. Rev. Bras. Entomol. 56(1), 86–94.

Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R (2015). Annocript: a flexible pipeline for the annotation of transcriptomes also able to identify putative long noncoding RNAs. Bioinformatics, 31(13), 2199-2201.

Chapter 2

RNA-Seq reveals that mitochondrial genes and long noncoding RNAs may play important roles in the bivoltine generations of the non-social Neotropical bee *Tetrapedia diversipes*

Natalia S. Araujo, Priscila Karla F. Santos and Maria Cristina Arias

Submetido à: *Apidologie* (2017)

Keywords: transcriptome / bivoltinism / diapause / solitary bee / RNA-Seq

Short title: Genes involved in bivoltinism

Abstract

In animals, the voltinism is a result of evolutionary adaptations to environmental conditions. These evolutionary adaptations may affect the population structure and social organization level profoundly. To study the bivoltinism of the solitary bee *Tetrapedia diversipes* we have performed comparative transcriptomic analyses of foundresses and larvae from the two reproductive generations (G1 and G2) produced per year by this bee. Most of the differentially expressed genes (DEG) were found between the foundresses: 52 DEG between adults, but only one between the larvae. From the DEG in foundresses 46 were higher expressed in G1 and most of them (38) have no function defined in the database. Interestingly, mitochondrial genes and lncRNAs are the only identified transcripts in this set of up-regulated genes. These results highlight the importance of developing new studies for non-model species and suggest that maternal genes in *T. diversipes* might be important in larvae diapause.

Introduction

The number of reproductive generations presented by a species in one year determines its voltinism (Corbet et al. 2006). In insects, the voltine trait is considered the result of evolutionary adaptations to environmental conditions, such as temperature, humidity, latitude and food resources (Corbet et al. 2006; Altermatt 2010; Cardoso & Silveira 2012; Hunt 2012). However, this attribute may have had profound consequences at the levels of population structure and social organization (Hunt & Amdam 2005). This is especially noticeable in species where sociality has evolved recently as in Halictinae bees (Brady et al. 2006). Within the Halictinae, it has been demonstrated that social behavior may be facultative and directly related to voltine generations (Yanega 1988). *Halictus rubicundus*, for example, presents solitary behavior when individuals are univoltine, nonetheless they may become social when individuals are bivoltine (two generations in the year) and enter in diapause (Soucy 2002). Indeed, bivoltinism encompassing diapause in one offspring generation is the biological system in which Hunt's bivoltine ground plan hypothesis on the evolution of insect sociality in temperate climates is based (Hunt & Amdam 2005). According to this hypothesis, social behavior could evolve from solitary species if females from the first reproductive generation remained in the nest, and changes in environmental conditions suppressed the entry into prepupal diapause in the following generation (Hunt & Amdam 2005; Hunt 2012). Thus, studies of ecological and molecular mechanisms involved in voltinism can be important pieces to understand the puzzle of bee social behavior and evolution.

In the tropics, bivoltinism has been observed in different solitary bees (Silveira et al. 2002; Alves-dos-Santos et al. 2007) including *Tetrapedia diversipes* Klug 1810 (subfamily Apinae, tribe Tetrapediini, Michener, 2007). This oil-collecting bee, native of the Neotropical region, has two main reproductive generations during the year, each presenting different development times (Alves-dos-Santos et al. 2002). Individuals of the first generation (G1) have a direct development from egg to adults, emerging within few weeks during the hot and wet months. Yet, individuals from the second generation (G2) halt their development in the fifth larval instar during the cold and dry season and emerge as adults only after a diapause period (Camillo 2005; Alves-dos-Santos et al. 2006). Due to this diapause, the development time of G2 individuals may be up to four times longer in comparison to G1 (Alves-dos-Santos et

al. 2002). In addition to this interesting developmental aspect, *T. diversipes* is an attractive model species to study because it easily nidifies in trap nests (Camillo 2005; Alves-dos-Santos et al. 2006; Menezes et al. 2012; Neves et al. 2012; Rocha-Filho & Garófalo 2015), it has molecular markers that are already characterized (Arias et al. 2016), and its genome is currently being sequenced. Combined, all these characteristics make *T. diversipes* a promising emerging model for developmental and evolutionary studies on the transition from a solitary to a social lifestyle in tropical climates.

Herein we report the complete transcriptomes of foundress females and non-diapause larvae of *T. diversipes* and a comparative gene expression analysis between the two reproductive generations using next generation sequencing (RNA-Seq). This sequencing approach has been widely used in gene expression studies, especially for non-model species. As it provides a non-directional way of obtaining data on the relative frequencies of mRNAs with no need of species-specific probes or a reference genome (Wang et al. 2009).

Material and Methods

Sample collection

Wooden trap nests (Alves-dos-Santos et al. 2002) were placed near a small secondary semi-deciduous forest containing native and ornamental plants (Alves-dos-Santos et al. 2006; Araujo et al. 2016) in São Paulo city, Brazil. Foundresses were collected between 10:00 and 12:00 in front of their nests while constructing by using an entomological net. Larvae were collected directly from inside the nests that had been completed and closed by the foundresses. All instars (1^{st} to 5^{th}) in the non-diapause state (from G1, and G2 before entering diapause) were sampled. Individuals were immediately frozen in liquid nitrogen. G1 samples were collected from November to December (mean temperature 26.4 °C; mean humidity 62.2%) and G2's from March to beginning of July (mean temperature 24 °C; mean humidity 59.2%).

RNA extraction and Sequencing

Total RNA was extracted from the individuals whole body using the RNeasy® kit (Qiagen) and following the manufacturer's protocol. RNA quality was verified by the sequencing facility (Macrogen) using a Bionalyzer® system, and results were

interpreted as discussed in Winnebeck et al. (2010). Nine foundresses and nine larvae from each generation were selected for RNA sequencing. Samples were divided into three replicates, each containing the RNA of three individuals from the same developmental stage. This sampling approach was adopted to improve gene identification and differential expression analyses (Hart et al. 2013; Lin et al. 2016). Altogether 12 samples were sequenced, 6 from G1 (3 pools of adults and 3 of larvae) and 6 from G2 (3 pools of adults and 3 of larvae), using a HiSeq2000® sequencer (Illumina). Paired-end reads of 100 bp were sequenced and about 50 million paired reads were obtained per sample. Sequencing and library preparation were performed by Macrogen.

Transcriptome assembly

Reads quality assessment was performed with the FastQC program (version 0.11.2, Andrews 2010) before and after cleaning. The FASTX Toolkit (version 0.0.14 - Hannon Lab 2009) was used to trim the first 14 bp of all reads because of the initial GC bias (Hansen et al. 2010). Low quality bases (phred score below 30) and small reads (less than 31 bp) were removed by SeqyClean program (version 1.9.3, Zhbannikov 2013).

Previous to assembly, the data were digitally normalized (20x coverage) to increase assembly efficiency using the Brown et al. (2012) protocol incorporated in Trinity (version 2.0.6 - Grabherr et al. 2013). Normalized data from each reproductive generation and development stage were independently assembled *de novo* using Trinity with default parameters. Assemblies were then concatenated with the CD-Hit program (version 4.6, Huang et al. 2010) at 95% similarity. Then all cleaned samples were realigned to this concatenated assembly by TopHat2 program (version 2.1.0, Kim et al. 2013) and these realignments were used as input in Corset (version 1.03, Davidson & Oshlack 2014) with minimum coverage of 50x to improve transcripts identification.

Transcripts were then annotated with Annocript (version 1.2, Musacchia et al. 2015) using the UniProt Reference Clusters (UniRef) database (version February 2016, Suzek et al. 2015). Transcripts with significant blast hits (e-value < 1e-5) against possible contaminants (plants, fungus, mites and bacteria) in UniRef were removed from the final dataset and were also used to identify other contaminants

based on cluster analysis from Corset, as described in Araujo et al. (2016). Quality assessments of the final assembly were performed with QUAST (version 4.0, Gurevich et al. 2013), BUSCO (version 2, Simão et al. 2015) and Qualimap (version 2.2, García-Alcalde et al. 2012).

Differential expression analysis

The Trinity script was used to automate the differential expression analyses using the Bowtie2 program (version 2.2.5, Langmead & Salzberg 2012) to realign all cleaned samples to the final transcriptome; the RSEM program (version 1.2.22, Li & Dewey 2011) to count the realigned reads; and the edgeR package (version 3.14.0 - Robinson et al. 2009) for the statistical analyses (minimum FDR p-value < 1e-5).

Results

Transcriptome assembly

Main parameters from final transcriptome assemblies of foundresses and larvae are simplified in Table I. The final adult transcriptome (after removal of contaminants) contains 44,486 transcripts, of which 27,098 are reported by at least one blast hit. The Annocript pipeline also identified that from all the transcripts, 709 are probably long noncoding RNAs (lncRNAs) and 32,037 have coding potential based on their open reading frame (ORF) or annotation. The larval transcriptome had similar results, presenting a total of 41,354 transcripts, of which 27,393 had at least one blast hit, 30,975 are potentially coding and 539 were identified as potential lncRNAs. Final transcriptome assemblies and annotation tables, containing all the statistics for coding and lncRNAs, are available at GitHub (https://github.com/nat2bee/bivoltine_apidologie).

The database for Hymenoptera orthologs available via BUSCO software has a total of 4,415 genes of which 3,645 (82.6%) were identified as complete in the *T. diversipes* adult transcriptome and 3,432 (77.7%) in the larval one. A small portion was identified as fragmented genes, 268 (6.1%) and 358 (8.1%) in the adult and larval transcriptome, respectively. Thus, only 502 (11.3%) and 625 (14.2%) of all Hymenoptera orthologs were missing in the adult and larval final dataset, respectively.

Table I Main quality parameters from the complete transcriptome assembly of *T. diversipes* foundresses and non-diapause larval stages (1st to 5th instars). Annotated transcripts refer to transcripts with at least one blast hit against the UniRef database

	Foundresses	**Larvae**
Total Number of transcripts	44,486	41,354
N50	3,243	3,028
%GC	37.95%	37.03%
Mean Coverage per sample	48x	53x
Annotated transcripts	27,098	27,393
Complete BUSCO orthologous	82.6%	77.7%

Differential expression analyses

Gene expression analyses with the edgeR program reported 52 differentially expressed genes (DEG) between the two adults generations (Online Resource 1). Among these, 46 genes were highly expressed in adult G1 females and 6 had elevated levels of expression in females from G2 (Fig. 1). Four G1 up-regulated genes (*cytochrome b*; *COII*; *NADH4* and one predicted uncharacterized protein) and three up-regulated genes from G2 (*defensin-2*; *DHRS11* and one uncharacterized protein) had a significant blast hit against the UniRef database (Fig. 1 and Online Resource 2). Other five of the higher represented transcripts in G1 females were also reported as possible lncRNAs by Annocript according to their ORF (! 100 bp), annotation to databases and sequence length (" 200 bp) (Fig. 1 and Online Resource 2). When comparing the larval generations, only one DEG was found, and this was higher expressed in individuals from G2 (Online Resource 1). This gene was annotated as a *replicase polyprotein* according to the UniRef database (Online Resource 2).

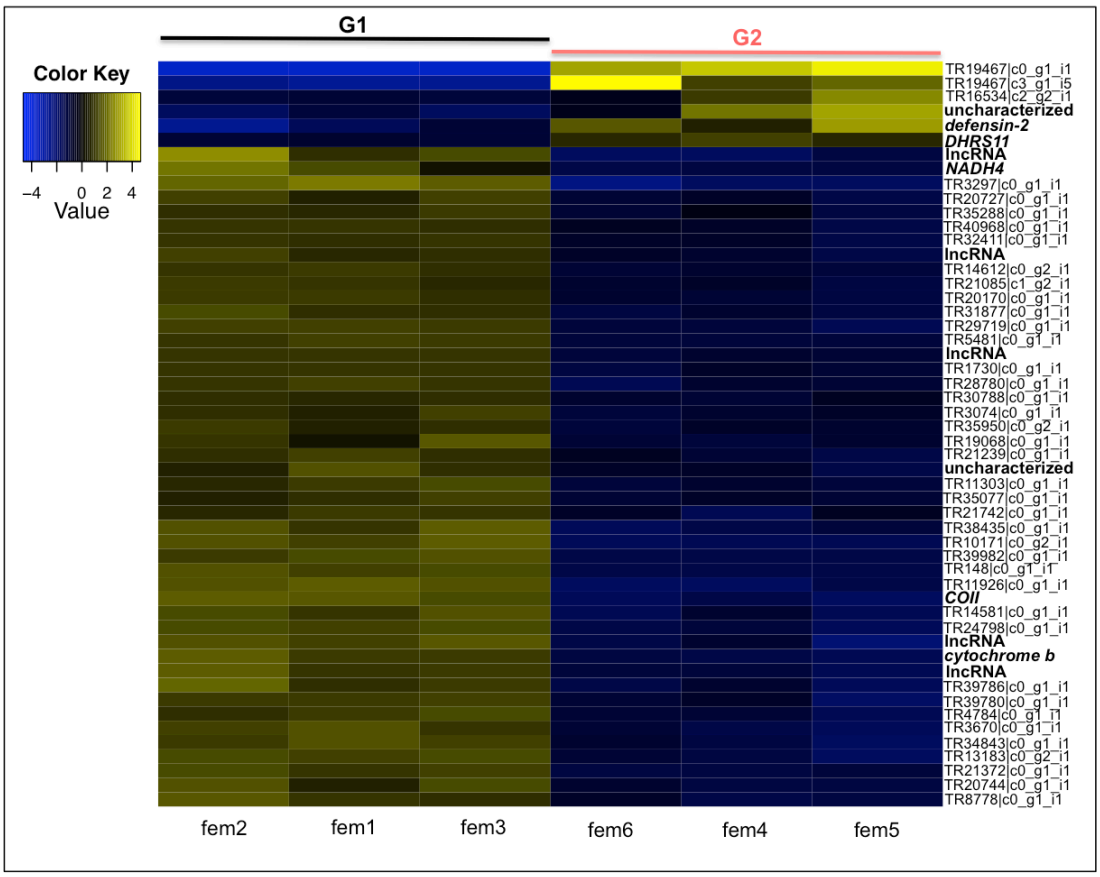


Fig. 1 Heatmap of the differentially expressed genes among *T. diversipes* adult females. fem1, fem2 and fem3 are replicated samples from G1 foundresses; fem4, fem5 and fem6 are replicated samples from G2 foundresses. Indicated in bold on the y-axis are the names of the annotated transcripts and lncRNAs. Uncharacterized refers to transcripts that were retrieved as uncharacterized proteins by blast queries against the UniRef database. Expression scale is log2

Discussion

We report for the first time the transcriptome assembly of female foundresses and non-diapause larvae of the solitary bee *T. diversipes*, which is an emerging model for voltinism and its possible consequences for the evolution of sociality. Quality analyses of the transcripts indicate that the assembly is comparable to other bee transcriptomes (Colgan et al. 2011; Kocher et al. 2013; Rehan et al. 2014; Harrison et al. 2015), and the high sequencing coverage (about 150 x per transcriptome) allowed the identification of alternative transcripts, as described elsewhere (Klerk & Hoen 2015). In both life cycle stages more than 60% (60.9% in adult females and 66.2% in larvae) of the transcripts have a significant blast hit against a sequence in the UniRef database, especially against bees and other Hymenoptera sequences (Online Resource 3). Moreover, it is expected that an additional 8% of the non annotated transcripts are

protein coding according to their ORF characteristics, but they did not have a significant blast hit, representing genes that are either novel or significantly divergent from any database sequence. Leaving about 30% of possible miss assemblies, non-coding RNAs, or transcriptional noise (retained introns and similar) assembled in *T. diversipes* transcriptomes.

Differential expression analyses revealed that between non-diapause larvae from G1 and G2 only the *replicase polyprotein* was up-regulated in G2. The *replicase polyprotein* or *RNA-dependent RNA polymerase* (*RdRp*) is a well known gene that encodes for a protein essential for genome replication of RNA viruses (Ahlquist 2002; Iyer et al. 2003), but recently it has been demonstrated that many eukaryotes also have one or more *RdRp* copies in their genomes (Nishikura 2001; Ahlquist 2002; Anantharaman et al. 2002; Chapman & Carrington 2007). In these cases, *RdRps* assume a different biological function and act directly in post-transcriptional regulatory mechanisms (especially in RNA silencing) and host response (Ahlquist 2002; Anantharaman et al. 2002; Chapman & Carrington 2007). Thus the higher expression of this gene in G2 larvae may indicate either a viral infection or a change in gene expression regulation when compared to larvae in G1, depending on the gene provenance. It is possible, for example that once G2 larvae will enter diapause, the higher expression of the *RdRp* may be involved with gene expression cascade changes necessary for diapause onset. However, in several arthropod genomes, including *Drosophila*, no homologs of the *RdRp* gene have been detected (Zong et al. 2009). Thus, although post-transcriptional regulatory mechanisms related to *RdRps* are functional in insects, it is not clear if other proteins are involved in the process instead (Ahlquist 2002; Zong et al. 2009). Therefore, identifying the true origin of the reported *RdRp* will only be possible after completion of the *T. diversipes* genome assembly.

Most of the DEG between the two reproductive generations were reported in foundresses. In G2 foundresses, whose offspring will enter diapause, six transcripts were higher represented. One of these, the *defensin-2* gene, is known to be expressed in honey bee fat body and is responsible for immune defense against a number of parasites (Klaudiny et al. 2005). For *Apis mellifera*, it has been reported that the level of *defensing-2* expression only increases in response to infections (Ilyasov et al. 2012). Thus, our finding suggests that G2 females are more likely to be exposed to

pathogens than G1 females. Another higher expressed gene in G2 foundresses that also may be involved in some type of immune response is the *dehydrogenase/reductase SDR family member 11* (*DHRS11*), a NAD(P)(H)-dependent oxidoreductase gene that belongs to the large superfamily of *short-chain dehydrogenase/reductases* (Persson et al. 2009; Endo et al. 2016). Genes from this superfamily are known to be involved in the metabolism of lipids, carbohydrates, vitamins, drugs and xenobiotics (Endo et al. 2016). Specifically in honeybees, *DHRS11* has been found to play a role in host resistance against the Varroa mite (Parker et al. 2012), and in larval caste differentiation it appears as one of the genes possibly necessary for worker ovary differentiation (Guidugli et al. 2004; Lago et al. 2016). Also, *DHRS11* product is one of the proteins present in bee venom (Li et al. 2013).

Among the higher expressed genes identified in G1 foundresses are *COII*, *NADH4* and *cytochrome b*, genes involved in energy production through oxidative phosphorylation (Cooper 2000). This suggests that adult females from G1 have higher energy demands than G2 females, which can be driven by environmental factors such as food availability and temperature. It is also worthy of attention that other five higher expressed genes in females from G1 are putative lncRNAs. LncRNAs are poly(A) RNAs which, different form other mRNAs are not translated to a functional protein. Presently, lncRNAs are defined as probable non-coding RNAs longer than 200 bp (Rinn & Chang 2012). These transcripts have recently emerged as critical elements in many genetic regulatory mechanisms because they are proved to act in a number of processes, such as: epigenetic regulation; DNA imprinting; control of cell pluri-potency; transcriptional silencing; co-activation, among others (Rinn & Chang 2012; Kung et al. 2013). Including in honey bees, where one specific lncRNA, lncov-1, was shown to be associated with programmed cell death in the larval worker ovary (Humann et al., 2013).

Therefore, two important aspects in *T. diversipes* bivoltine generations emerged from our analyses of gene expression. First, the *RdRp* gene, if not related to viral infection, may be one of the first genes to have its expression pattern specifically regulated in pre-diapause larvae from G2. Second, most of the observed expression differences between G1 and G2 individuals occur on adult females. Thus, DEG of

foundresses might be important to trigger diapause in the subsequent offspring generation.

Diapause induced by changes in the mother's metabolism is common in many insects (Denlinger 2002). This induction may be direct, as in the case of the silkworm *Bombyx mori* that produces a specific diapause hormone (Akitomo et al. 2015), or indirect, where the correlation between progenitor gene expression and brood diapause is not so clearly established, because even changes during the mother's development may affect the offspring (Denlinger 1998, 2002). It is possible that in *T. diversipes*, for example, the higher metabolic rates or the lncRNAs in the foundresses provides a molecular or epigenetic signal that prevents the larvae from entering diapause. Similar mechanisms of maternal epigenetic suppression in diapause have already been described in the fly *Sarcophaga bullata* and in the GABAergic circuit of *Bombyx mori* (Çabej 2013; Reynolds et al. 2013, 2016). Also lncRNAs have been reported as paramount in glucose and lipid metabolism, which are important energetic pathways for diapause (Denlinger 2002; Lang-Ouellette et al. 2014).

Furthermore it is important to notice that most of the DEG are new and their function therefore unknown. Thus, the other 42 non-characterized DEG reported in foundresses may also play an important role in inducing/ suppressing larval diapause, but we were unable to infer their biological relevance based on the sequences already available in genome databases. Although we have seen an expressive improvement in databases content over the last years (O'Leary et al. 2016), they are still highly biased to model species. Therefore, studies with non-model species are needed to improve our understanding of multiple biological systems.

Conclusions

Herein we report the complete transcriptome assembly from two different life stages of *T. diversipes*. Using these assemblies we compared gene expression differences between the two reproductive generations of *T. diversipes*. These analyses have shown that the *RdRp* is the only gene highly expressed in larvae from G2 before entering diapause, and that genes expressed in the mother are likely to influence the development time of the larval offspring. This can be inferred based on the fact that most of the gene expression differences observed occur in the adult life stage. Among

the higher expressed genes in G1 adult females, the only identified transcripts are the ones involved in oxidative phosphorylation and lncRNAs, while in G2 adult females higher expressed genes are possibly related to immune response.

Acknowledgments

The authors would like to thank Isabel Alves-dos-Santos and Guraci D. Cordeiro for the support during bee collection, to Susy Coelho for technical assistance, to FAPESP (São Paulo Research Foundation, process numbers 2013/12530-4 and 2012/18531-0) for financial support and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for scholarship to PKFS. This work was developed in the Research Centre on Biodiversity and Computing (BioComp) of the Universidade de São Paulo (USP), supported by the USP Provost's Office for Research. Part of the bioinformatic analyses was performed at the cloud computing service from USP.

References

Ahlquist, P. (2002) RNA-dependent RNA polymerases, viruses, and RNA silencing. Science, DOI:10.1126/science.1069132

Akitomo, S., Egi, Y., Nakamura, Y., Suetsugu, Y., Oishi, K., Sakamoto, K. (2015) Genome-wide microarray screening for *Bombyx mori* genes related to transmitting the determination outcome of whether to produce diapause or nondiapause eggs. Insect Sci., DOI:10.1111/1744-7917.12297

Altermatt, F. (2010) Climatic warming increases voltinism in European butterflies and moths. Proc. R. Soc. B, doi:10.1098/rspb.2009.1910

Alves-dos-Santos, I., Machado, I.C., Gaglianone, M.C. (2007) História Natural das Abelhas Coletoras de Óleo. Oecol. Bras., **11** (4), 544–557

Alves-dos-Santos, I., Melo, G.A.R., Rozen Jr, J.G. (2002) Biology and Immature Stages of the Bee Tribe Tetrapediini (Hymenoptera : Apidae). Am. Museum Nat. Hist., DOI:http://dx.doi.org/10.1206/0003-0082(2002)377

Alves-dos-Santos, I., Naxara, S.R.C., Patrício, E.F.L.R.A. (2006) Notes on the Morphology of *Tetrapedia diversipes* KLUG 1810 (Tetrapedini, Apidae), an oil-collecting bee. Braz. J. Morphol. Sci., **23** (3–4), 425–430

Anantharaman, V., Koonin, E.V., Aravind, L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res., DOI:10.1093/NAR/30.7.1427

Andrews, S. (2010) FastQC: A Quality Control tool for High Throughput Sequence Data. Babraham Bioinformatics [online] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 05 February 16)

Araujo, N.S., Zuntini, A.R., Arias, M.C. (2016) Getting Useful Information from RNA-Seq Contaminants: A Case of Study in the Oil-Collecting Bee *Tetrapedia diversipes* Transcriptome. Omics, DOI:10.1089/omi.2016.0054

Arias, M.C., Aulagnier, S., Baerwald, E.F., Barclay, R.M.R., Batista, J.S. et al. (2016) Microsatellite records for volume 8, issue 1. Conserv. Genet. Resour., DOI:10.1007/s12686-016-0522-2

Brady, S.G., Sipes, S., Pearson, A., Danforth, B.N. (2006) Recent and Simultaneous Origins of Eusociality in Halictid Bees. Proc. R. Soc. B, DOI:10.1098/rspb.2006.3496

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., Brom, T.H. (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. Genome Announc., **2** (4), e00802-14-e00802-14

Çabej, N. (2013) Living and Adapting to Its Own Habitat. In: Building the most complex structure on Earth : an epigenetic narrative of development and evolution of animals, pp. 193–238. Elsevier.

