Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Montagem e anotação do genoma de *Scaptotrigona postica*, uma importante abelha nativa sem ferrão

André Augusto Stella

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2023

André Augusto Stella
Engenheiro Agrônomo

Montagem e anotação do genoma de *Scaptotrigona postica*, uma importante abelha nativa sem ferrão

versão revisada de acordo com a Resolução CoPGr 6018 de 2011

Orientadora:
Profa. Dra. **MARIA IMACULADA ZUCCHI**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação

DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Stella, André Augusto

Montagem e anotação do genoma de *Scaptotrigona postica*, uma importante abelha nativa sem ferrão / André Augusto Stella - - versão revisada de acordo