Camillo, E. (2005) Nesting biology of four *Tetrapedia* species in trap nests (Hymenoptera: Apidae: Tetrapedinii). Rev. Biol. Trop., **53**, 175–186

Cardoso, C.F., Silveira, F.A. (2012) Nesting biology of two species of *Megachile* (Moureapis) (Hymenoptera: Megachilidae) in a semideciduous forest reserve in southeastern Brazil. Apidologie, DOI:10.1007/s13592-011-0091-z

Chapman, E.J., Carrington, J.C. (2007) Specialization and evolution of endogenous small RNA pathways. Nat. Rev. Genet., DOI:10.1038/nrg2179

Colgan, T.J., Carolan, J.C., Bridgett, S.J., Sumner, S., Blaxter, M.L., Brown, M.J. (2011) Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, *Bombus terrestris*. BMC Genomics, DOI:10.1186/1471-2164-12-623

Cooper, G. (2000) The Mechanism of Oxidative Phosphorylation. In: The Cell: A Molecular Approach. Sinauer Associates, Sunderland (MA).

Corbet, P.S., Suhling, F., Soendgerath, D. (2006) Voltinism of odontada: a review. Int. J. Odonatol., **9** (1), 1–44

Davidson, N.M., Oshlack, A. (2014) Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol., DOI:10.1186/s13059-014-0410-6

Denlinger, D.L. (1998) Maternal control of Fly Diapause. In: Maternal effects as adaptations eds Mousseau TA, Fox CW, pp. 275–286. Oxford University Press.

Denlinger, D.L. (2002) Regulation of diapause. Annu. Rev. Entomol., DOI:10.1146/annurev.ento.47.091201.145137

FASTX Toolkit (2009) by Hannon Lab [online] http://hannonlab.cshl.edu/fastx_toolkit/index.html (accessed on 05 February 16)

Endo, S., Miyagi, N., Matsunaga, T., Hara, A., Ikari, A. (2016) Human dehydrogenase/reductase (SDR family) member 11 is a novel type of 17B-hydroxysteroid dehydrogenase. Biochem. Biophys. Res. Commun., DOI:10.1016/j.bbrc.2016.01.190

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., et al. (2012) Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics, DOI:10.1093/bioinformatics/bts503

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., et al. (2013) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol., DOI:10.1038/nbt.1883.Trinity

Guidugli, K.R., Hepperle, C., Hartfelder, K. (2004) A member of the short-chain dehydrogenase/reductase (SDR) superfamily is a target of the ecdysone response in honey bee (Apis mellifera) caste development. Apidologie, DOI:10.1051/apido:2003068

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics, DOI:10.1093/bioinformatics/btt086

Hansen, K.D., Brenner, S.E., Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res., DOI:10.1093/nar/gkq224

Harrison, M., Hammond, R., Mallon, E. (2015) Reproductive workers show queen-like gene expression in an intermediately eusocial insect, the buff-tailed bimble bee *Bombus terrestris*. Mol. Ecol., **24**, 3043–3063

Hart, S.N., Therneau, T.M., Zhang, Y., Poland, G.A., Kocher, J.P. (2013) Calculating Sample Size Estimates for RNA Sequencing Data. J. Comput. Biol., **20**, 1–9

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W. (2010) CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics, DOI:10.1093/bioinformatics/btq003

Humann, F.C., Tiberio, G.J., Hartfelder K. (2013) Sequence and Expression Characteristics of Long Noncoding RNAs in Honey Bee Caste Development – Potential Novel Regulators for Transgressive Ovary Size. Plos One, DOI:10.1371/journal.pone.0078915

Hunt, J.H. (2012) A conceptual model for the origin of worker behaviour and adaptation of eusociality. J. Evol. Biol., DOI:10.1111/j.1420-9101.2011.02421.x

Hunt, J.H., Amdam, G.V. (2005) Bivoltinism as an Antecedent to Eusociality in the Paper Wasp Genus *Polistes*. Science, DOI:308/5719/264 [pii]\r10.1126/science.1109724

Ilyasov, R., Gaifullina, L., Saltykova, E., Poskryakov, A., Nikolenko, A. (2012) Review of the Expression of Antimicrobial Peptide Defensin in Honey Bees *Apis Mellifera* L. J. Apic. Sci., DOI:10.2478/v10289-012-0013-y

Iyer, L.M., Koonin, E.V., Aravind, L. (2003) Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. BMC Struct. Biol., DOI:10.1186/1472-6807-3-1

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol., DOI:10.1186/gb-2013-14-4-r36

Klaudiny, J., Albert, S., Bachanova, K., Kopernicky, J., Simuth, J. (2005) Two structurally different defensin genes, one of them encoding a novel defensin isoform, are expressed in honeybee *Apis mellifera*. Insect Biochem. Mol. Biol.,

DOI:10.1016/j.ibmb.2004.09.007
Klerk, E., Hoen, P.A.C. (2015) Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. Trends Genet., DOI:10.1016/j.tig.2015.01.001
Kocher, S.D., Li, C., Yang, W., Tan, H., Yi, S.V. et al. (2013) The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. Genome Biol., DOI:10.1186/gb-2013-14-12-r142
Kung, J.T.Y., Colognori, D., Lee, J.T. (2013) Long noncoding RNAs: past, present, and future. Genetics, DOI:10.1534/genetics.112.146704
Lago, D.C., Humann, F.C., Barchuk, A.R., Abraham, K.J., Hartfelder, K. (2016) Differential gene expression underlying ovarian phenotype determination in honey bee, *Apis mellifera L*, caste development. Insect Biochem. Mol. Biol., DOI:10.1016/j.ibmb.2016.10.001
Lang-Ouellette, D., Richard, T.G., Morin, P. (2014) Mammalian hibernation and regulation of lipid metabolism: A focus on non-coding RNAs. Biochem., DOI:10.1134/S0006297914110030
Langmead, B., Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, **9** (4), 357–359
Li, B., Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, DOI:10.1186/1471-2105-12-323
Li, R., Zhang, L., Fang, Y., Han, B., Lu, X., et al. (2013) Proteome and phosphoproteome analysis of honeybee (*Apis mellifera*) venom collected from electrical stimulation and manual extraction of the venom gland. BMC Genomics, DOI:10.1186/1471-2164-14-766
Lin, Y., Golovnina, K., Chen, Z.X., Lee, H.N., Negron, Y.L.S., et al. (2016) Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. BMC Genomics, **17**, 1–20
Michener, C.D. (2007) The Bees of the World. JHU Press.
Menezes, G.B., Gonçalves-Esteves, V., Bastos, E.M.A.F., Augusto, S.C., Gaglianone, M.C. (2012) Nesting and use of pollen resources by *Tetrapedia diversipes* Klug (Apidae) in Atlantic Forest areas (Rio de Janeiro, Brazil) in different stages of regeneration. Rev. Bras. Entomol., DOI:10.1590/S0085-56262012000100014
Musacchia, F., Basu, S., Petrosino, G., Salvemini, M., Sanges, R. (2015) Annocript: A flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics, DOI:10.1093/bioinformatics/btv106
Neves, C.M. de L., Carvalho, C.A.L. de, Souza, V.A., Junior, C.A.L. (2012) Morphometric Characterization of a Population of *Tetrapedia diversipes* in Restricted Areas in Bahia, Brazil (Hymenoptera: Apidae). Sociobiology, **59** (3), 767–782
Nishikura, K. (2001) A short primer on RNAi: RNA-directed RNA polymerase acts as a key catalyst. Cell, DOI:10.1016/S0092-8674(01)00581-5
O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D. et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., DOI:10.1093/nar/gkv1189
Parker, R., Guarna, M.M., Melathopoulos, A.P., Moon, K.M., White, R., et al. (2012) Correlation of proteome-wide changes with social immunity behaviors provides insight into resistance to the parasitic mite, Varroa destructor, in the honey bee (*Apis mellifera*). Genome Biol., DOI:10.1186/gb-2012-13-9-r81
Persson, B., Kallberg, Y., Bray, J.E., Bruford, E., Dellaporta, S.L. et al. (2009) The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. Chem. Biol. Interact., DOI:10.1016/j.cbi.2008.10.040
Rehan, S.M., Berens, A.J., Toth, A.L. (2014) At the brink of eusociality: transcriptomic correlates of worker behaviour in a small carpenter bee. BMC Evol. Biol., DOI:10.1186/s12862-014-0260-6
Reynolds, J.A., Bautista-Jimenez, R., Denlinger, D.L. (2016) Changes in histone acetylation

as potential mediators of pupal diapause in the flesh fly, *Sarcophaga bullata*. Insect Biochem. Mol. Biol., DOI:10.1016/j.ibmb.2016.06.012

Reynolds, J.A., Clark, J., Diakoff, S.J., Denlinger, D.L. (2013) Transcriptional evidence for small RNA regulation of pupal diapause in the flesh fly, *Sarcophaga bullata*. Insect Biochem. Mol. Biol., DOI:10.1016/j.ibmb.2013.07.005

Rinn, J.L., Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. Annu. Rev. Biochem., DOI:10.1146/annurev-biochem-051410-092902

Robinson, M.D., McCarthy, D.J., Smyth, G.K. (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, DOI:10.1093/bioinformatics/btp616

Rocha-Filho, L.C., Garófalo, C.A. (2015) Natural History of *Tetrapedia diversipes* (Hymenoptera: Apidae) in an Atlantic Semideciduous Forest Remnant Surrounded by Coffee Crops, Coffea arabica (Rubiaceae). Ann. Entomol. Soc. Am., DOI:doi: 10.1093/aesa/sav153

Silveira, F.A, Melo, G.A.R., Almeida, E.A.B. (2002) Abelhas Brasileiras: Sistemática e Identificação. Belo Horizonte.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, DOI:10.1093/bioinformatics/btv351

Soucy, S.L. (2002) Nesting Biology and Socially Polymorphic Behavior of the Sweat Bee *Halictus rubicundus* (Hymenoptera : Halictidae). Entomol. Soc. Am., **95** (1), 57–65

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., UniProt Consortium (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, DOI:10.1093/bioinformatics/btu739

Wang, Z., Gerstein, M., Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., DOI:10.1038/nrg2484

Winnebeck, E.C., Millar, C.D., Warman, G.R. (2010) Why does insect RNA look degraded? J. Insect Sci., DOI:10.1673/031.010.14119

Yanega, D. (1988) Social plasticity and early-diapausing females in a primitively social bee. Proc. Natl. Acad. Sci. U. S. A., DOI:10.1073/pnas.85.12.4374

Zhbannikov, I. (2013) SeqyClean [online] https://bitbucket.org/izhbannikov/seqyclean.git (accessed on 05 February 16)

Zong, J., Yao, X., Yin, J., Zhang, D., Ma, H. (2009) Evolution of the RNA-dependent RNA polymerase (RdRP) genes: Duplications and possible losses before and after the divergence of major eukaryotic groups. Gene, DOI:10.1016/j.gene.2009.07.004

Attachments

Online resource 1 - Table I Matrix of differentially expressed genes counts reported by the edgeR program. lar1, lar2 and lar3 are replicated samples from G1 larvae; and lar4, lar5 and lar6 from G2. fem1, fem2 and fem3 are replicated samples from G1 foundresses; and fem4, fem5 and fem6 from G2

	Transcript ID	Sample code					
Larvae		lar1	lar2	lar3	lar4	lar5	lar6
	TR15541|c0_g1_i1	0	0.014355293	0	3.355298233	0.718964336	1.497229129
		fem1	fem2	fem3	fem4	fem5	fem6
	TR8778|c0_g1_i1	2.354170026	2.955312803	2.268434394	0	0.182692298	0.606915942
	TR19467|c3_g1_i5	0.169925001	0.365132593	0.33571191	4.15534447	4.768025487	7.551046835
	TR19145|c0_g2_i1	10.58305549	10.72578205	10.31958745	7.657175513	7.245343536	7.33800538
	TR11926|c0_g1_i1	3.454439135	3.2124137	3.161887682	0	0.289244285	0
	TR28780|c0_g1_i1	5.078225543	4.905158322	4.90636151	3.027154052	2.856188951	2.393965276
	TR4784|c0_g1_i1	2.742437445	2.58183327	3.067466629	0.695102986	0.207892852	0.668119125
	TR148|c0_g1_i1	2.65718266	2.869476634	2.818032475	0	0	0
	TR35077|c0_g1_i1	2.272918995	2.001802243	2.577972569	0.669933836	0.367371066	0.361768359
	TR20635|c0_g2_i1	0.670840336	0	0	3.83975808	4.722793872	1.266636643
	TR20744|c0_g1_i1	1.927896454	2.773996325	2.659239572	0	0	0.403813062
	TR39780|c0_g1_i1	3.985864701	3.949534933	4.087887101	2.064193062	0.982582947	1.787014474
	TR24798|c0_g1_i1	5.028569152	5.208478242	5.176960992	2.710833979	2.18142064	2.407080775
	TR3297|c0_g1_i1	4.573556306	4.311938997	4.187530234	0.626672753	0.623866862	0
	TR39786|c0_g1_i1	2.42867841	3.467149032	2.713475875	0.669933836	0	0.192825404
	TR3074|c0_g1_i1	2.895302621	3.040015679	3.462837601	1.328262029	1.523060062	1.13553487
	TR34843|c0_g1_i1	3.972508859	3.592277624	3.819668183	1.279174108	0.809825996	1.705314441
	TR18788|c1_g1_i1	2.823137919	1.939226578	2.226816765	0.618238656	0	0.588804567
	TR35288|c0_g1_i1	2.386259141	2.49441561	2.749748779	1.250961574	0.420078116	0.727702673
	TR1730|c0_g1_i1	2.180147861	2.168321116	2.187451054	0.429749851	0.158337027	0
	TR21372|c0_g1_i1	2.272918995	2.673782534	2.593114696	0	0.207892852	0
	TR4367|c0_g1_i1	2.472487771	3.22897257	2.639926713	0.389566812	0	0.372952098
	TR20170|c0_g1_i1	2.874993639	2.89452677	2.577972569	0.745022149	0.521050737	0.884402553
	TR14581|c0_g1_i1	3.096261853	3.379759569	3.567423758	1.222186307	0.521050737	0.503857533
	TR31877|c0_g1_i1	2.101986173	2.597173585	2.0590091	0.325386415	0	0
	TR21085|c1_g2_i1	3.489542936	3.545721311	3.278728213	1.709731759	1.275007047	1.578214165
Foundresses	TR11303|c0_g1_i1	2.3305584	1.85239842	2.481040556	0.314986485	0	0
	TR32411|c0_g1_i1	2.442280035	2.505382849	2.491853096	0.799916203	0.195347598	0.768078435
	TR20727|c0_g1_i1	1.966799585	2.624802765	2.615651705	0.564622052	0	0.296310561
	TR21742|c0_g1_i1	2.731183242	2.354170026	2.552377071	0.166072676	0.96790637	0.86155842
	TR5481|c0_g1_i1	2.357270476	2.154777612	2.180784391	0	0	0
	TR42559|c0_g1_i1	5.794233893	6.607211934	4.877302541	3.163337708	3.210856386	3.037557935
	TR30788|c0_g1_i1	1.769771739	1.822118275	1.901880564	0	0.377401431	0.13093087
	TR35950|c0_g2_i1	1.82984956	2.384049807	2.091530593	0.459431619	0.344828497	0.119024103
	TR29719|c0_g1_i1	3.970669625	3.888499736	3.809002775	1.624802765	1.23388806	1.569004927
	TR16534|c2_g2_i1	0.115033243	0	0	2.13060145	3.570098568	0.588804567
	TR40968|c0_g1_i1	4.156072961	4.17990909	4.063589225	2.438559007	1.827819025	2.611408481
	TR35987|c0_g1_i1	2.799501814	3.302465287	2.895302621	1.113700499	0.641546029	1.317304068
	TR3670|c0_g1_i1	3.394376945	3.158337027	2.9053509	0.554834396	0.301002256	0.913798965
	TR19068|c0_g1_i1	1.577247536	2.185549435	2.783247097	0	0.334568276	0.168642036
	TR7633|c0_g1_i1	2.561448342	4.30648099	3.053111336	0	0.549915554	0
	TR16379|c0_g1_i1	4.275379596	3.297778355	4.88757401	6.506605116	8.675921759	7.491965302
	TR15108|c0_g1_i2	0.779890039	0.732052073	0.722466024	2.950468414	2.462314127	2.497229129
	TR34994|c0_g1_i1	2.735738782	2.730531274	2.794103899	0.853596506	0.7031007	0.685267407
	TR13183|c0_g2_i1	4.720606811	4.861409537	4.816855662	2.196921734	1.683471893	2.497229129
	TR39982|c0_g1_i1	3.031748063	2.807148808	3.178714641	0.259423152	0.254594043	0.240008965
	TR10171|c0_g2_i1	3.207424368	3.424922088	3.753925385	0.518031493	0.510961919	0.318461465
	TR31013|c0_g1_i1	3.303196247	3.608809243	3.695548468	1.168642036	0	0.668119125
	TR19467|c0_g1_i1	0	0	0	7.268696379	8.014360873	6.593981153
	TR21239|c0_g1_i1	4.857184806	4.451936399	4.491853096	2.567545448	2.178236585	2.901108243
	TR38435|c0_g1_i1	2.935082523	3.427740246	3.73020518	0.545968369	0.841570637	0.285698126
	TR8782|c0_g2_i1	8.508135456	9.13854572	8.491772944	6.047385701	5.836328134	6.138466365
	TR14612|c0_g2_i1	3.729879012	3.570462931	3.502203284	1.714135594	1.473527177	1.559736524

Online resource 2 - Table II Annotation result from UniRef, pfam domains, ORF and lncRNA probability of all the differentially expressed genes as reported by the Annocript program. Columns descriptions are below the Table. Possible lncRNAs are highlighted

		TranscriptName	TransLength	HSPNameUf	HSPLengthUf	HSPEvalueUf	HITLengthUf	QCoverageUf	HCoverageUf
Larvae	UP G2	TR15541|c0_g1_i1	6561	UniRef90_F1KPN1	4464	8.00E-137	3085	68.03840878	48.58995138
Foundresses	UP G1	TR30788|c0_g1_i1	621	-	-	-	-	-	-
		TR11303|c0_g1_i1	355	-	-	-	-	-	-
		TR34843|c0_g1_i1	969	-	-	-	-	-	-
		TR1730|c0_g1_i1	570	-	-	-	-	-	-
		TR32411|c0_g1_i1	481	-	-	-	-	-	-
		TR8778|c0_g1_i1	508	-	-	-	-	-	-
		TR3670|c0_g1_i1	364	-	-	-	-	-	-
		TR11926|c0_g1_i1	366	-	-	-	-	-	-
		TR24798|c0_g1_i1	636	-	-	-	-	-	-
		TR39786|c0_g1_i1	485	-	-	-	-	-	-
		TR10171|c0_g2_i1	714	-	-	-	-	-	-
		TR14612|c0_g2_i1	580	-	-	-	-	-	-
		TR3297|c0_g1_i1	241	-	-	-	-	-	-
		TR28780|c0_g1_i1	338	-	-	-	-	-	-
		TR38435|c0_g1_i1	573	-	-	-	-	-	-
		TR7633|c0_g1_i1	257	-	-	-	-	-	-
		TR3074|c0_g1_i1	732	-	-	-	-	-	-
		TR14581|c0_g1_i1	262	-	-	-	-	-	-
		TR42559|c0_g1_i1	442	UniRef90_A0A0S2LT52	132	7.00E-09	88	29.86425339	50
		TR20727|c0_g1_i1	357	-	-	-	-	-	-
		TR19068|c0_g1_i1	523	-	-	-	-	-	-
		TR35950|c0_g2_i1	674	-	-	-	-	-	-
		TR35288|c0_g1_i1	718	-	-	-	-	-	-
		TR31013|c0_g1_i1	228	-	-	-	-	-	-
		TR35987|c0_g1_i1	739	-	-	-	-	-	-
		TR20744|c0_g1_i1	437	-	-	-	-	-	-
		TR39982|c0_g1_i1	405	-	-	-	-	-	-
		TR39780|c0_g1_i1	399	-	-	-	-	-	-
		TR13183|c0_g2_i1	261	-	-	-	-	-	-
		TR40968|c0_g1_i1	452	-	-	-	-	-	-
		TR21372|c0_g1_i1	462	-	-	-	-	-	-
		TR18788|c1_g1_i1	521	UniRef90_UPI00021A793B	114	5.00E-12	72	21.88099808	52.77777778
		TR19145|c0_g2_i1	634	UniRef90_B8Q9C2	345	4.00E-54	230	54.41640379	50
		TR31877|c0_g1_i1	343	-	-	-	-	-	-
		TR35077|c0_g1_i1	483	-	-	-	-	-	-
		TR34994|c0_g1_i1	456	-	-	-	-	-	-
		TR20170|c0_g1_i1	805	-	-	-	-	-	-
		TR21239|c0_g1_i1	716	-	-	-	-	-	-
		TR21085|c1_g2_i1	1542	-	-	-	-	-	-
		TR21742|c0_g1_i1	551	-	-	-	-	-	-
		TR4784|c0_g1_i1	469	-	-	-	-	-	-
		TR8782|c0_g2_i1	2205	UniRef90_A0A0S1SA08	939	2.00E-75	379	42.58503401	82.58575198
		TR29719|c0_g1_i1	435	-	-	-	-	-	-
		TR4367|c0_g1_i1	312	-	-	-	-	-	-
		TR5481|c0_g1_i1	370	-	-	-	-	-	-
		TR148|c0_g1_i1	317	-	-	-	-	-	-
	UP G2	TR20635|c0_g2_i1	614	UniRef90_T1JAE3	360	7.00E-08	712	58.63192182	16.85393258
		TR16379|c0_g1_i1	789	UniRef90_Q5MQL3	315	4.00E-23	104	39.92395437	99.03846154
		TR19467|c0_g1_i1	275	-	-	-	-	-	-
		TR15108|c0_g1_i2	926	UniRef90_A0A0L7QTW9	741	2.00E-153	247	80.02159827	100
		TR19467|c3_g1_i5	792	-	-	-	-	-	-
		TR16534|c2_g2_i1	615	-	-	-	-	-	-

DescriptionUf	Taxonomy	BPId	BPDesc	MFId	MFDesc
Replicase polyprotein (Fragment)	Ascaris suum	GO:0006351	transcription, DNA-dependent	GO:0005524]---[GO:0003723]---[GO:0003724]---[GO:0003968]---[GO:0005198	ATP binding]---[RNA binding]---[RNA helicase activity]---[RNA-directed RNA polymerase activity]---[structural molecule activity
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
NADH dehydrogenase subunit 4L	Bombus sylvestris	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
PREDICTED uncharacterized protein LOC100643583	Bombus	-	-	-	-
Cytochrome c oxidase subunit 2	Bombus	GO:0022900	electron transport chain	GO:0005507]---[GO:0004129	copper ion binding]---[cytochrome-c oxidase activity
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
Cytochrome b	Calameuta idolon	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
Uncharacterized protein	Strigamia maritima	GO:0006030	chitin metabolic process	GO:0008061	chitin binding
Defensin-2	Apis	GO:0042742]---[GO:0045087	defense response to bacterium]---[innate immune response	-	-
-	-	-	-	-	-
Dehydrogenase/reductase SDR family member 11	Habropoda laboriosa	-	-	GO:0016491	oxidoreductase activity
-	-	-	-	-	-
-	-	-	-	-	-

domBPId	domBPDesc	domMFId	domMFDesc	LongOrfLength	LongOrfStrand	LongOrfFrame	ProbToBeNonCoding
GO:0006351	transcription, DNA-dependent	GO:0003723]---[GO:0003968]---[GO:0003724	RNA binding]---[RNA-directed RNA polymerase activity]---[RNA helicase activity	2144	+	2	0.00289942
-	-	-	-	70	-	0	0.557141
-	-	-	-	52	+	1	0.969662
-	-	-	-	53	-	2	0.0560917
-	-	-	-	63	+	1	0.336716
-	-	-	-	51	+	0	0.356673
-	-	-	-	105	+	1	0.607173
-	-	-	-	53	+	1	0.898429
-	-	-	-	58	-	1	0.854597
-	-	-	-	122	-	1	0.335559
-	-	-	-	57	+	1	0.233136
-	-	-	-	53	+	0	0.812127
-	-	-	-	47	+	1	0.866475
-	-	-	-	54	+	2	0.930003
-	-	-	-	62	-	2	0.87884
-	-	-	-	63	+	1	0.19075
-	-	-	-	44	-	0	0.952348
-	-	-	-	58	+	0	0.176825
-	-	-	-	39	+	1	0.89834
GO:0042773]---[GO:0055114	ATP synthesis coupled electron transport]---[oxidation-reduction process	GO:0016651	oxidoreductase activity, acting on NADH or NADPH	61	-	0	0.48465
-	-	-	-	45	-	1	0.862489
-	-	-	-	53	+	2	0.201812
-	-	-	-	68	+	1	0.141576
-	-	-	-	57	+	0	0.404806
-	-	-	-	52	+	2	0.971521
-	-	-	-	50	+	2	0.962495
-	-	-	-	49	-	0	0.796269
-	-	-	-	33	+	2	0.779156
-	-	-	-	53	-	2	0.871325
-	-	-	-	51	+	2	0.635036
-	-	-	-	56	+	2	0.209588
-	-	-	-	57	-	0	0.505554
-	-	-	-	94	+	1	0.845779
-	-	GO:0004129]---[GO:0005507	cytochrome-c oxidase activity]---[copper ion binding	54	+	2	0.128523
-	-	-	-	50	-	0	0.856726
-	-	-	-	71	+	1	0.28349
-	-	-	-	77	-	1	0.710905
-	-	-	-	59	+	1	0.878434
-	-	-	-	70	-	0	0.852243
-	-	-	-	59	+	1	0.172866
-	-	-	-	43	-	1	0.924255
-	-	-	-	62	-	0	0.791673
GO:0055114	oxidation-reduction process	-	-	126	-	0	0.0520242
-	-	-	-	45	-	0	0.705617
-	-	-	-	55	+	2	0.95612
-	-	-	-	62	+	1	0.931307
-	-	-	-	68	-	0	0.891739
GO:0019058	viral infectious cycle	-	-	190	-	2	0.264359
GO:0006952	defense response	-	-	112	-	2	0.51932
-	-	-	-	71	-	2	0.0961918
-	-	-	-	250	+	0	0.000839972
-	-	-	-	113	-	0	0.00382777
-	-	-	-	84	-	2	0.877917

lncRNA4Annocript
0
0
1
0
0
0
0
0
0
0
0
0
0
0
0
0
1
0
0
0
0
0
0
1
1
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
1
0
0
0
0
0
0
0
0

TranscriptName: transcript ID as in the assembly fasta file
TransLength: transcript length in nucleotides units
HSPNameUf: hit sequence ID (HSP) with lowest e-value as given from the blastx output against UniRef
HSPLengthUf: corresponding length (in nucleotides) of the HSP
HSPEvalueUf: e-value assigned to the HSP
HITLengthUf: length of the HIT as given from BLASTx output
QCoverageUf: percentage of the query (transcript) covered from the HSP
HCoverageUf: percentage the HSP covers the HIT
DescriptionUf: description of the HSP
Taxonomy: taxonomic group or species corresponding to the UniRef result
BPId: Biological processes IDs corresponding to the lowest e-value result between SwissProt and UniRef (separated by]---[)
BPDesc: Biological processes descriptions as in BPId
MFId: molecular functions IDs with same sorting as in BPId
MFDesc: molecular descriptions with same sorting as in BPId
domBPId: Biological processes IDs corresponding to the results of rpsblast against pfam domains (each pfam may contain more GO terms separated with ';'. Pfam results are in turn separated by]---[)
domBPDesc: Biological processes descriptions with same sorting as in domBPId
domMFId: molecular functions IDs with same sorting as in domBPId
domMFDesc: molecular descriptions with same sorting as in domBPId
LongOrfLength: length of the longest ORF as found from dna2pep
LongOrfStrand: strand of the longest ORF as found from dna2pep
LongOrfFrame: frame of the longest ORF as found from dna2pep
ProbToBeNonCoding: probability of the sequence be non-coding from Portrait
lncRNA4Annocript: Annocript's heuristic analysis. Score 1 means it is a lncRNA

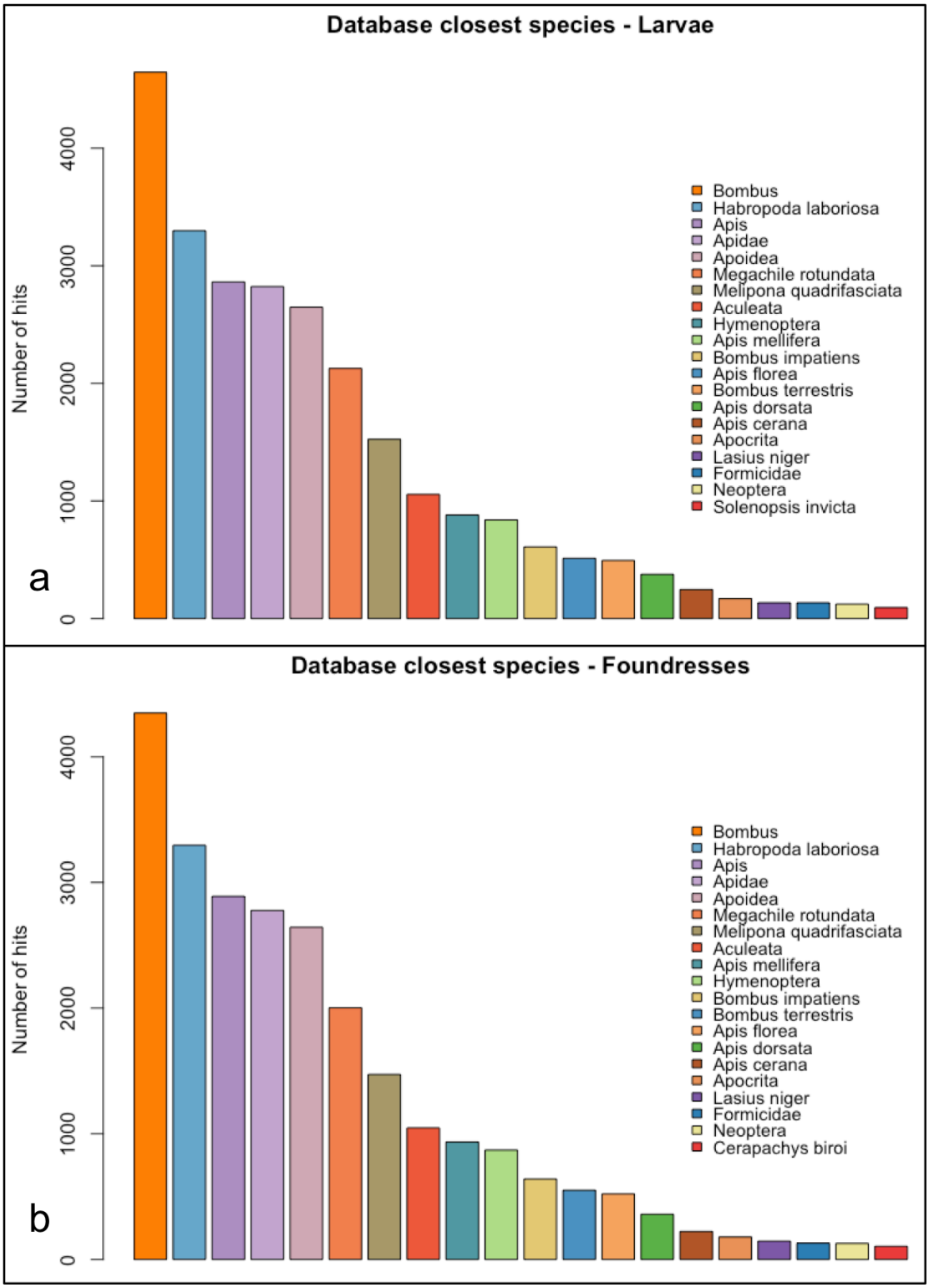


Fig. 1 Blast hits of *Tetrapedia diversipes* transcriptome against UniRef database. The twenty most frequent taxonomic groups are represented. **a-** In the larvae final transcriptome. **b-** In the foundresses final transcriptome

Chapter 3

Gene expression and epigenetic analyses in worker task division of eusocial bees

Natalia de Souza Araujo, Yannick Wurm* and Maria Cristina Arias

* Author(s) not consulted about this first manuscripts version

Revista de interesse: *Scientific Reports* (trabalho ainda não submetido)

Abstract

The existence of individuals who abdicate reproduction to take care of their siblings is one of the most intriguing characteristics of eusociality and a great challenge for evolutionary biology. In eusocial bees, female workers execute most of the colony tasks, except reproduction. Workers are roughly divided into two main specialized subcastes: nurses, are responsible for brood care and colony maintenance; and foragers take care of foraging and nest-guarding. Studies with honey bee workers suggest that specific genes/ networks might be commonly involved in subcaste differentiation. However, analyses in other eusocial bee species are scarce to support this hypothesis. We investigated the transcriptomic and the DNA methylation profile from nurses and foragers of two eusocial species, *B. terrestris* and *T. angustula*. Between nurses and foragers of *B. terrestris* 1,203 genes were differently expressed, while in *T. angustula* 241 genes were found. Mean DNA methylation of these genes was 0.78% in *Bombus terrestris*, and 1.30% in *Tetragonisca angustula*. Comparative analyses revealed that genes involved in worker task division are lineage specific, but the biological processes in which these genes are involved converged in eusocial bees. Results suggest that the specialization observed in worker subcastes occurred later in the evolution of eusocial insects.

Introduction

The specialization of castes in eusocial insects is a notorious example of poliphenism, where multiple phenotypes (morphological and behavioural) can emerge from the same genotype[1,2]. In bees, ants and wasps, queen and worker castes perform distinct functions in the colonies. While queens undertake reproductive duties, workers execute all the other tasks necessary for nest maintenance and growth[1,3]. To fulfil these multiple tasks, workers are usually divided into subcastes, that can exhibit a distinct external morphology[2,4] or not[5]. In eusocial bees, workers are roughly divided in two main subcastes; nurses and foragers[5,6]. Nurses are responsible for comb construction, offspring/ queen care and internal colony maintenance, while foragers perform all the other tasks related to external colony defence and resources provisioning[5,7]. For highly eusocial species, worker subcastes are mainly determined by the bee age: younger bees are nurses and older bees are foragers[8,9]. But in primitively eusocial species, specialization in worker subcastes is less fixed[10,11].

In order to investigate differences between bee worker subcastes, many studies have been developed in the honey bee, *Apis*. Gene expression comparisons have identified a number of expression changes related to the worker behaviour[1,5,8,12]. Which could even be used to identify a network of specific neurogenomic states underlying behaviour in individual bees[13]. Also epigenetic marks, as DNA methylation, were shown to directly affect the worker task division[14]. Some specific genes are more or less methylated according to the worker subcaste, and foragers that are forced to revert to the nursing behaviour restore more than half of the nursing-specific DNA methylation marks lost[15,16].

These studies are great examples of how new molecular tools can be used to solve primordial biological questions. However, to fully comprehend the division of tasks in bee workers it is essential to expand the diversity of studied species[17,18]. In the primitively eusocial bumblebees, a largely studied biological model, molecular aspects related to workers subcastes have been understudied and restricted to a few genes, leaving many open questions[10,19–22] . This is especially due to the difficulty of properly determining the dynamic and less specialized worker subcastes in these bees[19,22]. But even in other highly eusocial bee groups (as the stingless bees), which have an age division of labour, investigations about task division are still scarce. For example, to our best knowledge, no expression or epigenetic studies have been

developed in stingless bees to address this question. Herein we contribute to the understanding of these knowledge gaps through the analyses of expression and global methylation patterns of genes involved in worker task division from two eusocial species, the primitively eusocial bumblebee, *Bombus terrestris*, and the highly eusocial stingless bee, *Tetragonisca angustula*. Using these data we were able to better comprehend the task division behaviour in these species and to look for conserved genes and global patterns of gene expression control common to eusocial bees.

Results

Transcriptome assembly

Final workers transcriptome for *B. terrestris* had 27,987 transcripts of which 431 are potentially lncRNAs and 21,638 (77,3%) had a significant blast hit against the UniRef database. In *T. angustula*, workers final assembly had 33,065 transcripts, wherein 26,623 (80,5%) were annotated and 347 were considered lncRNAs. Additionally, of the 4,415 orthologous genes in the BUSCO Hymenoptera database, 91.9% and 86.2% were found complete in *B. terrestris* and *T. angustula* data, respectively. While 3.9% of the transcripts in *B. terrestris* and 7.9% in *T. angustula* were fragmented. A summary of major quality parameters from both assemblies can be found in S1 - Table I.

Differential expression analyses in *B. terrestris*

Because *B. terrestris* task division in workers is a plastic behaviour[10,19], we performed a principal component analysis of the read counts as an additional verification step to validate our sampling method. In Figure 1 it is shown the two main components in samples from both species, and it is possible to notice a clear division between nurses and foragers. This indicates that our sampling method was efficient to get two distinct groups in bumblebee workers, here considered as nurses and foragers. *B. terrestris* nurses had 436 transcripts highly expressed (S2) and foragers 767 (S3), comprising a total of 1,203 differentially expressed genes (DEG) between the two worker groups. Nurses had 77.3% transcripts annotated and three lncRNAs among its highly expressed genes. While in foragers 72.6% of the over expressed genes had a significant blast hit, and one lncRNA were identified in the

dataset. However, most of the DEG were annotated as "uncharacterized" or "predicted" proteins. Which possibly reflects the absence of Gene Ontology (GO) terms significantly over or underrepresented among the DEG.

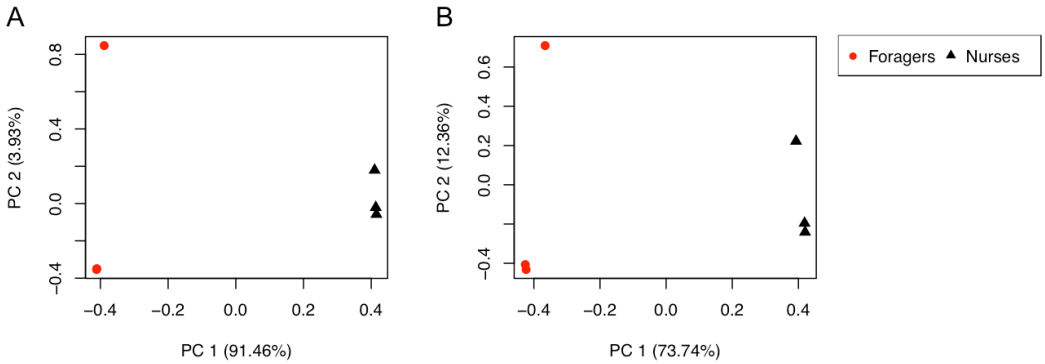


Figure 1 Principal Component Analyses using transcript counts normalized by DESeq2. The two principal components involved in sample differences are shown. A- *B. terrestris* data. B- *T. angustula* data.

Differential expression analyses in *T. angustula*

In workers of *T. angustula* 241 transcripts had different levels of expression between nurses and foragers (S1 – Figure 2). In nurses, 179 genes were highly expressed and of these 157 reported a blast hit against the database (S4). Some of these are genes related to the mitochondrial metabolism (*COI*, *COII*, *COIII*, *NAD-IDH*, *NADH5*, *CytB* and *28S ribosomal protein S30*), DNA replication and integration (*DNA2-helicase*, *transposon TNT 1-94* and *Tc1 transposase*), lipid metabolism (*lipase 3*) and epigenetic alterations (*histone H3*, *histoneH2B*, and a predicted protein that have a methyltransferase activity according to domain analysis). Foragers had 62 transcripts reported as highly expressed (S1 – Figure 2), of which 59 were annotated against the database (S5). However, most of the reported blast hits were against "predicted" or "uncharacterized" proteins. Enrichment analyses, after GOSlim, revealed 53 GO terms as differentially represented in the tested set of DEG when compared to the whole transcriptome (complete data in S1 - Table II). Most of the enriched GO terms were overrepresented and all the enriched biological process (BP) reported can be seen in Table I.

Table I Biological Processes terms, among *T. angustula* DEG, significantly enriched in a two-sided Fisher's exact test (FDR < 0.05). All GO processes were over represented in the tested gene set of DEG when compared to the entire transcriptome annotation.

GO ID	GO Term	FDR
GO:0045333	Cellular respiration	0.0029
GO:0022900	Electron transport chain	0.0041
GO:0055114	Oxidation-reduction process	0.0041
GO:0015980	Energy derivation by oxidation of organic compounds	0.0041
GO:0006091	Generation of precursor metabolites and energy	0.0041
GO:0005975	Carbohydrate metabolic process	0.0078
GO:0009060	Aerobic respiration	0.0086
GO:0022904	Respiratory electron transport chain	0.0092
GO:0006119	Oxidative phosphorylation	0.0134
GO:0006665	Sphingolipid metabolic process	0.0194

Comparative analyses

When comparing the annotated DEG from *T. angustula* and *B. terrestris* workers, we have found that 10 and 3 unique genes were commonly highly expressed in nurses and foragers respectively. However, simulations have shown that these results are not different from what is expected by chance (p-value 0.99). Suggesting that different genes drive the division of tasks in different eusocial bees. Then we asked if the same BP could be involved in the worker behaviour of both species instead. These comparative analyses were performed in two ways; first, we searched for common GO terms only among the highly expressed genes in one of the two workers group, i.e. common BP of genes highly expressed in one subcaste (nurses or foragers) of *T. angustula* and *B. terrestris*; secondly, we compared all differentially expressed BP in both species, regardless of being over or under expressed in one of the subcastes.

For the first analysis, we considered, as common BP, the ones that commonly appeared in the same worker subcaste of both species and that were at least 2 times more frequent in this subcaste than in the other. We have found that in nurses of *T. angustula* and nurses of *B. terrestris* six biological processes were commonly more frequent (Figure 2), which is significantly more than what is due by chance (p-value = 0). However, when foragers were compared no significant results were found.

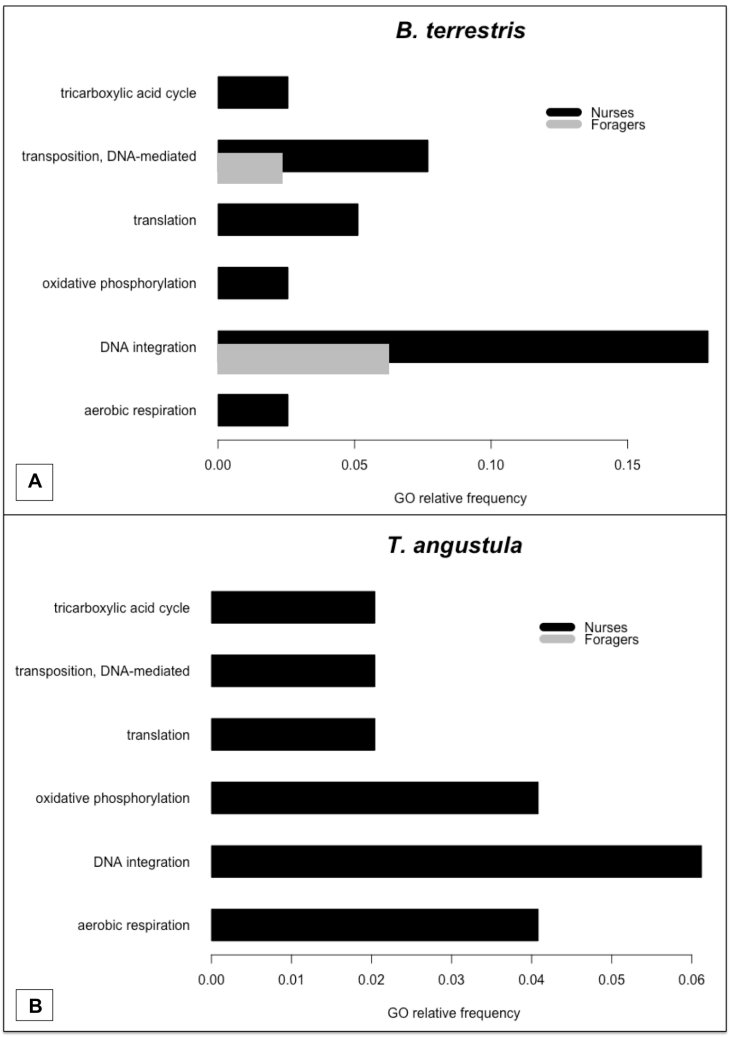


Figure 2 Gene Ontology Biological Processes that were at least two times more frequent in nurses highly expressed genes than in foragers highly expressed genes. When no grey bars appear it means no gene was reported as involved in this specific BP in foragers.

The second round of comparisons have shown that among all of the 29,418 BP registered[23], 15 terms are commonly find in the DEG of both species, regardless of their expression pattern in nurses or foragers. Being them: DNA integration; aerobic respiration; carbohydrate metabolic process; cellular amino acid metabolic process; intracellular protein transport; intracellular signal transduction; lipid metabolic process; nucleocytoplasmic transport; oxidative phosphorylation; regulation of transcription [DNA-dependent]; small GTPase mediated signal transduction; translation; transposition [DNA-mediated]; tricarboxylic acid cycle; and vesicle-mediated transport. This is significantly more than what is expected by chance (p-value = 0). Moreover, after semantic similarity clustering, these common processes clustered with most of the species-specific terms (S1 – Figure 3). Additional

comparative analyses, but using only orthologous data, follow in supplementary results (S1).

We also compared the DEG in *B. terrestris* and *T. angustula* with previous findings in *Apis* workers. This comparison is not straightforward because different methodologies of sampling, expression estimation and analyses pipeline were used in previous studies, so the datasets are not truly comparable. However, we focused in some specific genes/ molecular pathways largely discussed in literature. These comparisons will be detailed and discussed further in the text.

DNA methylation in the workers genes

Whole bisulfite sequencing (WBS) from *B. terrestris* and *T. angustula* nurses were used for DNA methylation analyses in workers transcriptome. In general, distinct DNA contexts are involved in global methylation pattern of workers transcripts in each species; in *B. terrestris* CG methylation contributes more to the general methylation profile and in *T. angustula* CHH methylation are the major contributor (Figure 3). Furthermore, transcripts are less methylated in nurses of *B. terrestris* (mean methylation 0.66%) than in *T. angustula* nurses (mean methylation 1.24%). In both species, DEG were more methylated than the overall mean (S1 – Table III). However, while in *B. terrestris* this increase was mostly due to higher levels of methylation in genes over expressed in nurses (mean methylation level of the highly expressed transcripts in *B. terrestris* nurses was 43.93% higher than the mean methylation level of the whole transcriptome), in *T. angustula* the genes highly expressed in foragers were the ones showing higher levels of methylation (mean methylation of 1.57%). Also, as can be seen in Figure 3, the methylated sites are variable between genes highly expressed in each worker subcaste.

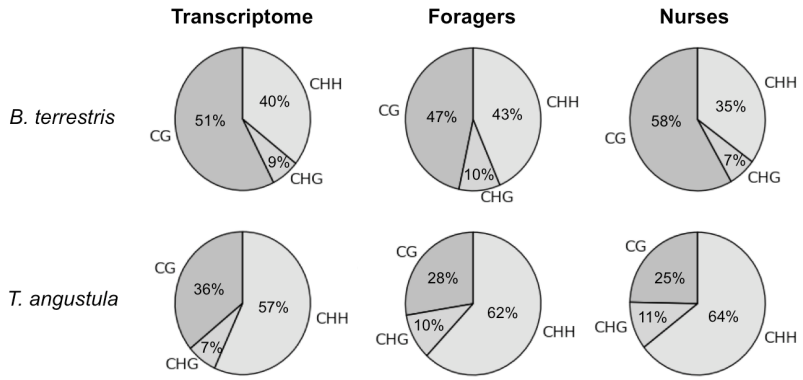


Figure 3 Contribution (in percentage) of each methylation type in the general methylation profile of *B. terrestris* and *T. angustula*. Transcriptome – refers to all the transcripts assembled for workers; Foragers – refers to transcripts highly expressed in foragers; Nurses – refers to transcripts highly expressed in nurses. DNA methylation data are from whole bisulfite sequencing of nurses genomic DNA.

Discussion

Transcriptomic analyses presented here revealed some unique molecular patterns involved in the worker behaviour of *B. terrestris* and *T. angustula*. For example, genes related to growth and circadian rhythm, were only over expressed among foragers of *B. terrestris*. Suggesting that rhythm mighty be important in labour division of bumblebees, as shown for honeybees[24,25], but not on the stingless bee. On the other hand, nurses of *T. angustula* had highly expressed genes involved in innate immune response when compared to foragers, nonetheless similar genes were not highly expressed in any of *B. terrestris* subcastes. When results were compared to genes previously reported in literature, the same pattern of lineage specific gene expression levels were observed (Table II).

Table II Summary of main genes/ molecular pathways discussed in literature as involved in honey bee worker task division and a brief comparison to present findings in *B. terrestris* and *T. angustula*.

***vitellogenin* (*vg*)**	This yolk precursor gene is related to egg production in many insects[26]. In honey bees it interacts with juvenile hormone (JH) in a double repressor network, and its expression is reduced in foragers[5,6,26]. But for bumblebees, this double repressor network apparently does no exist[21]. In our *B. terrestris* data, two genes highly expressed in foragers have *vg* transcription factors domains. Which might be related to the primitively eusocial behaviour in these bees, since workers may dispute reproductive status with queens further in colony cycle. In nurses of *T. angustula*, a vg receptor is highly expressed, as in honey bees. Nevertheless, it is worth noticing that stingless bees workers usually produce trophic eggs[27], so vg might be involved in this process or even have alternative unknown roles

	in stingless bee, as suggested in[28].
juvenile hormone (JH)	These hormones are important regulators in honey bee maturation; JH have higher levels in foragers than nurses[5,6]. Only one DEG was considered as possibly related to JH pathways, the gene predicted as "takeout-like". This gene has a JH biding domain and is highly expressed in *B. terrestris* foragers. In *Drosophila*, the takeout protein is implicated in circadian control of feeding behaviour[29] and in honey bees the JH is connected with foraging onset[6]. Thus it is possible that JH could be involved in *B. terrestris* foraging.
***foraging* (*for*)**	Although this gene have been reported as highly expressed in foragers of honey bees[30] and bumblebees[22], there are controversial results in literature. In honey bees, this gene expression was not among the best predictors of work behavioural transition[5,8] and in bumblebees, its expression was higher in nurses than foragers in one study[19]. Herein, this gene was not differentially expressed in the studied species.
***period* (*per*)**	This gene is related to rhythm in bees and has been reported as over expressed in honey bees foragers[24,25]. This specific gene does not appear among the DEG datasets. However, *B. terrestris* foragers have other highly expressed rhythm genes (*protein quiver* or *sleepless*) that are related to sleep, rhythmic process, and regulation of circadian sleep/wake cycle. This suggests that in the primitively eusocial *B. terrestris*, and in *Apis*, rhythm genes are more relevant to nurse/forager behaviours than in the highly eusocial *T. angustula*.
Insulin/ Insulin-like signalling (IIS)	Genes involved in this pathway are important regulators of metabolism and feeding-related behaviour in bee workers[31,32]. In both species studied, there are genes related to insulin metabolism (genes containing insulin domains, transcription factor and regulators). This indicates that the insulin pathway is commonly important to worker subcaste specialization in different eusocial bee species.
Energetic metabolism	In general, genes related to energetic metabolism are expected to be involved in bee worker behaviour because feeding circuits are basal pathways to different bee activities[33]. Indeed, many genes related to energetic metabolism are DEG between nurses and foragers of both species, being some of the BP terms commonly found in both species related to this pathway. Specific examples of genes involved in energetic pathways (besides JH and IIS) studied in honey bees are *malvolio* and *royal jelly*[34,35]. The first was not differentially expressed in our data; the second were related to some DEG in *B. terrestris*. In nurses of this species, two highly expressed genes were predicted as protein yellow genes (which have a major royal jelly protein family domain), and in foragers other two over expressed genes had major royal jelly protein family domains.
Transcription factors (TF)	Different TF are believed to be involved in the dynamic changes related to behaviour in eusocial bees[32]. Corroborating this hypothesis, different TF genes, were identified among the DEG of both species. Specifically the *ultraspiracle* (*usp*) TF, known to participate in honey bee worker task division transition because of its interaction with JH[36], was not among them.
DNA methylation/ epigenetic modifications	DNA methylation is known to be involved in nurse to forager transition in honey bees[15,16]. In the two species investigated in the present study genes possibly related to epigenetic marks were also differentially expressed. In *T. angustula*, histone genes (H3 and H2B) and a methyltransferase were DEG; and in *B. terrestris* the DEG were histone H3-K4 demethylation and lncRNAs. All these genes were highly expressed in nurses, except for one lncRNA, that was over expressed in *B. terrestris* foragers.

The involvement of lineage specific genes in task division is not completely unexpected since previous studies have reported that, even within a genus, genes

involved in the same process can vary[12]. An explicit implication of this observation is that function and networks of well-studied genes, involved in worker behaviour of honey bees, may not be directly extrapolated for other species. One example is the JH network, which have been largely studied in *Apis* and its interaction with other genes, such as *vg* and *usp*, are known to affect the worker behaviour[5,36]. Honey bee nurses have higher levels of vg and lower levels of JH when compared to foragers. When nurses become foragers their levels of JH increases, which in turn, represses the *vg* expression in a double repressor network[6,36]. Previous studies have demonstrated that for bumblebees this network apparently is not regulated in the same manner[21]. Our data corroborates this assumption, indicating that the vg/JH network described for *Apis* is not similarly regulated in worker subcastes of the bumblebee. As described in Table II, in *B. terrestris*, genes related to JH and vg are both highly expressed. For *T. angustula*, we found evidences supporting the importance of *vg* in nursing behaviour but no genes related to JH were highly expressed in foragers (Table II). Also for stingless bees, evidences suggests that *vg* is involved in different regulatory mechanisms[28]. Thus it is likely that the double repressor network between vg/JH is not functional in *T. angustula* as well.

Changes in the JH network have been reported even for honey bees under selective pressure[37]. Honey bee strains selected for low pollen hoarding develop the pollen hoarding syndrome, in which a number of phenotypic responses are affected[37]. One of these responses is the feedback negative control of vg and JH; low pollen hoarding strains do not increase the JH expression nor forage precociously after down regulation of *vg*, as normally occur in wild honey bees. This study illustrates how specific gene interactions can rapidly change under selection, explaining why specific genes seem to have distinct roles in the worker behaviour of different species.

Nevertheless, because a significant number of BP were common between *B. terrestris* and *T. angustula* DEG, the toolkit hypothesis still holds for the existence of specific networks involved in behaviour. This scenario suggests that, labour division in workers is a derived behaviour that evolved independently in each species, due to selective pressures to maintain eusociality. During this process, the same BP were modified. In this perspective, it is also interesting to notice that BP of highly expressed genes in nurses are more conserved than in foragers. Since foragers are

more exposed to specific environmental challenges in each lineage, the greater convergency between nurses might be related to behavioural differences in subcastes.

This mosaic pattern of unique genetic features involved in common molecular mechanisms is also observed in the epigenetic mechanisms underneath worker behaviour. The involvement of DNA methylation and other epigenetic factors in subcaste differentiation of both species is supported by results from transcriptomic and WBS data. Suggesting that as in *Apis* these processes are crucial in the expression changes necessary for task division[15,16]. Genes involved in epigenetic alterations were found among the DEG of *T. angustula* and *B. terrestris*, and WBS indicates that, in nurses, genes related to the worker behaviour are more methylated (S1 - Table III). However, a closer investigation in these data revealed that the specific epigenetic mechanism affecting the behaviour is not conserved.

In *B. terrestris*, DEG involved in epigenetic mechanisms are histone modifiers and lncRNAs, while in *T. angustula* they are histone genes and a methyltransferase. Combining these findings to the greater number of lncRNAs reported for the complete transcriptome of the bumblebee and the higher level of overall methylation in the complete transcriptome of the stingless bee, it becomes clear that the type of epigenetic mechanism used in both species varies. Apparently, in *B. terrestris* lncRNAs are especially relevant while in *T. angustula* DNA methylation plays a greater role. Additionally, WBS shows that genes of both species differ in the DNA context and in the amount of methylation. Genes highly expressed in nurses of *B. terrestris* were more methylated than genes over expressed in foragers, but the opposite was true for *T. angustula*. In bees high levels of methylation in genes are related to an increase in gene expression, i.e. genes with more methylation have higher expression levels[14]. Considering that all DNA methylation data are from nurses, these results exemplify a common molecular mechanism affecting gene expression differently.

The hypothesis that labour division evolved independently in different eusocial species have been raised by previous studies[38,39]. Genes involved in caste differences (queen and workers) are more conserved among different species[39], while genes that influence worker traits are under positive selection pressures in eusocial bees[38]. These evidences suggest that genes related to workers phenotypes are major

factors in species adaptation, while caste genes are more basal to sociality. Our transcriptomic and WBS results are consistent with these assumptions.

Through the analyses of the transcriptomic and methylation patterns of worker subcastes from two different eusocial bee groups, in a highly comparative way, we provide an important dataset for the study of social behaviour evolution. Main findings support the hypothesis that worker task division evolved later in bee sociality through adaptations of specific genetic mechanisms in each lineage. Distinct evolutionary constraints selected unique genetic and epigenetic mechanisms that resulted in similar phonotypical behaviours. However, despite these independent evolutionary backgrounds, common biological processes were affected in different lineages, especially the ones associated with metabolism and gene expression control. This contributed for a mosaic pattern in worker task division, where unique and shared features can be found.

Material and Methods

Sample collection and sequencing

Workers were classified according to observations and age control. Foragers were sampled while foraging, and nurses were defined based on two methods: 1- *T. angustula* nurses were defined by age; 2- *B. terrestris* nurses were selected based on behavioural observations (sampling methods are detailed in S1). All individuals were sampled between 10h-12h for both species and whole worker bodies were used for RNA and DNA extraction. For RNA-Seq, total RNA were extracted from workers using Qiagen® extraction kit (RNeasy Mini Kits). RNA quality and quantification were verified using Bionalyzer® and the Nanodrop®, respectively. Three colonies per species were used for RNA extraction, each considered as a sample replicate. Six *T. angustula* workers, from the same colony and subcaste, were pooled as one sample. *B. terrestris* samples were a pool of three workers per subcaste. Samples were posteriorly used for RNA sequencing in the Illumina® HiSeq 2000. For whole bisulfite sequencing, total DNA from one nurse bee of each species was extracted using a phenol-chloroform protocol[40]. WBS were performed following the protocol described in[41] using the Illumina® NextSeq500.

Transcriptome assembly and differential expression analyses

Reads were cleaned and normalized as detailed in S1. Transcriptome assembly were performed differently for each species. For *B. terrestris*, its genome[42] was used as reference for the assembly of the cleaned reads. We performed two distinct assemblies. First, using HISAT2[43] (v2-2.0.3) and StringTie[44] (v1.2.2) a reference assembly was obtained. Secondly, the Trinity[45] (v2.1.1) program was used to perform a reference guided *de novo* assembly. The two resulting assemblies were merged using CD-Hit[46] (v4.6), Corset[47] (v1.05) and Lace[48] (v0.80). This combined approach was necessary since results of using only the reference assembly performed poorly in the quality tests. A similar approach was employed for *T. angustula* transcriptome assembly, but with some adaptations. The genome of this species is not available on databases, so we used the genome from another stingless bee, *Melipona quadrifasciata*[49], and only the Trinity program was used. Two assemblies were made with the normalized reads, a reference guided *de novo* assembly and a complete *de novo* assembly. Subsequently, assemblies were merged as in *B. terrestris*. Complete description of these methods and the parameters used are in S1. Differential expression analyses were performed adapting scripts available in the Trinity package. Bowtie2[50] (v2.2.5), RSEM[51] (v1.2.22) and DESeq2[52] (p-value < 1e-3) were used to identify differentially expressed genes. We focused our discussion on genes highly expressed in each subcaste, but it is important to keep in mind that genes over expressed in one group have reduced levels of expression in the other. And this reduced expression is also biologically relevant. To identify the enriched GO terms, a two-tailed Fisher's exact test was performed using Blast2GO®. Comparisons between species results followed the workflow described in S1 – Figure 1.

DNA methylation analyses

Cleaning and adapter trimming of the bisulfite converted reads were performed using the Trim Galore[53] (v 0.4.3) wrapper script with default parameters. Trim Galore uses Cutadapt[54] (v 1.13) for cleaning. Complete transcriptome assemblies were used as reference for each species, so the DNA methylation of coding regions could be analysed, since these sequences are the main methylation targets in bees and other Hymenoptera[14]. PCR bias filtering, alignment of the cleaned reads and methylation call were performed using the BS-Seeker2[55] (v 2.1.0), because this

program allows the use of Bowtie2 in local alignment mode. CGmapTools[56] (v 0.0.1) was used to filter low coverage methylated sites (minimum of 10x) and statistics.

Acknowledgements

Authors would like to thank Dr. Isabel Alves-dos-Santos and MSc. Sheina Koffler from the Laboratório de Abelhas (Universidade of São Paulo) for the support during *T. angustula* sampling, and to Dr. Lars Chittka and Dr. Stephan Wolf from Bee Sensory and Behavioural Ecology Lab (Queen Mary University of London) for the support during *B. terrestris* sampling. We also would like to thank Dr Bob Schmitz (University of Georgia) for whole bisulfite sequencing and library preparation during this study and to Susy Coelho for technical assistance.

Additional Information

This study was financed by FAPESP (São Paulo Research Foundation, process numbers 2013/12530-4 and 2012/18531-0) and developed at the Research Centre on Biodiversity and Computing (BioComp) of the Universidade de São Paulo (USP), supported by the USP Provost's Office for Research. Part of the bioinformatic analyses was performed at the cloud computing service from USP and from QMUL.

Authors declare they have no competing financial interests.

References

1. Grozinger, C. M., Fan, Y., Hoover, S. E. R. & Winston, M. L. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (Apis mellifera). *Mol. Ecol.* **16,** 4837–4848 (2007).
2. Grüter, C. *et al.* Repeated evolution of soldier sub-castes suggests parasitism drives social complexity in stingless bees. *Nat. Commun.* **8,** e4 (2017).
3. Robinson, G. E., Fahrbach, S. E. & Winston, M. L. W. Insect societies and the molecular biology of social behavior. *Bioessays* **19,** 1099–1108 (1997).
4. Fjerdingstad, E. J. & Ross, H. C. The Evolution of Worker Caste Diversity in Social Insects. *Am. Nat.* **167,** 390–400 (2006).
5. Whitfield, C. W. *et al.* Genomic dissection of behavioral maturation in the honey bee. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 16068–16075 (2006).
6. Guidugli, K. R. *et al.* Vitellogenin regulates hormonal dynamics in the worker caste of a eusocial insect. *FEBS Lett.* **579,** 4961–4965 (2005).
7. Engels, W. & Imperatriz-Fonseca, V. L. in *Social Insects: An Evolutionary Approach to Castes and Reproduction* (ed. Engels, P. D. W.) 167–230 (Springer Berlin Heidelberg, 1990).
8. Whitfield, C. W., Cziko, A.-M. & Robinson, G. E. Gene Expression Profiles in the Brain Predict Behavior in Individual. *Science.* **296,** 296–299 (2003).
9. Hrncir, M., Jarau, S. & Barth, F. G. Stingless bees (Meliponini): senses and behavior. *J. Comp. Physiol. A Neuroethol. Sensory, Neural, Behav. Physiol.* 1–5 (2016). doi:10.1007/s00359-016-1117-9
10. Goulson, D. *et al.* Can alloethism in workers of the bumblebee, *Bombus terrestris*, be explained in terms of foraging efficiency? *Anim. Behav.* **64,** 123–130 (2002).
11. Russell, A. L., Morrison, S. J., Moschonas, E. H. & Papaj, D. R. Patterns of pollen and nectar foraging specialization by bumblebees over multiple timescales using RFID. *Sci. Rep.* **7,** 42448 (2017).
12. Sen Sarma, M., Whitfield, C. W. & Robinson, G. E. Species differences in brain gene expression profiles associated with adult behavioral maturation in honey bees. *BMC Genomics* **8,** (2007).
13. Chandrasekaran, S. *et al.* Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 18020–18025 (2011).
14. Yan, H. *et al.* DNA Methylation in Social Insects: How Epigenetics Can Control Behavior and Longevity. *Annu. Rev. Entomol.* **60,** 435–452 (2015).
15. Lockett, G. A., Kucharski, R. & Maleszka, R. DNA methylation changes elicited by social stimuli in the brains of worker honey bees. *Genes, Brain Behav.* **11,** 235–242 (2012).
16. Herb, B. R. *et al.* Reversible switching between epigenetic states in honeybee behavioral subcastes. *Nat. Neurosci.* **15,** 1371–1373 (2012).
17. Kapheim, K. M. Genomic sources of phenotypic novelty in the evolution of eusociality in insects. *Curr. Opin. Insect Sci.* **13,** 24–32 (2016).
18. Toth, A. L. & Rehan, S. M. Molecular Evolution in Insect Societies: An Eco-Evo-Devo Synthesis. *Annu. Rev. Entomol.* **62,** 419–442 (2017).
19. Kodaira, Y., Ohtsuki, H., Yokoyama, J. & Kawata, M. Size-dependent foraging gene expression and behavioral caste differentiation in Bombus ignitus. *BMC Res. Notes* **2,** (2009).
20. Goulson, D. *Bumblebees : behaviour, ecology, and conservation. Bumblebees their behaviour and ecology* (Oxford University Press, 2010).
21. Amsalem, E., Malka, O., Grozinger, C. & Hefetz, A. Exploring the role of juvenile hormone and vitellogenin in reproduction and social behavior in bumble bees. *BMC Evol. Biol.* **14,** 1–13 (2014).
22. Tobback, J., Mommaerts, V., Vandersmissen, H. P., Smagghe, G. & Huybrechts, R.

Age- and task-dependent *foraging* gene expression in the bumblebee *Bombus terrestris*. *Arch. Insect Biochem. Physiol.* **76,** 30–42 (2011).
23. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25,** 3045–3046 (2009).
24. Toma, D. P., Bloch, G., Moore, D. & Robinson, G. E. Changes in period mRNA levels in the brain and division of labor in honey bee colonies. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 6914–6919 (2000).
25. Bloch, G., Rubinstein, C. D. & Robinson, G. E. *period* expression in the honey bee brain is developmentally regulated and not affected by light, flight experience, or colony type. *Insect Biochem. Mol. Biol.* **34,** 879–891 (2004).
26. Nelson, C. M., Ihle, K. E., Fondrk, M. K., Page, R. E. & Amdam, G. V. The gene *vitellogenin* has multiple coordinating effects on social organization. *PLoS Biol.* **5,** 0673–0677 (2007).
27. Koedam, D. & Tienen, P. G. M. Van. The regulation of worker-oviposition in the stingless bee. *Insectes Soc.* **44,** 229–244 (1997).
28. Dallacqua, R. P., Simões, Z. L. P. & Bitondi, M. M. G. Vitellogenin gene expression in stingless bee workers differing in egg-laying behavior. *Insectes Soc.* **54,** 70–76 (2007).
29. So, W. V *et al. takeout*, a novel *Drosophila* gene under circadian clock transcriptional regulation. *Mol. Cell. Biol.* **20,** 6935–6944 (2000).
30. Ben-Shahar, Y., Robichon, A., Sokolowski, M. B. & Robinson, G. E. Influence of Gene Action Across Different Time Scales on Behavior. *Science (80-.).* **296,** 741–744 (2002).
31. Ament, S. A., Corona, M., Pollock, H. S. & Robinson, G. E. Insulin signaling is involved in the regulation of worker division of labor in honey bee colonies. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4226–4231 (2008).
32. Dolezal, A. G. & Toth, A. L. Honey bee sociogenomics: A genome-scale perspective on bee social behavior and health. *Apidologie* **45,** 375–395 (2014).
33. Fischer, E. K. & O'Connell, L. A. Modification of feeding circuits in the evolution of social behavior. *J. Exp. Biol.* **220,** 92–102 (2017).
34. Ben-Shahar, Y., Dudek, N. L. & Robinson, G. E. Phenotypic deconstruction reveals involvement of manganese transporter *malvolio* in honey bee division of labor. *J Exp Biol* **207,** 3281–3288 (2004).
35. Buttstedt, A., Moritz, R. F. A. & Erler, S. Origin and function of the major royal jelly proteins of the honeybee (*Apis mellifera*) as members of the *yellow* gene family. *Biol. Rev.* **89,** 255–269 (2014).
36. Ament, S. A. *et al.* The Transcription Factor *Ultraspiracle* Influences Honey Bee Social Behavior and Behavior-Related Gene Expression. *PLoS Genet.* **8,** e1002596 (2012).
37. Page, R. E., Rueppell, O. & Amdam, G. V. Genetics of Reproduction and Regulation of Honeybee (*Apis mellifera L.*) Social Behavior. *Annu. Rev. Genet.* **46,** 97–119 (2012).
38. Harpur, B. a *et al.* Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 2614–2619 (2014).
39. Jones, B. M., Kingwell, C. J., Wcislo, W. T., Robinson, G. E. & Jones, B. M. Caste-biased gene expression in a facultatively eusocial bee suggests a role for genetic accommodation in the evolution of eusociality. (2017). doi:10.1098/rspb.2016.2228
40. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162,** 156–159 (1987).
41. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10,** 475–483 (2015).

42. Sadd, B., Barribeau, S. & Bloch, G. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol*. **16,** 1–32 (2015).
43. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat*. *Methods* **12,** 357–360 (2015).
44. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat*. *Biotechnol*. **33,** 290–295 (2015).
45. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat*. *Biotechnol*. **29,** 644–652 (2013).
46. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* **26,** 680–682 (2010).
47. Davidson, N. M. & Oshlack, A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol*. **15,** 410 (2014).
48. Hawkins, A. D. K., Oshlack, A. & Davidson, N. M. SuperTranscript : a reference for analysis and visualization of the transcriptome. *bioRxiv* (2016). doi:10.1101/077750
49. Kapheim, K. M. *et al*. Genomic signatures of evolutionary transitions from solitary to group living. *Science (80)*. **348,** 1139–1143 (2015).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat*. *Methods* **9,** 357–359 (2012).
51. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 1–16 (2011).
52. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15,** 1–34 (2014).
53. Krueger, F. Trim Galore. (2012). Available at https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
54. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
55. Guo, W. *et al*. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* **14,** 1–8(2013).
56. Guo, W. & Zhu, P. CG-maptools: Command-line Toolset for Bisulfite Sequencing Data Analysis. (2017). Available at https://cgmaptools.github.io

Attachments

S1 - Supplementary File 1

Supplementary Results

Differential expression analyses based on orthologous

To investigate if the lack of similarity among species were due to differences in transcriptome assembly or annotation, we performed differential expression analyses using only the conserved orthologous protein genes. For that, 2,940 unique orthologous were identified using the OMA[1]. Differential expression analyses with this data set identified 43 (31 highly expressed in foragers and 12 highly expressed in nurses) and 11 (one highly expressed in foragers and 10 highly expressed in nurses) DEG in *B. terrestris* and *T. angustula*, respectively (S6). Most of these genes have already being reported in the total analyses. After comparisons, neither the over expressed genes nor their BP terms were common between workers with the same task in the colony from both species. The reduced number of DEG and the complete absence of similarities between both species DEG data suggest that genomic changes involved in bee workers labour division have many species-specific genes and isoforms. Thus conserved orthologous are not the most representative genes to study worker subcastes.

Supplementary Material and Methods

Sample collection and sequencing

Three *B. terrestris* colonies were obtained from commercial suppliers (Biobest®) and kept in lab condition at Queen Mary University of London (England). Colonies were marked and housed in wood boxes attached to foraging arenas. After 16 days of adaptation all recently born workers in the colony received an individual number tag used for later identification. Bumblebee workers usually do not forage right after emergency[2], thus we waited for additional five days before observing their behaviour. Thus all collected bees were in between 6-22 days older. After this period colonies were observed for one day during all its active foraging period (6 hours uninterruptedly). All the bees detected in the foraging areas were considered as foragers and the bees that stayed inside of the nest for the whole day were considered nurses. In the following day, between 10h – 12h, all nurses and foragers were

collected and immediately frozen in liquid nitrogen. Foragers were sampled first, while collecting nectar in the foraging arena. Nurses were posteriorly collected inside of the colonies.

Three *T. angustula* colonies, regularly kept at the Laboratório de Abelhas (University of São Paulo – Brazil), were used for sample collection. All workers were collected between 10h – 12h. Foragers were collected while leaving and returning to the colonies from foraging trips. To avoid the guard subcaste[3], workers standing in front of the colony entrance were avoided. To collect nurses, brood cells (close to emergency) were removed from the colonies and transferred to an incubator with controlled temperature and humidity. Upon emergency female workers were marked with specific colours using a water based ink and immediately returned to the colony. Ten to twelve days after their emergency and reintroduction, colonies were opened and marked individuals were sampled. During this age worker bees from *T. angustula* present nursing behaviour[4]. Some of the foragers were collected before nurse sampling and others after this period, but no foragers were sampled while nurses were been marked and collected to avoid grater colony disturbance. Nurses from different colonies were collected in different days.

RNA-Seq sequencing generated about 50 million paired reads (100bp) per sample (colony replicate). *B. terrestris* workers were sequenced at the Genome Center at Queen Mary University of London, and *T. angustula* samples were sequenced at LACTAD (Unicamp). Whole bisulfite sequencing (WBS) returned 60-70 million single reads (150 bp) per sample, sequencing were at University of Georgia.

Transcriptome assembly and differential expression analyses

Reads quality assessment was performed with the FastQC program[5] (v0.11.2) before and after cleaning. The FASTX Toolkit[6] (v0.0.14) was used to trim the first 14 bp of all reads because of the initial GC bias[7]. Low quality bases (phred score below 30) and small reads (less than 31 bp) were removed using SeqyClean[8] (v1.9.3). To increase the *T. angustula* transcriptome assembly efficiency the cleaned reads were digitally normalized[9] (20x coverage) previously. Programs used in transcriptome assembly were run using default parameters. CD-Hit was used to merge transcripts with more than 95% similarity and Corset was set to keep transcripts with a minimum of 50x coverage.

Quality parameters of the final assemblies were analysed using QUAST[10] (v4.0), BUSCO[11] (v2) and Qualimap[12] (v2.2). Transcripts were then annotated with Annocript[13] (v 1.2) using the UniProt Reference Clusters (UniRef) database[14] (June 2016 version). Transcripts with significant blast hits (e-value < 1e-5) against possible contaminants (plants, fungus, mites and bacteria) in UniRef were removed from the final dataset. This annotation pipeline was used for both transcriptome assemblies so the dataset could be comparable, the use of a different annotation approach for each transcriptome would hinder comparisons.

During analyses of the sequenced and cleaned reads we identified a possible batch effect in samples from *T. angustula*, one nurse and one forager replicate were sequenced in different lanes and it seemed to affect sample correlation. This effect was corrected during differential expression analyses following the suggested protocol in DESeq2 documentation. No batch effect was identified in *B. terrestris* samples. *T. angustula* and *B. terrestris* samples were independently analysed (Figure 1).

Comparative analyses

Comparative analyses followed the workflow detailed in Figure 1. First, each species transcriptome were independently assembled. Second, cleaned reads were realigned to each assembled transcriptome. Third, transcripts with different levels of expression were identified in each species. Forth, DEG were compared. DEG were compared using custom python and bash scripts. Statistical tests of significance for comparisons were based on simulations using R scripts, p-value smaller than 0.05 were considered significant.

For analyses of orthology the OMA program[1] (v1.0.6) was used to identify orthologous among all transcripts assembled for both species. Differential expression analyses of the orthologous were performed as described for the complete transcriptome but using the reduced set of 2,940 unique orthologous as reference.

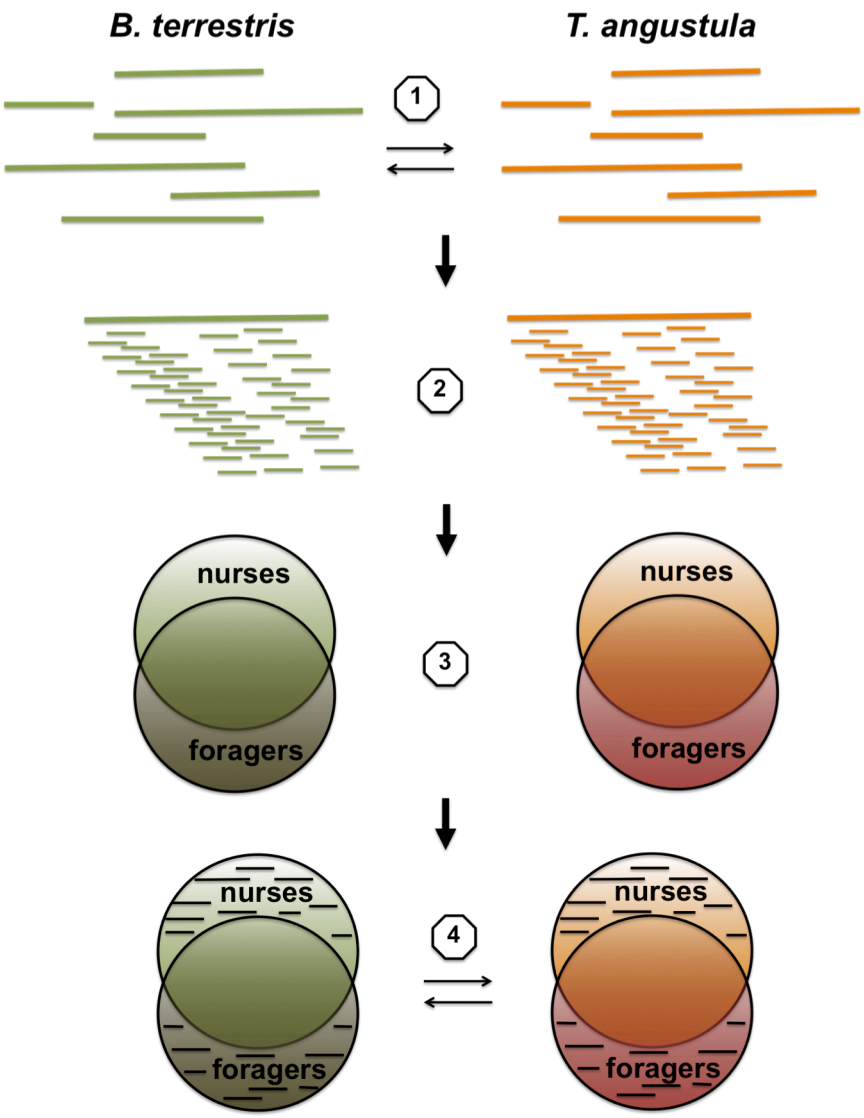


Figure 1. Workflow of analyses comparing nurses and foragers of *T. angustula* and *B. terrestris*. **1** – Transcriptome assembly. **2** - Realignment of the cleaned reads to the assembled transcripts for expression count. **3** - Differential expression analyses between the distinct worker subcastes within species. **4** - Comparison of differentially expressed genes involved in the work labour division of both species and its expression levels in nurses and foragers.

Supplementary Tables

Table I Main quality parameters from the complete transcriptome assembly of *T. angustula* and *B. terrestris* workers. Annotated transcripts refer to transcripts with at least one blast hit against the UniRef database

	T. angustula	***B. terrestris***
Total Number of transcripts	33,065	27,987
N50	3,614	5,490
%GC	38.11	37.19
Mean sample coverage	>100x	>130x
Annotated transcripts	26,623	21,638
lncRNA	347	431
Complete BUSCO orthologous	82.6%	91.9%

Table II Gene Ontology terms from the DEG in *T. angustula* with significant results in a two-sided Fisher's exact enrichment test. OVER – indicates the GO over represented in the tested set of differentially expressed genes when compared to the whole transcriptome annotation. UNDER – indicates the GO under represented in the tested set of differentially expressed genes when compared to the whole transcriptome annotation.

GO ID	GO Term	GO category	FDR	Over/ Under
GO:0005739	mitochondrion	Cellular Component	3.04E-04	OVER
GO:0070469	respiratory chain	Cellular Component	3.04E-04	OVER
GO:0045333	cellular respiration	Biological Process	0.002926225	OVER
GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	Molecular Function	0.003233042	OVER
GO:0022900	electron transport chain	Biological Process	0.004121204	OVER
GO:0004348	glucosylceramidase activity	Molecular Function	0.004121204	OVER
GO:0019866	organelle inner membrane	Cellular Component	0.004121204	OVER
GO:0005743	mitochondrial inner membrane	Cellular Component	0.004121204	OVER
GO:0055114	oxidation-reduction process	Biological Process	0.004121204	OVER
GO:0031966	mitochondrial membrane	Cellular Component	0.004121204	OVER
GO:0015980	energy derivation by oxidation of organic compounds	Biological Process	0.004121204	OVER
GO:0006091	generation of precursor metabolites and energy	Biological Process	0.004121204	OVER
GO:0009055	electron carrier activity	Molecular Function	0.004476534	OVER
GO:0016675	oxidoreductase activity, acting on a heme group of donors	Molecular Function	0.006554425	OVER
GO:0004129	cytochrome-c oxidase activity	Molecular Function	0.006554425	OVER
GO:0016676	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor	Molecular Function	0.006554425	OVER
GO:0005740	mitochondrial envelope	Cellular Component	0.006554425	OVER

GO:0015002	heme-copper terminal oxidase activity	Molecular Function	0.006585827	OVER
GO:0005975	carbohydrate metabolic process	Biological Process	0.00789253	OVER
GO:0009060	aerobic respiration	Biological Process	0.008634894	OVER
GO:0022904	respiratory electron transport chain	Biological Process	0.009295788	OVER
GO:0043169	cation binding	Molecular Function	0.009295788	OVER
GO:0006119	oxidative phosphorylation	Biological Process	0.013483612	OVER
GO:0008233	peptidase activity	Molecular Function	0.015364632	OVER
GO:0031975	envelope	Cellular Component	0.015364632	OVER
GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	Molecular Function	0.015364632	OVER
GO:0031967	organelle envelope	Cellular Component	0.015364632	OVER
GO:0015078	hydrogen ion transmembrane transporter activity	Molecular Function	0.017905695	OVER
GO:0044429	mitochondrial part	Cellular Component	0.018418226	OVER
GO:0008236	serine-type peptidase activity	Molecular Function	0.019382871	OVER
GO:0017171	serine hydrolase activity	Molecular Function	0.019382871	OVER
GO:0006665	sphingolipid metabolic process	Biological Process	0.019405302	OVER
GO:0070011	peptidase activity, acting on L-amino acid peptides	Molecular Function	0.021221896	OVER
GO:0016491	oxidoreductase activity	Molecular Function	0.034586684	OVER
GO:0004252	serine-type endopeptidase activity	Molecular Function	0.045137481	OVER
GO:0043168	anion binding	Molecular Function	0.004121204	UNDER
GO:0032549	ribonucleoside binding	Molecular Function	0.004121204	UNDER
GO:0001883	purine nucleoside binding	Molecular Function	0.004121204	UNDER
GO:0035639	purine ribonucleoside triphosphate binding	Molecular Function	0.004121204	UNDER
GO:0032550	purine ribonucleoside binding	Molecular Function	0.004121204	UNDER
GO:0001882	nucleoside binding	Molecular Function	0.004121204	UNDER
GO:0097367	carbohydrate derivative binding	Molecular Function	0.004121204	UNDER
GO:0032555	purine ribonucleotide binding	Molecular Function	0.004121204	UNDER
GO:0017076	purine nucleotide binding	Molecular Function	0.004121204	UNDER
GO:0032553	ribonucleotide binding	Molecular Function	0.004121204	UNDER
GO:0005524	ATP binding	Molecular Function	0.006585827	UNDER
GO:0030554	adenyl nucleotide binding	Molecular Function	0.006585827	UNDER
GO:0032559	adenyl ribonucleotide binding	Molecular Function	0.006585827	UNDER
GO:0036094	small molecule binding	Molecular Function	0.01231152	UNDER
GO:1901265	nucleoside phosphate binding	Molecular Function	0.015364632	UNDER
GO:0000166	nucleotide binding	Molecular Function	0.015364632	UNDER
GO:0016740	transferase activity	Molecular Function	0.019382871	UNDER

Table III Mean methylation level of genes in *B. terrestris* and *T. angustula*. Transcriptome – refers to the complete transcriptome; Differentially expressed genes – refers to all differentially expressed genes between nurses and foragers; Highly expressed in foragers – refers to the genes with higher expression in foragers when compared to nurses; Highly expressed in nurses – refers to the genes with higher expression in nurses when compared to foragers. Methylation estimations are based on WBS of nurses.

	B. terrestris	***T. angustula***
Transcriptome	0.66%	1.24%
Differentially expressed genes	0.78%	1.30%
Highly expressed in foragers	0.73%	1.57%
Highly expressed in nurses	0.95%	1.25%

Supplementary Figures

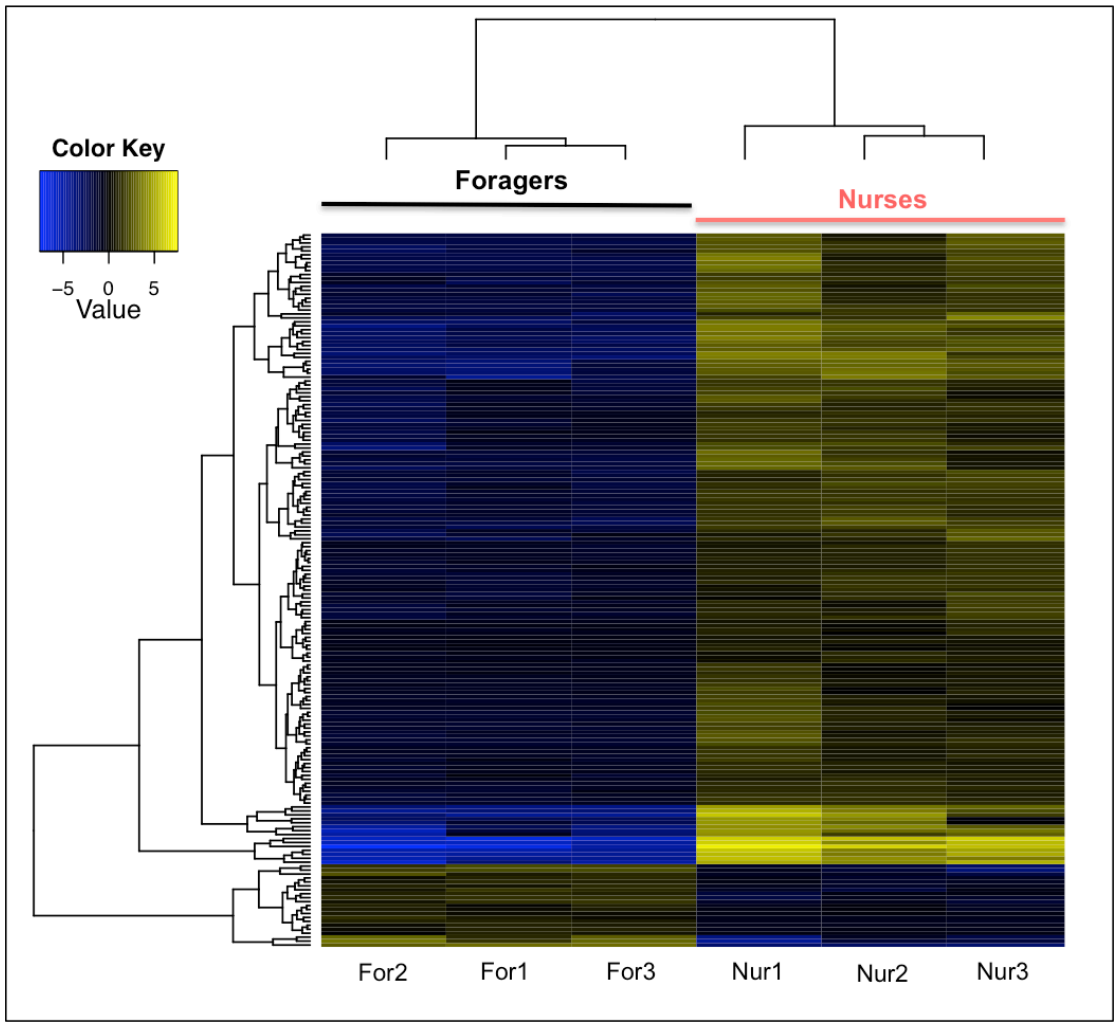


Figure 2 Heatmap of the 241 differentially expressed genes between foragers and nurses of *T. angustula*. **For1, For2** and **For3** – are replicated pool samples from foragers; **Nur1, Nur2** and **Nur3** – are replicated pool samples from nurses. Expression scale is on log2

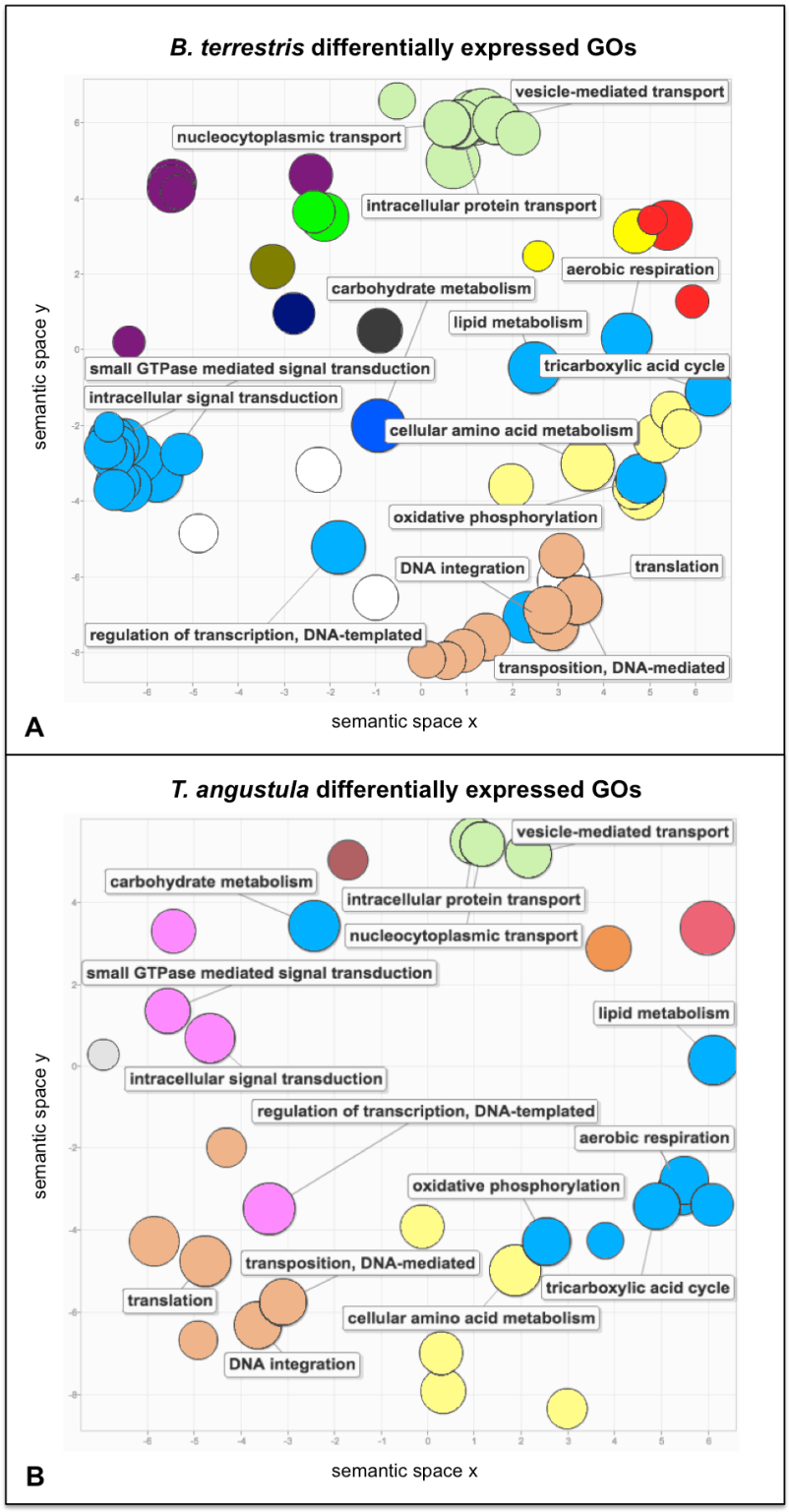


Figure 3 Scatterplot of the GO Biological Processes in which all the DEG between nurses and foragers are involved. GO terms are clustered in a two dimensional space based on semantic similarities among terms (i.e. closer terms are more similar). GO terms common to both species datasets are the only ones indicated by names. Colours refer to "superclusters" of loosely related terms, circles with the same colour means the different terms are from the same superclusters. Superclusters were determined by REVIGO[15]. Circle size indicates the frequency of the GO term in the UniProt database (more general terms have larger circles).

S2 to S6 follow at https://github.com/nat2bee/Suppl_PhDthesis

References cited

1. Altenhoff, A. M. *et al*. The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res*. **43,** D240–D249 (2015).
2. Woodard, S. H., Bloch, G. M., Band, M. R. & Robinson, G. E. Molecular heterochrony and the evolution of sociality in bumblebees (*Bombus terrestris*). *Proc. R. Soc. B Biol. Sci*. **281,** (2014).
3. Grüter, C. *et al*. Repeated evolution of soldier sub-castes suggests parasitism drives social complexity in stingless bees. *Nat. Commun*. **8,** e4 (2017).
4. Koedam, D. & Tienen, P. G. M. Van. The regulation of worker-oviposition in the stingless bee. *Insectes Soc*. **44,** 229–244 (1997).
5. FASTQC developed by Babraham Bioinformatics. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
6. Hannon lab. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html
7. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*. **38,** 1–7 (2010).
8. SEQYCLEAN developed by Ilya Zhbannikov.
9. Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *Genome Announc*. **2,** (2012).
10. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–5 (2013).
11. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).
12. García-Alcalde, F. *et al*. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28,** 2678–9 (2012).
13. Musacchia, F., Basu, S., Petrosino, G., Salvemini, M. & Sanges, R. Annocript: A flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. *Bioinformatics* **31,** 2199–2201 (2015).
14. Suzek, B. E. *et al*. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–32 (2015).
15. Supek, F., Bo#njak, M., $kunca, N. & $muc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* **6,** e21800 (2011).

Chapter 4

Unveiling the expression dynamics of genes involved in bee sociality

Natalia de Souza Araujo, Yannick Wurm*, Bob Schmitz* and Maria Cristina Arias

* Author(s) not consulted about this first manuscripts version

Revista de interesse: *PNAS – Proceedings of the National Academy of Sciences of the United States of America* (trabalho ainda não submetido)

Abstract

Bees are a great model to study the evolution of social behaviour since in several taxa within this group sociality seems to have evolved independently. Which originated a great diversity of social life styles. The tribes Apini and Meliponini are comprised only by highly eusocial bee species, whereas various levels of sociality can be detected in other tribes, being the vast majority of bees indeed solitary. Although the molecular evolution of eusociality has been the subject of many studies, the genetic changes involved in this behaviour have not been completely understood. Fundamental questions about shared and derivate gene pathways involved in the different social systems are still open. Recently new sequencing technologies have allowed gene expression studies of non model and model organisms in a deep and non-directional way, which is promising for evolutionary studies of complex behavioural traits. Herein, some of these new molecular tools were used to investigate the gene expression profile of different bee species with distinct behaviours. Using a unique approach, designed to avoid biological and technical confounding factors, we report 787 genes possibly involved in bee sociality. Many of these genes were successfully found in the transcriptome of other bee species in a moderate expression level. Therefore, future comparative studies with these "behavioural genes", in different lineages, are straightforward. Additionally, DNA methylation analyses indicate that the overall amount of methylation is not directly correlated to bee

behaviour, as previously stated, instead the DNA methylation context is more likely related to differential methylation involved in social dynamics.

Introduction

Interspecific interactions such as female choice, species recognition and labour division are examples of social activities (1). In some species individuals interact only sporadically, while in others, each individual is part of a complex social structure (2). Understanding the evolutionary mechanisms beneath these structured animal societies is one of the main questions of the modern biology (3–5). Social leaving species evolved independently multiple times, and although their social communities phenotypically converge in many features (6), sociality is still a very diverse behaviour, controlled by numerous genes (2, 7–9). Only among the insects, sociality has emerged at least 12 independent times (6), and distinct levels of social complexity can be observed (1).

Bees comprise a remarkable group to study the various aspects of sociality. In this group, social organization is greatly variable and has evolved multiple times, which allows studying the evolution of the different social forms comparatively (10–12). Bee species can be roughly classified as solitary, subsocial, communal, quasisocial, primitively eusocial or highly eusocial (13). The vast majority of bees are solitary (14), which means that a single female performs all tasks related to its own existence and reproduction, abandoning the nest right after its completion (13). In subsocial species, females are essentially solitary, but after offspring emergency, they still feed and care for their brood (15). In communal and quasisocial nests, multiple fertile females construct reproductive cells in a common location. However, while in communal species all females build and use their own cells, in quasisocial lineages one individual is dominant and usurp the cells built by other females to lay its eggs (13). Eusociality is the most complex form of social behaviour observed in bees (12). In these species there are two castes, queen and worker (normally sterile), that contribute to colony maintenance (10). Primitively eusocial species, such as the bumblebees, live in smaller colonies (of a few hundred individuals) with annual cycles and less specialized castes (16, 17). Conversely, highly eusocial bees (from the

tribes Apini and Meliponini) have the largest (comprising thousands of bees) and longest living colonies maintained by exceptionally specialized castes (13, 18).

Most elucidative studies about common molecular bases for sociality, include the analyses of large amounts of data, from different species, in a comparative way (7, 19, 20). These studies have provided valuable information concerning specific and convergent aspects of the social behaviour. However, many social systems are understudied. Meaning that, there are still many gaps concerning the molecular mechanisms involved in the evolution of the different forms of sociality to support a broad evolutionary hypothesis (9, 21, 22). Therefore, it is essential to expand the diversity of species and social life styles in studies applying recent molecular approaches.

Based on eco-evolutionary ideas and molecular data, it was hypothesized the existence of a genetic toolkit underneath the social behaviour of different species (1, 23). This hypothesis has guided many studies on the evolution of sociality in the last years (19, 24–27). Recently, analyses of genomic data have suggested that the toolkit hypothesis should be adjusted to accommodate the existence of common gene networks in different social species, instead of a set of specific genes (7, 9, 20). And an increasing number of evidences indicate that also mechanisms of gene expression control would be particularly important to regulate insects social behaviour (20, 28, 29). However the comparative studies performed so far were not conclusive about the existence of a genetic toolkit for sociality.

Because behavioural changes are dynamic, and not always involve alterations in the DNA sequence (30), studying the molecular mechanisms involved in the social behaviour is especially challenging (21). However, studies concerning gene expression changes and its control mechanisms are promising (29, 31). Herein, we aim to add valuable information about the convergent molecular features involved in different social forms, through the first comparative analyses of transcriptomic and epigenetic data from a number of bee species. We employed an innovative methodology to perform these comparisons. Altogether, twelve bee species were used in this study (S1 – Table I). Three of them, *Tetrapedia diversipes* (a solitary species), *Bombus terrestris* (a primitively eusocial species) and *Tetragonisca angustula* (a highly eusocial species), were used as models for comparative expression analyses, which allowed the identification of possible behavioural genes and the DNA

methylation pattern analyses (Figure 1). The remaining nine bee species were used to verify the occurrence of these genes in other species and to analyse their general expression profile within each lineage (Figure 2). The expression and DNA methylation profile of different bee species, with distinct behaviours, were analysed in highly controlled conditions to avoid biological and technical confounding factors in the identification of a possible genetic toolkit for sociality.

Figure 1. Bee species used as models for behavioural genes identification. From left to right: *Tetrapedia diversipes*, a solitary bee; *Bombus terrestris*, a primitively eusocial bee; *Tetragonisca angustula*, a highly eusocial bee. Photos credit: Guraci D. Cordeiro (*T. diversipes*); Julio Beeman – The Beekeeping Family (*B. terrestris*); Alex Wild (*T. angustula*).

Results

Behavioural Genes identification - Differential expression analyses of the three species

To identify genes possibly related to behaviour, the levels of gene expression in the solitary bee *T. diversipes*, in the primitively eusocial bee *B. terrestris* and in the highly eusocial bee *T. angustula*, from adult females (founder in *T. diversipes* and nurses in the eusocial bees) and larvae (from multiple instars), were compared. The final transcriptome assembly of each species was obtained by compiling data from adults and larvae and including all assembled isoforms. For *T. diversipes* it were identified 19,778 coding transcripts; 20,777 for *B. terrestris*; and 26,157 for *T. angustula* (S1 – Table II). All these transcripts had at least one blast hit against the UniRef or the Pfam databases. Orthologous search reported 6,413 unique orthologous among the three species. This represents around half of all genes been expressed in these species, as indicated by the realignment ratio of the reads. Where about 42.83% of adults and 44.92% of the larvae cleaned reads from *T. diversipes* realigned to the

orthologous; in *B. terrestris*, realignment ratio was 57.88% for adults reads and 60.40% for larvae; and in *T. angustula*, 47.51% and 52.08% of the cleaned reads from adults and larvae, respectively, realigned to the orthologous coding transcripts (S1 – Figure 1).

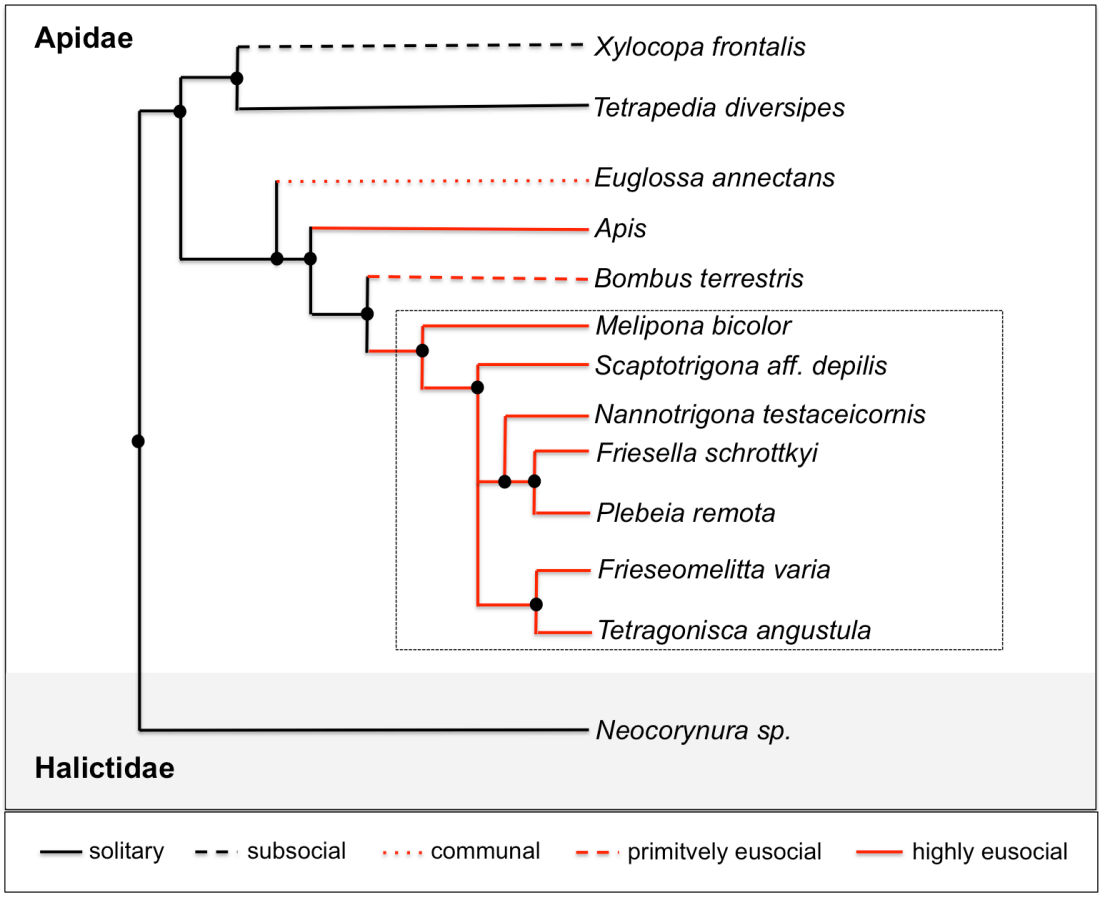


Figure 2. Phylogenetic relationship among species used in this study and their behaviour. The honey bee genus, *Apis*, was included in this phylogeny just for comparison, but no species of this genus was investigated here. Highlighted, inside of the doted box, are the species from the highly eusocial tribe, Meliponini. Figure based on 12, 18, 32

Differential expression analyses were performed in two ways. First, differentially expressed genes (DEG) among the three model species (Figure 1) were identified. As expected, because of the phylogenetic distance (Figure 2), most of the DEG were between *T. diversipes* and the other two eusocial bees: 1,770 DEG between *T. diversipes* and *B. terrestris*, and 1,679 DEG between *T. diversipes* and *T. angustula*. Still, between the eusocial bees 1,321 genes were differentially expressed. Altogether 2,839 genes, from the 6,413 orthologous, were differentially expressed. The amount of genes highly expressed in selected groups are illustrated in Figure 3, some genes were always over expressed in one species when compared to the two

others, while some genes were commonly highly expressed in two of the three species.

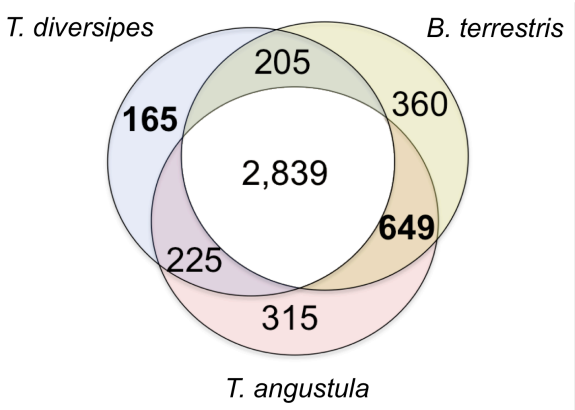


Figure 3. Orthologous genes differentially expressed among *T. diversipes*, *B. terrestris* and *T. angustula*. In the middle of the diagram total number of DEG are represented. Other numbers refers to the amount of genes exclusively highly expressed in each group (e.g. 165 in *T. diversipes*, indicates that 165 genes were only highly expressed in *T. diversipes* when compared to the other bees). Intersections represent genes that are commonly highly expressed in two of the three species. In bold are highlighted genes over expressed only in the solitary bee and commonly highly expressed only in the eusocial species; these genes were used for behavioural genes identification.

The second approach used for differential expression analyses was to group the eusocial species for comparisons. The two eusocial bees, *B. terrestris* and *T. angustula*, comprised the eusocial group and the solitary bee, *T. diversipes*, represented the solitary group. We then checked for orthologous transcripts with different levels of expression between the solitary and the eusocial groups. In this case, 987 genes were commonly highly expressed in the eusocial species and 373 were over expressed in the solitary bee.

For the final gene selection, we compared results from both differential expression analyses. Thus, genes identified as highly expressed only in the solitary bee (165 from first analyses; 373 – from second analyses) were compared. Only genes reported in both analyses were selected. The same was done for the genes highly expressed in the eusocial species (649 from first analyses; 987 – from second analyses). A total of 632 genes were selected as over expressed only in eusocial bees and 155 were over expressed in the solitary species. These 787 genes were considered as behavioural genes (BG), which are genes possibly involved in bee social behaviour.

Functional analyses of the Behavioural Genes – Comparison among the three species

BG from each species sometimes blasted to different genes in database, so Gene Onthology (GO) terms were slighted different in each species dataset. Nevertheless, the same biological processes were always among the most common GO terms: regulation of transcription [DNA-dependent]; transcription [DNA-dependent]; signal transduction and translation (S2 – S7). However, no GO terms were significantly enriched in the BG set when compared to all orthologous.

DNA methylation analyses of the Behavioural Genes – Comparison among the three species

Whole bisulfite sequencing (WBS) of *T.* diversipes, *B. terrestris* and *T. angustula* adult females were performed to investigate the DNA methylation pattern of BG genes. Samples of the eusocial species (nurses) and of the solitary bee (founder) were sequenced. The methylation analyses were performed only for coding sequences (see material and methods for details). Since most of the DNA methylated sites occur within coding regions in Hymenoptera (33–35), these regions are reliable representatives of bees whole genome methylation. In average, 1.88% of all cytosine nucleotide from orthologous genes of *T. diversipes*, were methylated. For *B. terrestris* and *T. angustula*, this ratio was 1.02% and 1.42%, respectively. In *T. diversipes* and *T. angustula* the mean methylation level of the BG was slightly higher than these values; 2.05% in *T. diversipes* and 1.51% in *T. angustula* (S1 – Table III). However, the DNA context of the methylated sites in each species varied in a similar way for all bees (Figure 4). When compared to total orthologous, in BG the CG sites have reduced levels of methylation (Figure 4A) and consequently, contributed less to the overall methylation profile (Figure 4B) in all three species. This pattern was equally observed in eusocial and in solitary highly expressed BG, but it was more pronounced among the BG highly expressed in the solitary bee (S1 – Table III).

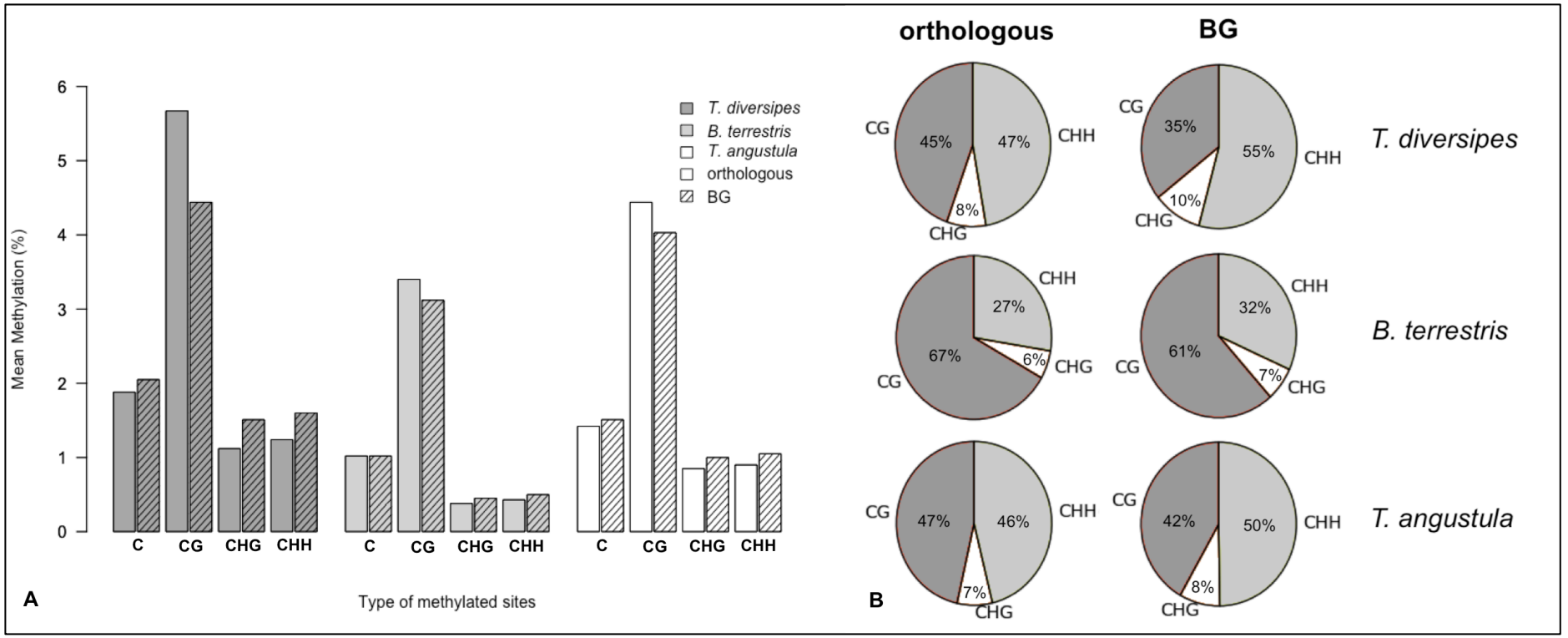


Figure 4. Methylation profile of *T. diversipes*, *B. terrestris* and *T. angustula* obtained with whole body bisulfite sequencing (minimum of 10X coverage). H may be A, T or C nucleotides. A – Percentage of each methylation site (cytosine or cytosine groups). C refers to the mean global level of methylation. B – Contribution of each methylation type in the general methylation profile. Orthologous refers to all common genes found in the three species. BG are behavioural genes.

Behavioural Genes in other bee species

The existence and general expression pattern of the BG were also determined for other nine bee species with different levels of social organization (S1 – Table I). In this case, samples from adult females were sequenced with reduced coverage and without replicates, so levels of gene expression for orthologous were only ranked. Total transcriptome assembly data for these bees are reported in S1 – Table II. Number of identified orthologous varied from 4,947 (in *Neocorynura sp.*) to 5,498 (in *Frieseomelitta varia*) (Table I). The relative orthologous gene expression was estimated as fold changes in log2 from the species mean (S1 – Figure 2). This type of analyses does not allow differential expression comparisons among species, instead it reveals the expression pattern of the genes within species.

Altogether, 349 BG were identified in the nine bee species, and of these, some had different levels of expression within species. The number of BG reported with different expression levels varied from 93 (in *S. aff. depilis*) to 174 (in *X. frontalis*) (Table I), including 26 genes that were reported as differentiated in all nine bees (S1 – Table IV). In general, BG reported as commonly highly expressed in the eusocial species, *B. terrestris* and *T. angustula*, were more common among the genes with different levels of expression (GDE) than the ones highly expressed in the solitary *T. diversipes* (Table I). Analyses of the BG highly expressed in each social group also indicate no correlation between the within species expression pattern of the BG and the bee behaviour. If this correlation existed, it would be expected that, for all the highly eusocial species studied (*F. schrottkyi*, *F. varia*, *M. bicolor*, *N. testaceicornis*, *P. remota* and *S. aff. depilis*), the BG commonly highly expressed in *B. terrestris* and *T. angustula* would also be over represented and the BG highly expressed in *T. diversipes* would be down represented. But the opposite would be expected in solitary bees. However, as can be seen in Table I, almost all species presented the pattern expected for eussocial species, regardless their behaviour. *N. testaceicornis*, a highly eusocial bee, was the only species showing a different pattern.

Table I. Results of the identification of behavioural genes and their expression pattern in other bee species. Values represent total number of genes in each group. Orthologous – indicate how many of the 6,413 orthologous reported among the three main species are also find in the different bee species. GDE – are genes with different levels of expression from the species mean. BG – are the 787 behavioural genes identified when comparing gene expression data from the three main species. Eusocial BG – are the BG commonly highly expressed in the two eusocial bee species (*B. terrestris* and *T. angustula*) in the three main species comparison. Solitary BG – are the BG highly expressed only in the solitary bee *T. diversipes* in the three main species comparison. OVER – means the within species level of expression of these genes were equal to or above 2 log2(fold changes). UNDER – means the within species level of expression of these genes were equal to or bellow -2 log2(fold changes).

Species	Orthologous	GDE	BG in GDE	BG not in GDE	Eusocial BG OVER	Eusocial BG DOWN	Solitary BG OVER	Solitary BG DOWN
Neocorynura sp.	4,947	985	136	446	58	53	8	17
Xylocopa frontalis Olivier, 1789	5,080	1,315	174	437	76	65	15	18
Euglossa annectans Dressier, 1982	5,241	1,226	159	462	74	62	8	15
Friesella schrottkyi (Friese, 1900)	5,431	1,256	159	492	78	58	7	16
Frieseomelitta varia (Lepeletier, 1836)	5,498	1,224	160	508	73	59	10	18
Melipona bicolor Lepeletier, 1836	5,177	929	122	497	70	30	10	12
Nannotrigona testaceicornis Lepeletier, 1836	5,220	1,293	164	448	59	73	18	14
Plebeia remota (Holmberg, 1903)	5,308	873	118	506	75	34	8	1
Scaptotrigona aff. depilis (Moure, 1942)	5,000	767	93	499	67	14	8	4

Discussion

Molecular comparative analyses of dynamic features, such as the social behaviour, among different species is challenging because a number of variables might affect gene identification (21). But despite these risks, comparative studies are still one of the most promising ways to unravel conserved molecular paths in evolution (7, 19, 20, 29). To avoid some misleading results in gene expression comparisons of the three bee species reported here, we employed a unique approach of analyses to account for many biological and technical variables. To deal with biological variance, we analysed multiple individuals of different bee species in distinct developmental stages, controlling for behaviour and sampling period (36, 37). To avoid technical confounding factors, multiple samples were used as replicates with a high sequencing coverage and low p-value for DEG identification (38).

Specially, the combination of larvae and adult transcriptomic data was important to improve gene expression comparisons among species. Because by combining different developmental stages data we were able to account for conserved and species specific gene expression profiles, since gene expression during larvae development are highly conserved among different species (36) and gene expression in adult bee workers are more variable (34, see Chapter 3). In this manner, the comparative approach used to determine the BG herein mitigates the number of confounding factors in the analyses, which in turn, reduces the identification of genes falsely related to bee behaviour. Nevertheless, to eliminate completely the identification of false positive genes in gene expression studies is very unlikely, even in within species comparisons (38).

The comparative analyses of the three main species provided an important dataset to comprehend the molecular mechanisms involved in bee social behaviour. Gene expression comparisons successfully identified conserved and unique expression patterns among different bee species and behaviours, as indicated in Figure 3. And although some of the BG identified may reflect the species phylogenetic relationship, this gene set also represents a relevant group of candidate genes to be involved in bee sociality. Many of these genes are implicated in biological processes that are likely to be involved in behavioural modulation, including transcription, regulation of transcription, signal transduction and translation. This results supports previous studies that indicate that genes involved in the regulation of gene expression

machinery are important in social behaviour (21, 29)

Additionally to using these comparisons to identify BG, the different identified gene sets illustrated in Figure 3, could also be explored to investigate other evolutionary questions. For example, the genes highly expressed only in *T. angustula* and only in *B. terrestris* represent important lineage specific adaptations that could reveal molecular mechanisms beneath the differences observed between highly eusocial and primitively eusocial bees. Besides, the methodology used here is a promising way to investigate gene expression differences comparatively and could be used in different species to answer distinct questions.

Analyses of the BG in other nine bee species revealed that a great number of these genes could be found in these bees (Table I). This supports the use of the BG as candidate genes to investigate the molecular evolution of behaviour across different species. Specifically 349 BG were found expressed in all bees, even when using a reduced coverage and only one adult stage per sample (as it was the case for the nine bee species tested, see material and methods section for details). Furthermore, BG internal expression levels appear not to be directly related to the bee behaviour. Which means that genes important to behaviour are not necessarily over or down expressed in a species when compared to other genes also been expressed at the same time. It reveals that genes with moderated expression levels within species might be important in behavioural adaptations. So, in order to identify changes in expression related to the evolution of social behaviour in different lineages, analyses should always be performed comparatively among species.

After the publication of the ten complete bee genomes in 2015 (20), the role of DNA methylation in social behaviour has been in evidence (33, 39, 40). Prediction of methylated sites in the assembled genomes, suggested that an increase in social complexity would also mean an increase in the levels of gene methylation (20). Whole bisulfite sequencing for the three bee species used in our study does not support this hypothesis. Overall, the female founder of the solitary bee *T. diversipes* had higher levels of methylation in its genes than nurses of the other two eusocial bees, *B. terrestris* and *T. angustula*. It indicates that genes of species with more complex social behaviour are not necessarily more methylated. The general mean level of gene methylation was also not remarkably different between all orthologous and the behavioural genes in these bees. These evidences support the idea that the

overall amount of methylation is not directly correlated to eusociality. Nevertheless, these data do not discard the involvement of differential gene methylation of specific genes in sociality.

General patterns identified here, suggest that more than the overall amount of DNA methylation in genes, there are observable changes in the preferable sites for methylation in the BG. For all three species the DNA context of methylated sites varied in a similar way; there was a decrease in the amount of CG dinucleotide methylated and an increase of methylation in non-CG sites of the BG, when compared to all orthologous. The effects of differential DNA methylation contexts are poorly understand and underestimated (41). Methylation in CG and non-CG sites are mediated by different mechanisms (42), in honey bees, as in other mammals, methylation at CG sites are maintained by DNA methyltransferase 1 (Dnmt1) while non-CG methylation are kept by mechanisms of *de novo* methylation involving the DNA methyltransferase 3 (Dnmt3) (41). For most animals, including honey bee, wasps and the bee species studied here, most methylation occur in CG dinucleotides (35, 43) but it is also known that the amount of non-CG methylation is underestimated due to difficulties in the identification of these sites (34, 41).

A possible biological role for methylation in different DNA contexts were previously reported for ants (34) and honey bees (41). For *Apis*, non-CG methylation seems to be involved with alternative mRNA splicing and is especially enriched in genes related to behavioural responses (41). Although no direct correlation of this type of methylation to behaviour could be confirmed in this previous study (41), this data support our hypothesis that differential DNA methylation contexts are probably related to social behaviour in bees. It is even expected that the amount of non-CG sites methylated in the BG are underestimated in our study, since this type of methylation seems to be enriched in introns and not in exons, as for CG methylation (41). More studies are still needed to infer how specific differential methylation marks in some genes could affect behavioural dynamics, but these results suggests that, in future analyses, the DNA methylation context should also be considered in the study of social evolution.

In this study we tested the use of comparative expression analyses to identify genes possibly related to social behaviour in different bee species. Obtained results indicate that this type of comparative study is not only feasible but also essential to

identify the molecular mechanisms shared by different species. Behavioural genes identified here are good candidate to investigate the evolution of social behaviour in bees and to evaluate the toolkit hypothesis. They have a consistent gene expression level when using high coverage RNA sequencing of multiple individuals, samples and different life stages, and were expressed in several bee species, which will allow further comparative analyses to refine gene selection. Additionally, DNA methylation pattern of the BG suggested that, although the overall amount of methylation may not be directly related to social behaviour complexity, the DNA methylation context is probably involved in social dynamics.

Material and Methods

Sample collection and sequencing

Different developmental stages were identified according to observations and age control, depending on the species collected (detailed sampling methods in S1). Individuals were always sampled between 10h-12h and immediately frozen in liquid nitrogen. Whole bodies were used in all RNA and DNA extractions. For RNA-Seq, total RNA was extracted using the Qiagen® extraction kit (RNeasy Mini Kits). RNA quality and quantification were verified using Bionalyzer® and the Nanodrop®, respectively. The samples were used for RNA sequencing in the Illumina® HiSeq 2000 (Macrogen, South Korea; LACTAD, Unicamp, Brazil; or QMUL Genome Centre, UK).

T. diversipes presents two main reproductive generation during the year and, in each generation, developmental time greatly varies (44). So, to avoid differences in expression due to this developmental characteristic, for this bee, six samples were sequenced from each developmental stage, three samples from generation one and three from generation two. For *T. angustula* and *B. terrestris* three different colonies were used as sample replicates. Altogether 12 high coverage samples were sequenced for *T. diversipes* (6 founders samples and 6 larvae samples), and 6 high coverage samples were generated for each eusocial bee (3 nurses samples and 3 larvae samples). Sequenced samples were always a pool of individuals; three individuals were pooled for *T. diversipes* and *B. terrestris* samples and six individuals were pooled for *T. angustula*. Nurses and founders were chosen to represent the adult stage because previous expression analyses suggested that these would be the most

comparable phases (8). For whole bisulfite sequencing, total DNA from one nurse bee of each eusocial species and one founder from the solitary bee were extracted using a phenol-chloroform protocol (45). Whole bisulfite sequencing were performed following the protocol described in (46) using the Illumina® NextSeq500 (University of Georgia). RNASeq data from *T. diversipes* samples were previously used in Chapter 2, and *T. angustula* and *B. terrestris* nurses data in Chapter 3 analyses. WBS of *T. angustula* and *B. terrestris* were also used in Chapter 3.

Total RNA from the additional nine bee species was extracted from adult females. Nurses were sampled for the highly eusocial species (*F. schrottkyi*, *F. varia*, *M. bicolor*, *N. testaceicornis*, *P. remota* and *S. aff. depilis*); one fertile female of *E. annectans* that were sharing a nest with other bees was collected; founders that were still feeding and caring for their offspring were used for *X. frontalis*; and founders were sampled for *Neocorynura sp.*. One pooled sample, containing the RNA from 3 – 6 individuals (depending on the bee size) were sequenced per species. Except for *E. annectans*, in which only one bee was sequenced.

Transcriptome assembly and differential expression analyses – Comparisons among the three species

Reads quality assessment was performed with the FastQC program v0.11.2 (47) before and after cleaning. The FASTX Toolkit v0.0.14 (48) was used to trim the first 14 bp of all reads because of the initial GC bias (49). Low quality bases (phred score below 30) and small reads (less than 31 bp) were removed using SeqyClean v1.9.3 (50).

For each species the transcriptome assembly were performed differently. Because the complete genome of *B. terrestris* is available (51), it was used as reference in two assembling approaches: reference assembly – using HISAT2 v2-2.0.3 (52) and StringTie v1.2.2 (53); and reference guided *de novo* assembly – using Trinity v2.1.1 (54). For *T. angustula* the closest genome available, from *Melipona quadrifasciata* (20), was used for a reference guided *de novo* assembly. Normalized reads of this bee was also used for a traditional *de novo* assembly. In both cases, the Trinity program was used. For *T. diversipes* only the *de novo* assembly, with no reference genome, was performed with Trinity. Assemblies from each developmental stage (adults and larvae), and different generations in *T. diversipes* case, of all species

were performed independently to avoid chimera transcript assemblies (55). For traditional *de novo* transcriptome assemblies, cleaned reads were digitally normalized (20x coverage) previous to assembly, to increase software efficiency (56). Programs used in transcriptome assembly had default parameters. Later the independent assemblies were merged with CD-Hit v4.6 (57) at 95% similarity, Corset v1.05 (58) with minimum of 50x coverage, and Lace v0.80 (59).

Quality parameters of the final assemblies were analysed using QUAST v4.0 (60), BUSCO v2 (61) and Qualimap v2.2 (62). Transcripts were then annotated with Annocript v1.2 (63) using the UniProt Reference Clusters (UniRef) (64) and the Pfam databases (June 2016 version). Transcripts with significant blast hits (e-value < 1e-5) against possible contaminants (plants, fungus, mites and bacteria) in UniRef were removed from the final dataset. This annotation pipeline was used even for *B. terrestris* transcriptome dataset; using the annotation based on *B. terrestris* genome would hinder comparisons.

Differential expression analyses were performed adapting scripts available in the Trinity package. Bowtie2 v2.2.5 (65), RSEM v1.2.22 (66) and DESeq2 (67) (p-value < 1e-5) were used to identify differentially expressed genes. The design used in the analyses was "design = ~ stage + condition", where stage may be "adult" or "larvae" and condition was: "sp1", "sp2" or "sp3" in the first differential expression analyses; and "eusocial" or "solitary" in the second analyses. Identification of enriched GO terms was performed with a two-tailed Fisher's exact test using Blast2GO®.

Orthologous identification

Orthologous of the three main species were identified using tblastx v2.6.0 (68). All transcripts from one species were blasted against all transcripts of the others, and the best hit was selected (minimum e-value of 1e-10). Unique hits were chosen based on similarity, so when the same gene blasted with more than one transcript from the same species, only one hit pair per gene was maintained. For the identification of orthologous transcripts in the other nine species a similar approach was used, but instead of blasting all transcripts to all species, only the orthologous transcripts from the closest species was used. For all Meliponini, orthologous from *T.*

angustula were used in the blast searches. For the remaining species orthologous from *B. terrestris* were used.

DNA methylation analyses

Trim Galore v0.4.3 (69) wrapper script, with default parameters, was used to automate cleaning and adapter trimming of the bisulfite converted reads. Orthologous transcript and BG were used as reference for reads alignment, so only coding regions could be analysed. PCR bias filtering, alignment of the cleaned reads and methylation call were performed with BS-Seeker2 v2.1.0 (70), because this program allows the use of Bowtie2 in local alignment mode. CGmapTools v0.0.1 (71) was used to filter low coverage methylated sites (minimum of 10x) and statistics.

Transcriptome assembly and estimation of expression levels – Nine species data

None of the additional nine bee species sequenced have a reference genome, so we performed a similarity test to verify if other available bee genomes could be used. After cleaning, reads were aligned to the complete genome of the closest species available in database, using HISAT2 (52). When more than 20% of the reads were successfully aligned to the reference genome the reference guided *de novo* assembly and the traditional *de novo* assembly were performed with Trinity, and their results were merged as described for the three species assembly. This was the case for *E. annecans* (with *E. mexicana* reference genome) and most stingless bees (with *M. quadrifasciata* reference genome), except *N. testaceicornis* (S1 – Figure 3). For the remaining species only the traditional *de novo* assembly was performed with Trinity.

To estimate the levels of gene expression, cleaned reads were aligned to the assembled transcriptomes using Bowtie2. Read counts and normalization were then performed using RSEM. In order to remove outliers from the mean calculation, the mean level of gene expression for the species were determined using all TPM counts between the first and third quantiles. Fold change for each orthologous gene was then calculated dividing the TPM count of expression, per the species mean level of gene expression. Values were lately converted to log2. Genes considered to have higher or lower levels of expression varied at least in two log2(fold changes) from the species mean, which means these genes have at least four times more or less levels of

expression than the mean expression of orthologous genes. Analyses of different expression levels were performed in R with custom Rscripts.

Acknowledgements

For support during sampling of the different species the authors would like to thank: MSc. Priscila Karla Ferreira dos Santos, Larissa Logullo Piconi and to Leticia Eiko Kikuta from the LGEA (Universidade de São Paulo); MSc. Sheina Koffler, Dr. Isabel Alves-dos-Santos, Dr. Guaraci Duran Cordeiro, Priscilla Baruffaldi Bittar and Dr. Sergio Dias Hilário from the Laboratório de Abelhas (Universidade de São Paulo); to Dr. Lars Chittka and Dr. Stephan Wolf from Bee Sensory and Behavioural Ecology Lab (Queen Mary University of London); to Dr. Andres Arce and Dr. Richard Gill (Imperial College London – Silwood Park); to Dr. Denise de Araújo Alves (Universidade de São Paulo – ESALQ); to Dr. Solange Cristina Augusto (Universidade Federal de Uberlândia); and to the Brazilian beekeeper Antonio. Dr. Isabel Alves-dos-Santos identified the *Neocorynura sp*. bees, for which we are grateful. We also would like to thank Susy Coelho for technical assistance and Dr. Tatiana Teixeira Torres for collaborating in the initial design of this study. Additionally, we thank the funding agency FAPESP (São Paulo Research Foundation, process numbers 2013/12530-4 and 2012/18531-0) and the Research Centre on Biodiversity and Computing (BioComp) of the Universidade de São Paulo (USP), supported by the USP Provost's Office for Research, for financial support. Part of the bioinformatic analyses was performed at the cloud computing service from USP and at Queen Mary University of London computing cluster.

Authors declare they have no competing financial interests.

References

1. Robinson GE, Fahrbach SE, Winston MLW (1997) Insect societies and the molecular biology of social behavior. *Bioessays* 19(12):1099–1108.
2. Robinson GE, Ben-Shahar Y (2002) Social behavior and comparative genomics: new genes or new gene regulation? *Genes Brain Behav* 1(4):197–203.
3. Robinson GE (1999) Integrative animal behaviour and sociogenomics. *Trends Ecol Evol*. doi:10.1016/S0169-5347(98)01536-5.
4. Wilson EO (2000) *Sociobiology!: the new synthesis* (Belknap Press of Harvard University Press).
5. Nowak MA, Tarnita CE, Wilson EO (2010) The evolution of eusociality. *Nature* 466(7310):1057–1062.
6. Fischman BJ, Woodard SH, Robinson GE (2011) Molecular evolutionary analyses of insect societies. *Proc Natl Acad Sci U S A*:10847–54.
7. Berens AJ, Hunt JH, Toth AL (2014) Comparative transcriptomics of convergent evolution: Different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Mol Biol Evol*:1–14.
8. Toth AL, et al. (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* 318(5849):441–4.
9. Toth AL, Rehan SM (2017) Molecular Evolution in Insect Societies: An Eco-Evo-Devo Synthesis. *Annu Rev Entomol* 62(1):419–442.
10. Cardinal S, Danforth BN (2011) The antiquity and evolutionary history of social behavior in bees. *PLoS One* 6(6):e21086.
11. Peters RS, et al. (2017) Evolutionary History of the Hymenoptera. *Curr Biol* 27(7):1013–1018.
12. Bossert S, Murray EA, Blaimer BB, Danforth BN (2017) The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. *Mol Phylogenet Evol* 111:149–157.
13. Michener CD (2007) *The Bees of the World* (JHU Press). second.
14. Batra SWT (1984) Solitary Bees. *Sci Am* 250(2):120–127.
15. Camillo E, Garofalo CA (1989) Social Organization in reactivated nests of three species of *Xylocopa* (Hymenoptera, Anthophoridae) in southeastern Brasil. *Insectes Soc* 36(2):92–105.
16. Thompson GJ, Oldroyd BP (2004) Eavaluating alternative hypotheses for the origin of eusociality in corbiculate bees. *Mol Phylogenetics Evol 33* 33:452–456.
17. Woodard SH, Bloch GM, Band MR, Robinson GE (2014) Molecular heterochrony and the evolution of sociality in bumblebees (*Bombus terrestris*). *Proc R Soc B Biol Sci* 281(1780). doi:10.1098/rspb.2013.2419.
18. Grüter C, et al. (2017) Repeated evolution of soldier sub-castes suggests parasitism drives social complexity in stingless bees. *Nat Commun* 8(1):e4.
19. Woodard SH, et al. (2011) Genes involved in convergent evolution of eusociality in bees. *Proc Natl Acad Sci U S A* 108(18):7472–7477.
20. Kapheim KM, et al. (2015) Genomic signatures of evolutionary transitions from solitary to group living. *Science (80-)* 348(6239):1139–1143.
21. Kapheim KM (2016) Genomic sources of phenotypic novelty in the evolution of eusociality in insects. *Curr Opin Insect Sci* 13:24–32.
22. Søvik E, Bloch G, Ben-Shahar Y (2015) Function and evolution of microRNAs in eusocial Hymenoptera. *Front Genet* 6(MAY):1–11.
23. Robinson GE, Fernald RD, Clayton DF (2008) Genes and social behavior. *Science* 322(5903):896–900.
24. Grozinger CM, Fan Y, Hoover SER, Winston ML (2007) Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (Apis mellifera). *Mol Ecol* 16(22):4837–4848.
25. Robinson GE, Grozinger CM, Whitfield CW (2005) Sociogenomics: social life in

molecular terms. *Nat Rev Genet* 6(4):257–70.
26. Schulz DJ, Robinson GE (2001) Octopamine influences division of labor in honey bee colonies. *J Comp Physiol - A Sensory, Neural, Behav Physiol* 187(1):53–61.
27. Chandrasekaran S, et al. (2011) Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proc Natl Acad Sci U S A* 108(44):18020–18025.
28. Herb BR, et al. (2012) Reversible switching between epigenetic states in honeybee behavioral subcastes. *Nat Neurosci* 15(10):1371–1373.
29. Simola DF, et al. (2013) Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res* 23(8):1235–1247.
30. Robinson GE, Barron AB (2017) Epigenetics and the evolution of instincts. *Science (80-)* 356(6333):26–27.
31. Zhang ZH, et al. (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS One*:0–35.
32. Hedtke SM, Patiny S, Danforth BN (2013) The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. *BMC Evol Biol* 13:138.
33. Yan H, et al. (2015) DNA Methylation in Social Insects: How Epigenetics Can Control Behavior and Longevity. *Annu Rev Entomol* 60(1):435–452.
34. Bonasio R, et al. (2012) Genome-wide and caste-specific DNA methylomes of the ants camponotus floridanus and harpegnathos saltator. *Curr Biol* 22(19):1755–1764.
35. Lyko F, et al. (2010) The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biol* 8(11). doi:10.1371/journal.pbio.1000506.
36. Domazet-Loso T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468(7325):815–818.
37. Ometto L, Shoemaker D, Ross KG, Keller L (2011) Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Mol Biol Evol* 28(4):1381–92.
38. Lin Y, et al. (2016) Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* 17(1):28.
39. Lockett GA, Almond EJ, Huggins TJ, Parker JD, Bourke AFG (2016) Gene expression differences in relation to age and social environment in queen and worker bumble bees. *Exp Gerontol* 77:52–61.
40. Simola DF, et al. (2016) Epigenetic (re)programming of caste-specific behavior in the ant *Camponotus floridanus*. *Science (80)* 351(6268):aac6633-aac6633.
41. Cingolani P, et al. (2013) Intronic Non-CG DNA hydroxymethylation and alternative mRNA splicing in honey bees.
42. Bernatavichute Y V., Zhang X, Cokus S, Pellegrini M, Jacobsen SE (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PLoS One* 3(9). doi:10.1371/journal.pone.0003156.
43. Beeler SM, et al. (2014) Whole-genome DNA methylation profile of the jewel wasp (Nasonia vitripennis). *G3 (Bethesda)* 4(3):383–8.
44. Alves-dos-Santos I, Melo GAR, Rozen Jr JG (2002) Biology and Immature Stages of the Bee Tribe Tetrapediini (Hymenoptera : Apidae). *Am Museum Nat Hist* 3377:1–45.
45. Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* 162(1):156–159.
46. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR (2015) MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc* 10(3):475–483.
47. FASTQC developed by Babraham Bioinformatics. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

48. Hannon lab FASTX-Toolkit. Available at: http://hannonlab.cshl.edu/fastx_toolkit/index.html.
49. Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 38(12):1–7.
50. SEQYCLEAN developed by Ilya Zhbannikov.
51. Sadd B, Barribeau S, Bloch G (2015) The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol* 16(1):1–32.
52. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12(4):357–360.
53. Pertea M, et al. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33(3):290–295.
54. Grabherr MG, et al. (2013) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* 29(7):644–652.
55. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
56. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *Genome Announc* 2(4). Available at: http://arxiv.org/abs/1203.4802 [Accessed May 23, 2016].
57. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* 26(5):680–682.
58. Davidson NM, Oshlack A (2014) Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol* 15(7):410.
59. Hawkins ADK, Oshlack A, Davidson NM (2016) SuperTranscript : a reference for analysis and visualization of the transcriptome. *bioRxiv*. doi:10.1101/077750.
60. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–5.
61. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
62. García-Alcalde F, et al. (2012) Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28(20):2678–9.
63. Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R (2015) Annocript: A flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. *Bioinformatics* 31(13):2199–2201.
64. Suzek BE, et al. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6):926–32.
65. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
66. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(323). doi:10.1186/1471-2105-12-323.
67. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):1–34.
68. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
69. Krueger F (2012) Trim Galore. Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
70. Guo W, et al. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* 14(774). doi:10.1186/1471-2164-14-774.
71. Guo W, Zhu P (2017) CG-maptools: Command-line Toolset for Bisulfite Sequencing Data Analysis. Available at: https://cgmaptools.github.io.

Attachments

S1 - Supplementary File 1

Supplementary Material and Methods

Sample collection and sequencing

B. terrestris samples were collected from three colonies obtained from the commercial supplier Biobest® and kept in lab condition at Queen Mary University of London (England). Colonies were marked and housed in wood boxes attached to foraging arenas. After 16 days of adaptation all recently born workers in the colony received an individual number tag used for later identification. Bumblebee workers usually do not forage right after emergency (1), thus we waited for additional five days before observing their behaviour. After this period colonies were observed for one day during all its active foraging period (6 hours uninterruptedly). All the bees that remained inside of the nest for the whole day were considered nurses. In the following day after observations, between 10h – 12h, nurses and larvae in different developmental stages were collected and immediately frozen in liquid nitrogen.

T. angustula and *T. diversipes* were collected at the Laboratório de Abelhas (University of São Paulo – Brazil). Three *T. angustula* colonies, regularly kept in the lab were used. To collect *T. angustula* samples, brood cells were removed from the colonies. Larvae of different developmental stages were immediately collected from brood cells. Older brood cells, which contained adults closer to emergency, were transferred to an incubator with controlled temperature and humidity. Upon emergency female workers were marked with specific colours using a water based ink and immediately returned to the colony. Ten to twelve days after emergency of workers, colonies were opened and marked individuals were sampled. During this age worker bees from *T. angustula* present nursing behaviour (2). *T. diversipes* samples were collected directly from trap nests for solitary bees. For larvae collection nests under construction were accompanied to allow sampling of different larvae stages (3). Founders were collected in front of the trap nests while constructing their nests.

Other nine species were sampled in different locations (as listed in S1 – Table I). Highly eusocial bees were collected similarly to *T. angustula*, only *F. varia* and *F. schrottkyi* were not sampled using age control. For these species nurses were sampled

based on exoskeleton maturation, i.e. female workers with lighter exoskeleton were considered nurses (4).

RNA-Seq sequencing generated about 50 million paired reads (100bp) per sample (colony replicate) for the three main species and about 30 million paired reads (100bp) per sample for the remaining species. WBS returned 60-70 million single reads (150 bp) per each of the three main species.

Supplementary Tables

Table I – Sampling and behavioural information for the different bee species studied.

Species	Behaviour	Sample stage	Sample location	Transcriptome sequencing	Bisulfite sequencing
Tetragonisca angustula (Latreille, 1811)	highly eusocial	Adult / Larvae	São Paulo (SP) - Brazil	Yes	Yes - Nurse
Bombus terrestris (Linnaeus, 1758)	primitively eusocial	Adult / Larvae	London - United Kingdom	Yes	Yes - Nurse
Tetrapedia diversipes Klug, 1810	solitary	Adult / Larvae	São Paulo (SP) - Brazil	Yes	Yes - Founder
Neocorynura sp.	solitary	Adult	São Paulo (SP) - Brazil	Yes	No
Xylocopa frontalis Olivier, 1789	subsocial	Adult	Uberlandia (MG) - Brazil	Yes	No
Euglossa annectans Dressier, 1982	communal	Adult	São Paulo (SP) - Brazil	Yes	No
Friesella schrottkyi (Friese, 1900)	highly eusocial	Adult	Piracicaba (SP) - Brazil	Yes	No
Frieseomelitta varia (Lepeletier, 1836)	highly eusocial	Adult	Piracicaba (SP) - Brazil	Yes	No
Melipona bicolor Lepeletier, 1836	highly eusocial	Adult	São Paulo (SP) - Brazil	Yes	No
Nannotrigona testaceicornis Lepeletier, 1836	highly eusocial	Adult	São Paulo (SP) - Brazil	Yes	No
Plebeia remota (Holmberg, 1903)	highly eusocial	Adult	São Paulo (SP) - Brazil	Yes	No
Scaptotrigona aff. depilis (Moure, 1942)	highly eusocial	Adult	São Paulo (SP) - Brazil	Yes	No

Table II – Transcriptomic assembly data and statistics for all studied species. Mean coverage is given per sample; only one sample was sequenced for all species except *T. diversipes* (12 samples sequenced), *B. terrestris* (6 samples sequenced) and *T. angustula* (6 samples sequenced) – as detailed in Material and Methods section. In these three species mean coverage is from a representative sample (in this case an adult female sample pool). BUSCO Hymenoptera orthologous database has 4,415 genes, percentages refers to a ratio of this database.

Species	Number of transcripts	GC (%)	N50	Mean coverage (times)	Complete BUSCOs	Fragmented BUSCOs
Tetragonisca angustula (Latreille, 1811)	26,157	37.58	4,560	166.57	89.20%	8.70%
Bombus terrestris (Linnaeus, 1758)	20,777	36.43	5,432	214.22	89.00%	7.50%
Tetrapedia diversipes Klug, 1810	19,778	38.08	4,029	151.54	79.90%	10.00%
Neocorynura sp.	33,231	41.88	2,151	53.22	74.50%	11.20%
Xylocopa frontalis Olivier, 1789	39,992	40.51	2,172	80.35	75.20%	11.00%
Euglossa annectans Dressier, 1982	44,628	39.03	1,680	79.4	71.90%	14.30%
Friesella schrottkyi (Friese, 1900)	47,338	38.63	1,739	114.01	70.90%	15.10%
Frieseomelitta varia (Lepeletier, 1836)	44,362	39.14	1,909	90.53	76.30%	12.80%
Melipona bicolor Lepeletier, 1836	38,077	39.27	1,769	77.76	69.70%	12.70%
Nannotrigona testaceicornis Lepeletier, 1836	39,006	38.64	2,007	81.01	75.00%	13.00%
Plebeia remota (Holmberg, 1903)	41,112	39.68	1,450	79.09	65.30%	15.70%
Scaptotrigona aff. depilis (Moure, 1942)	44,461	39.65	989	108.9	54.50%	17.20%

Table III – Mean methylation in each DNA context. Values for cytosine (C) represents the global level of methylation in the data. Orthologous – all orthologous genes; BG – behavioural genes; OVER eusocial – BG highly expressed in the eusocial species; OVER solitary – BG highly expressed in the solitary bee.

DNA Context	Orthologous	BG	OVER eusocial	OVER solitary	Species
C	1.88%	2.05%	2.12%	2.02%	***T. diversipes***
CG	5.67%	4.44%	4.88%	4.09%	
CHG	1.12%	1.51%	1.56%	1.46%	
CHH	1.24%	1.60%	1.63%	1.61%	
C	1.02%	1.02%	1.04%	1.06%	***B. terrestris***
CG	3.40%	3.12%	3.22%	3.03%	
CHG	0.38%	0.45%	0.45%	0.50%	
CHH	0.43%	0.50%	0.51%	0.56%	
C	1.42%	1.51%	1.53%	1.60%	***T. angustula***
CG	4.44%	4.03%	4.12%	3.83%	
CHG	0.85%	1.00%	1.01%	1.13%	
CHH	0.90%	1.05%	1.06%	1.17%	

Table IV – Annotation data of the behavioural genes that presented a different level of expression in all nine bee species. Annotation was based on *B. terrestris* transcript sequences, since this bee is the only one with a complete genome available, among the ones analysed in this study. tblastx were performed against the NCBI non redundant database.

Orthologous	Description	Seq length	e-value	Similarity (%)	InterPro ID	InterPro GO IDs	InterPro GO names
NewId_1047	Bombus impatiens death-associated 1 (LOC100747804) mRNA	1203	0.00E+00	94	no IPS match	no IPS match	no IPS match
NewId_1620	Bombus impatiens 40S ribosomal S24 (LOC100743084) transcript variant misc_RNA	745	8.95E-136	96	G3DSA:3.30.70.330 (GENE3D); IPR001976 (PFAM); IPR001976 (PANTHER); IPR001976 (PRODOM); IPR012678 (SUPERFAMILY)	F:GO:0003735; C:GO:0005840; C:GO:0005622; P:GO:0006412	F:structural constituent of ribosome; C:ribosome; C:intracellular; P:translation
NewId_1797	Dufourea novaeangliae 60S ribosomal L8 (LOC107193484) mRNA	1301	1.55E-164	100	IPR002171 (SMART); IPR022669 (SMART); IPR002171 (PIRSF); IPR022666 (PFAM); IPR014722 (G3DSA:2.30.30.GENE3D); IPR022669 (PFAM); IPR014726 (G3DSA:4.10.950.GENE3D); G3DSA:2.40.50.140 (GENE3D); mobidb-lite (MOBIDB_LITE); IPR002171 (PANTHER); PTHR13691:SF28 (PANTHER); IPR008991 (SUPERFAMILY); IPR012340 (SUPERFAMILY)	F:GO:0003735; C:GO:0005840; C:GO:0005622; P:GO:0006412	F:structural constituent of ribosome; C:ribosome; C:intracellular; P:translation
NewId_1803	Apis florea forkhead box E1-like (LOC100866104) mRNA	4498	3.96E-19	76	IPR001766 (PRINTS); IPR001766 (SMART); IPR001766 (PFAM); IPR011991 (G3DSA:1.10.10.GENE3D); mobidb-lite (MOBIDB_LITE); PTHR11829:SF285 (PANTHER); PTHR11829 (PANTHER); IPR001766 (PROSITE_PROFILES); cd00059 (CDD); IPR011991 (SUPERFAMILY)	F:GO:0003700; P:GO:0006355; F:GO:0043565	F:transcription factor activity, sequence-specific DNA binding; P:regulation of transcription, DNA-templated; F:sequence-specific DNA binding
NewId_1912	Bombus impatiens uncharacterized LOC100744770 (LOC100744770) transcript variant mRNA	3943	0	94	no IPS match	no IPS match	no IPS match

NewId_2458	Apis dorsata 40S ribosomal S23-like (LOC102681136) mRNA	2022	1.24E-103	96	G3DSA:2.40.50.140 (GENE3D); IPR006032 (PIRSF); IPR006032 (PFAM); IPR005680 (TIGRFAM); mobidb-lite (MOBIDB_LITE); PTHR11652:SF30 (PANTHER); IPR006032 (PANTHER); IPR005680 (CDD); IPR012340 (SUPERFAMILY)	F:GO:0003735; C:GO:0015935; C:GO:0005840; C:GO:0005622; P:GO:0006412	F:structural constituent of ribosome; C:small ribosomal subunit; C:ribosome; C:intracellular; P:translation
NewId_2622	Apis mellifera elongation factor 1-alpha F2 (EF1a-F2) mRNA	2506	0	99	IPR000795 (PRINTS); IPR004160 (PFAM); PF00009 (PFAM); G3DSA:2.40.30.10 (GENE3D); G3DSA:2.40.30.10 (GENE3D); G3DSA:3.40.50.300 (GENE3D); IPR004161 (PFAM); IPR004539 (TIGRFAM); PTHR23115 (PANTHER); PTHR23115:SF218 (PANTHER); IPR004539 (HAMAP); IPR000795 (PROSITE_PROFILES); cd03693 (CDD); cd01883 (CDD); cd03705 (CDD); IPR009000 (SUPERFAMILY); IPR027417 (SUPERFAMILY); IPR009001 (SUPERFAMILY)	F:GO:0005525; F:GO:0003746; C:GO:0005737; F:GO:0003924; P:GO:0006414	F:GTP binding; F:translation elongation factor activity; C:cytoplasm; F:GTPase activity; P:translational elongation
NewId_2960	Bombus impatiens SUMO-conjugating enzyme UBC9-B (LOC105680548) mRNA	1313	0	98	no IPS match	no IPS match	no IPS match
NewId_3256	Bombus impatiens aldose reductase-like (LOC100748867) mRNA	4628	0	91	no IPS match	no IPS match	no IPS match
NewId_3259	Apis dorsata lethal(2)essential for life-like (LOC102670824) mRNA	2866	7.49E-97	96	IPR001436 (PRINTS); IPR002068 (PFAM); IPR008978 (G3DSA:2.60.40.GENE3D); mobidb-lite (MOBIDB_LITE); IPR031107 (PANTHER); PTHR11527:SF176 (PANTHER); IPR002068 (PROSITE_PROFILES); cd06526 (CDD); IPR008978 (SUPERFAMILY)	no GO terms	no GO terms

NewId_3308	Dufourea novaeangliae polyubiquitin (LOC107187242) transcript variant mRNA	2262	6.80E-151	98	IPR019956 (PRINTS); IPR000626 (SMART); G3DSA:3.10.20.90 (GENE3D); IPR000626 (PFAM); G3DSA:3.10.20.90 (GENE3D); G3DSA:3.10.20.90 (GENE3D); PTHR10666 (PANTHER); PTHR10666:SF217 (PANTHER); IPR000626 (PROSITE_PROFILES); IPR000626 (PROSITE_PROFILES); IPR000626 (PROSITE_PROFILES); cd01803 (CDD); cd01803 (CDD); cd01803 (CDD); IPR029071 (SUPERFAMILY); IPR029071 (SUPERFAMILY); IPR029071 (SUPERFAMILY)	F:GO:0005515	F:protein binding
NewId_3366	Bombus impatiens enolase (LOC100746870) mRNA	2592	0	99	IPR000941 (PRINTS); IPR020810 (SMART); IPR020811 (SMART); IPR020810 (PFAM); IPR029065 (G3DSA:3.20.20.GENE3D); IPR000941 (PIRSF); IPR000941 (TIGRFAM); IPR020811 (PFAM); IPR029017 (G3DSA:3.30.390.GENE3D); IPR000941 (PANTHER); PTHR11902:SF28 (PANTHER); IPR000941 (HAMAP); IPR000941 (CDD); IPR029017 (SUPERFAMILY); IPR029065 (SUPERFAMILY)	F:GO:0004634; F:GO:0000287; F:GO:0046872; P:GO:0006096; C:GO:0000015	F:phosphopyruvate hydratase activity; F:magnesium ion binding; F:metal ion binding; P:glycolytic process; C:phosphopyruvate hydratase complex
NewId_370	Habropoda laboriosa proteasome maturation (LOC108578983) mRNA	1486	3.40E-88	89	PF05348 (PFAM); IPR008012 (PANTHER)	P:GO:0043248	P:proteasome assembly
NewId_3728	Bombus terrestris 60S ribosomal L7a (LOC100646610) mRNA	1164	0	98	no IPS match	no IPS match	no IPS match
NewId_3886	Apis dorsata mitogen-activated kinase kinase kinase 15-like (LOC102671856) transcript variant mRNA	1040	1.02E-118	98	IPR003096 (PRINTS); IPR001715 (SMART); IPR000557 (PFAM); IPR001715 (G3DSA:1.10.418.GENE3D); IPR001715 (PFAM); PTHR18959 (PANTHER); PTHR18959:SF63	F:GO:0005515	F:protein binding

					(PANTHER); IPR000557 (PROSITE_PROFILES); IPR001715 (PROSITE_PROFILES); IPR001715 (CDD); IPR001715 (SUPERFAMILY)		
NewId_4049	Bombus impatiens 60S ribosomal L9 (LOC100741677) transcript variant mRNA	735	6.70E-135	99	IPR000702 (PIRSF); IPR020040 (PFAM); IPR020040 (G3DSA:3.90.930.GENE3D); IPR020040 (G3DSA:3.90.930.GENE3D); IPR000702 (PANTHER); PTHR11655:SF20 (PANTHER); IPR020040 (SUPERFAMILY); IPR020040 (SUPERFAMILY)	F:GO:0003735; F:GO:0019843; C:GO:0005840; P:GO:0006412	F:structural constituent of ribosome; F:rRNA binding; C:ribosome; P:translation
NewId_4078	Habropoda laboriosa apolipophorins (LOC108572790) mRNA	13058	0	81	IPR001747 (SMART); IPR015255 (SMART); IPR001846 (SMART); G3DSA:2.20.80.10 (GENE3D); G3DSA:1.25.10.20 (GENE3D); G3DSA:1.20.5.1230 (GENE3D); IPR009454 (PFAM); IPR001846 (PFAM); IPR001747 (PFAM); IPR015817 (G3DSA:2.20.50.GENE3D); IPR015255 (PFAM); IPR015816 (G3DSA:2.30.230.GENE3D); PTHR23345 (PANTHER); IPR001747 (PROSITE_PROFILES); IPR001846 (PROSITE_PROFILES); IPR015819 (SUPERFAMILY); SSF58113 (SUPERFAMILY); IPR015819 (SUPERFAMILY); IPR011030 (SUPERFAMILY)	P:GO:0006869; F:GO:0005319	P:lipid transport; F:lipid transporter activity
NewId_441	Dufourea novaeangliae DNA mismatch repair Msh2 (LOC107189119) mRNA	2315	2.14E-12	93	no IPS match	no IPS match	no IPS match
NewId_4774	Bombus terrestris 60S ribosomal L36 (LOC100644187) mRNA	584	4.27E-117	100	no IPS match	no IPS match	no IPS match
NewId_5492	Bombus terrestris plasminogen activator inhibitor 1 RNA-binding	3847	0	100	no IPS match	no IPS match	no IPS match

	(LOC100649628) mRNA						
NewId_5696	Apis cerana chitinase Idgf4 (LOC107996411) transcript variant mRNA	6852	0	92	IPR011583 (SMART); G3DSA:3.20.20.80 (GENE3D); IPR001223 (PFAM); IPR029070 (G3DSA:3.10.50.GENE3D); G3DSA:3.20.20.80 (GENE3D); PTHR11177:SF228 (PANTHER); PTHR11177 (PANTHER); cd02873 (CDD); IPR029070 (SUPERFAMILY); IPR017853 (SUPERFAMILY)	P:GO:0005975; F:GO:0008061	P:carbohydrate metabolic process; F:chitin binding
NewId_5939	Bombus impatiens cytochrome c iso-1 iso-2-like (LOC100740053) mRNA	1653	9.96E-32	72	IPR002327 (PRINTS); IPR009056 (G3DSA:1.10.760.GENE3D); IPR009056 (PFAM); PTHR11961:SF20 (PANTHER); IPR002327 (PANTHER); IPR009056 (PROSITE_PROFILES); IPR009056 (SUPERFAMILY)	F:GO:0009055; F:GO:0020037	F:electron carrier activity; F:heme binding
NewId_6040	Bombus impatiens paxillin (LOC100746469) transcript variant mRNA	9202	0	96	IPR002112 (PRINTS); IPR004827 (SMART); G3DSA:1.20.5.170 (GENE3D); G3DSA:1.10.880.10 (GENE3D); IPR004827 (PFAM); PTHR11462 (PANTHER); IPR004827 (PROSITE_PROFILES); cd14691 (CDD); SSF57959 (SUPERFAMILY)	F:GO:0003677; F:GO:0003700; P:GO:0006355; F:GO:0043565	F:DNA binding; F:transcription factor activity, sequence-specific DNA binding; P:regulation of transcription, DNA-templated; F:sequence-specific DNA binding
NewId_6278	Bombus impatiens uncharacterized LOC100747125 (LOC100747125) transcript variant mRNA	2574	0	97	IPR015897 (SMART); IPR004119 (PFAM); G3DSA:3.90.1200.10 (GENE3D); PTHR11012:SF37 (PANTHER); PTHR11012 (PANTHER); IPR011009 (SUPERFAMILY)	no GO terms	no GO terms
NewId_712	Bombus impatiens translation elongation factor 2 (LOC100750013) mRNA	5474	0	99	IPR000795 (PRINTS); IPR000640 (SMART); IPR005517 (SMART); IPR000640 (PFAM); G3DSA:3.40.50.300 (GENE3D); G3DSA:3.30.70.870 (GENE3D); IPR009022 (PFAM); IPR005517 (PFAM); IPR004161 (PFAM);	F:GO:0005525; F:GO:0003924	F:GTP binding; F:GTPase activity

					G3DSA:3.30.70.240 (GENE3D); G3DSA:2.40.30.10 (GENE3D); IPR005225 (TIGRFAM); IPR014721 (G3DSA:3.30.230.GENE3D); G3DSA:3.90.1430.10 (GENE3D); PF00009 (PFAM); PTHR42908 (PANTHER); PTHR42908:SF8 (PANTHER); IPR000795 (PROSITE_PROFILES); cd01681 (CDD); cd03700 (CDD); cd01885 (CDD); cd04096 (CDD); IPR020568 (SUPERFAMILY); IPR009022 (SUPERFAMILY); IPR027417 (SUPERFAMILY); IPR009000 (SUPERFAMILY); IPR009022 (SUPERFAMILY)		

Supplementary Figures

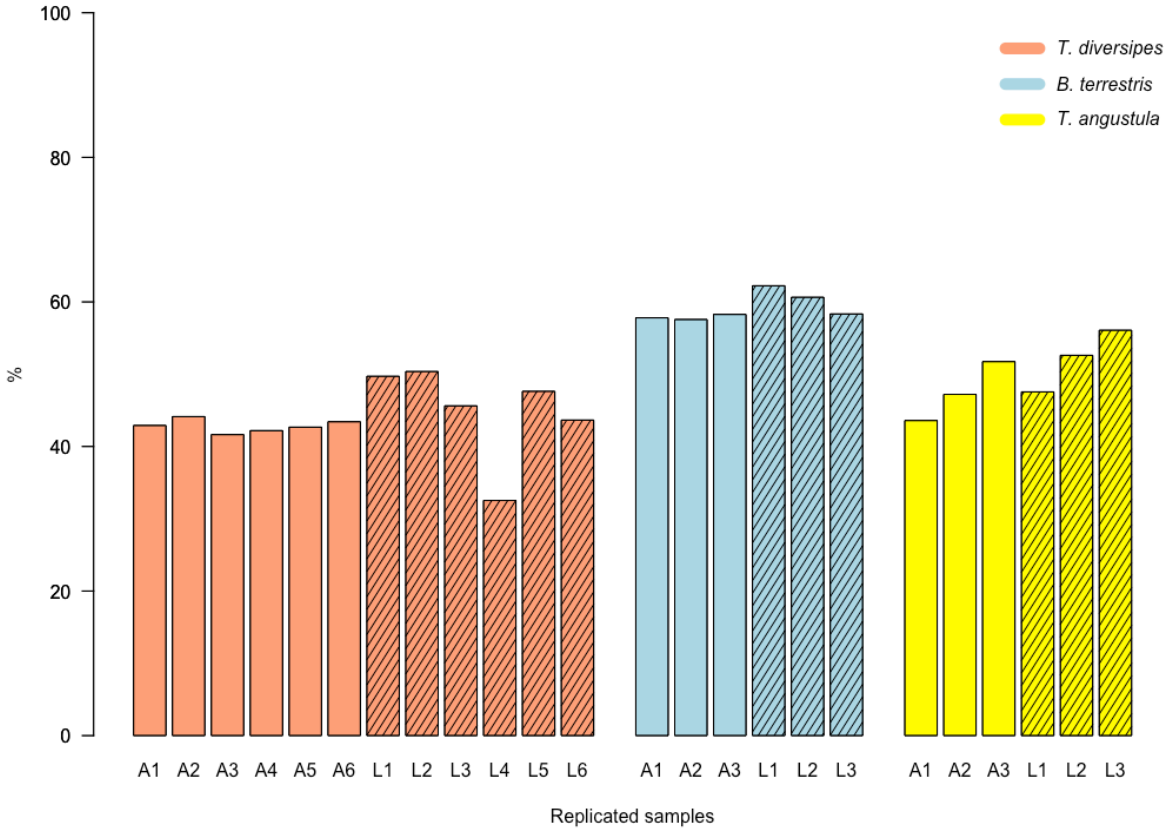


Figure 1 – Realignment ratio of cleaned reads of all replicated samples to the orthologous transcripts. A1-A6 indicate the replicated adult samples, and L1-L6 the replicated samples of larvae. Adults of *B. terrestris* and *T. angustula* were workers from the nurse subcaste, while *T. diversipes* adults were founder females. Two reproductive generations were used in *T. diversipes* sampling (see details in Material and Methods section).

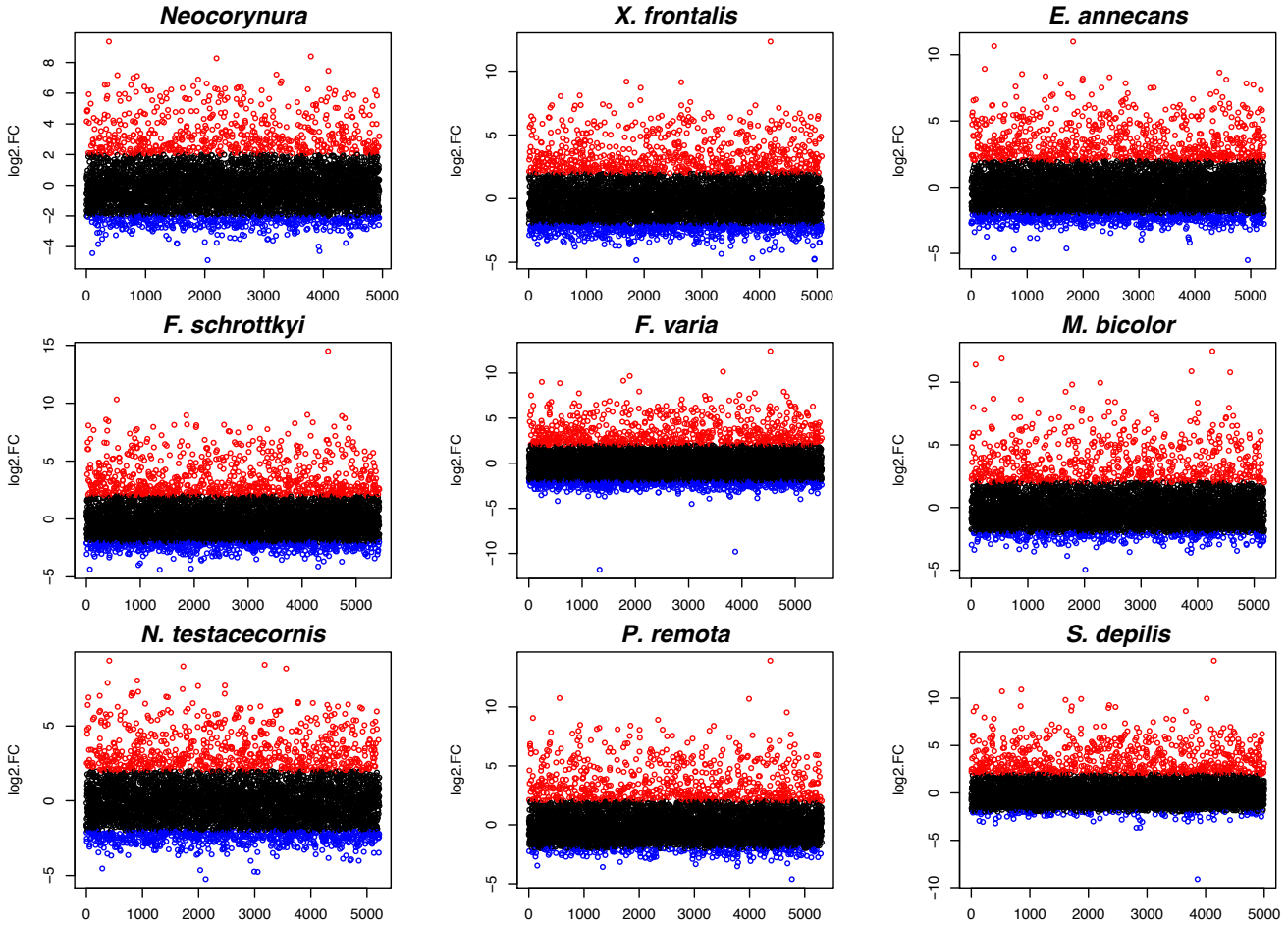


Figure 2 – Orthologous gene expression of the different bee species in log2 fold changes from the species mean expression level. Red – genes with expression at least 2 log2 fold changes greater than the species mean. Blue – genes with expression at least -2 log2 fold changes smaller than the species mean.

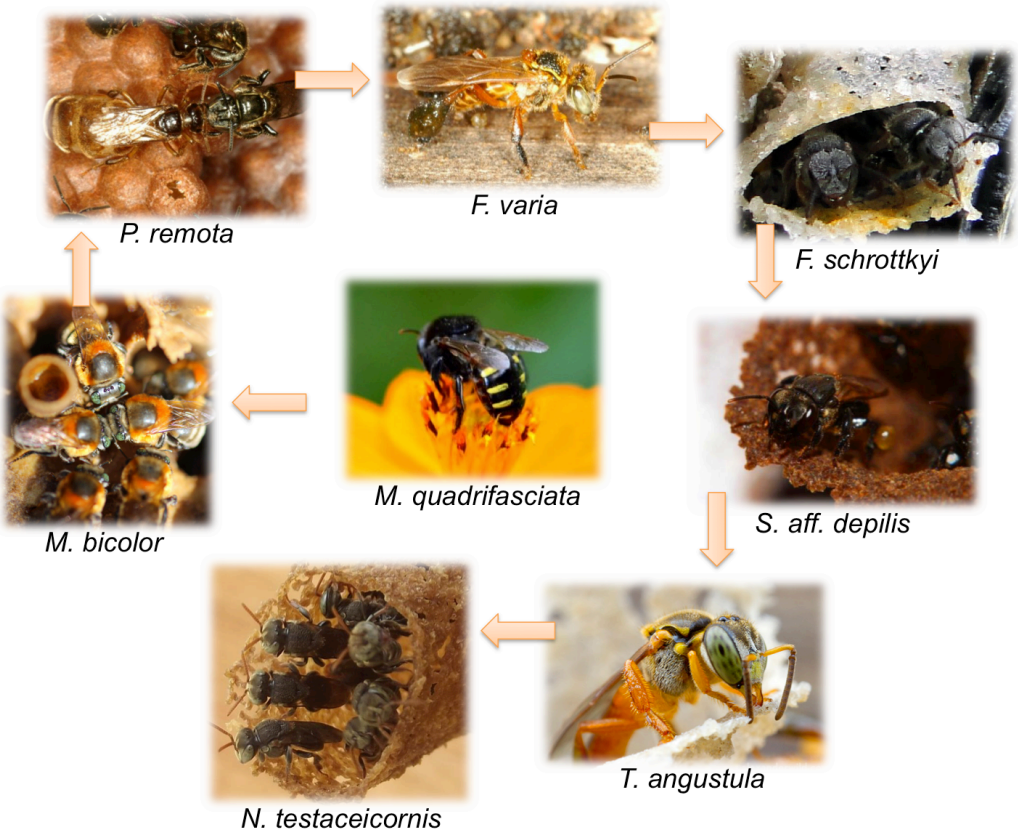


Figure 3 – Species of the tribe Meliponini used in this study and their genetic similarity based on the alignment rate of transcriptomic data to the reference genome of *M. quadrifasciata*. Alignment ratio is indicated by arrows in descending order, i. e. *M. bicolor* transcriptome had the greatest alignment ratio to *M. quadrifasciata* genome reference and *N. testaceicornis* had the smallest.

Photos credit: *M. quadrifasciata* (https://meliponariodamadecopas.blogspot.com.br/2014/12/curso-melipona-quadrifasciata-mandacaia.html); *M. bicolor* (http://www.abelhasemferrao.com/caixa-para-abelha-guaraipo-faca-voce-mesmo/); *P. remota* (https://hbjunior19.wordpress.com/2014/08/22/abelhas-sem-ferrao-descricao-das-especies-stingless-bees-description-of-species/); *F. varia* (https://www.cpt.com.br/cursos-criacaodeabelhas/artigos/abelhas-sem-ferrao-marmelada-amarela-frieseomelitta-varia); *F. schrottkyi* (Marcos Grangeiro); *S. aff. depilis* (Cristiano Menezes); *T. angustula* (Alex Wild); *N. testaceicornis* (Willian Henrique de Lima).

S2 to S5 follow at https://github.com/nat2bee/Suppl_PhDthesis

References cited

1. Woodard SH, Bloch GM, Band MR, Robinson GE (2014) Molecular heterochrony and the evolution of sociality in bumblebees (*Bombus terrestris*). *Proc R Soc B Biol Sci* 281(1780). doi:10.1098/rspb.2013.2419.
2. Koedam D, Tienen PGM Van (1997) The regulation of worker-oviposition in the stingless bee. *Insectes Soc* 44:229–244.
3. Alves-dos-Santos I, Melo GAR, Rozen Jr JG (2002) Biology and Immature Stages of the Bee Tribe Tetrapediini (Hymenoptera : Apidae). *Am Museum Nat Hist* 3377:1–45.
4. Elias-Neto M, et al. (2014) Heterochrony of cuticular differentiation in eusocial corbiculate bees. *Apidologie* 45(4):397–408.

Conclusões Gerais

As análises apresentadas neste trabalho trouxeram resultados de relevância metodológica e evolutiva. A técnica de RNA-Seq se mostrou uma poderosa ferramenta no estudo de diferentes características comportamentais em abelhas e, como foi exemplificado no Capítulo 1, esses dados podem ser analisados de diferentes maneiras para responder questões biológicas distintas. No referido capítulo, por exemplo, avaliamos as famílias de plantas possivelmente visitadas pela abelha solitária *Tetrapedia diversipes* e identificamos que este padrão é variável nas duas gerações reprodutivas. Durante a primeira geração reprodutiva, Amaranthacea e Euphorbiaceae foram as principais famílias de plantas visitadas. Enquanto que, durante a segunda geração, o principal recurso foram plantas da família Euphorbiaceae.

O padrão bivoltino, observado em *T. diversipes,* pode ser relevante para a evolução do comportamento social, tema central desta tese. Então, no Capítulo 2, utilizamos os dados de transcriptoma dessa espécie também para entender as diferenças de expressão entre cada geração. Com essas análises, identificamos que grande parte dos genes encontrados como diferencialmente expressos ocorrem entre adultos e não larvas. O que sugere que os níveis de expressão de genes maternos influenciam o tempo de desenvolvimento larval nessa abelha.

Uma das características mais estudadas da eussocialidade é a existência de operárias estéreis. No Capítulo 3, em busca de compreender melhor a divisão de trabalho nas abelhas operárias, comparamos os genes diferencialmente expressos entre nutrizes e forrageiras de duas espécies eussociais. Com estas análises identificamos um padrão interessante da eussocialidade: genes envolvidos na especialização de operárias são altamente específicos de cada linhagem mas atuam em processos biológicos comuns. Esse cenário sugere que a divisão de trabalho em operárias seja um processo que evolui posteriormente em diferentes linhagens sociais,

envolvendo os mesmos processos biológicos, formando um padrão mosaico de características únicas e compartilhadas a diferentes espécies.

Os resultados obtidos nas análises do Capítulo 3 foram importantes para o estudo dos mecanismos genéticos envolvidos na eussocialidade. No entanto, para responder a perguntas mais fundamentais do comportamento social, precisamos realizar análises comparativas mais abrangentes, incluindo espécies com comportamentos distintos em diferentes estágios do desenvolvimento. Para tal, no Capítulo 4, elaboramos uma método de análise que permitiu tais comparações e, com isso, identificamos 787 genes possivelmente envolvidos na origem do comportamento social das diferentes espécies de abelhas. Estes podem ser encontrados em múltiplas espécies e estão relacionados a funções biológicas relevantes para o comportamento, o que os torna potenciais candidatos para o *toolkit* social em abelhas. Ainda neste capítulo identificamos, com o sequenciamento bissulfito que, mais do que a quantidade, o contexto de metilação nos genes parece ser um fator importante para o comportamento social de abelhas.

Assim concluímos que o comportamento social é uma característica complexa que pode ser estudada por meio de diferentes abordagens de análises. Ao analisar a divisão de trabalho em operárias (Capítulo 3), por exemplo, podemos verificar os mecanismos moleculares envolvidos na manutenção do comportamento eussocial. Enquanto que com as análises do Capítulo 4, estudamos genes envolvidos com o surgimento do comportamento social. Ambas as abordagens são importantes para o estudo da evolução da socialidade, no entanto os padrões evolutivos em cada uma delas parece ser diferente. No primeiro caso, há uma pressão seletiva recente em genes distintos nas diferentes linhagens, assim o conceito de *toolkit* só pode ser aplicado aos processos biológicos envolvidos nesse comportamento. Já no segundo caso, verificamos que é sim possível a existência de um *toolkit* de genes específicos envolvido no comportamento de diferentes espécies, e que estes genes não estão envolvidos em alguns poucos processos biológicos específicos. Logo as respostas para as principais perguntas que guiaram esta tese podem variar de acordo com a abordagem de estudo utilizada.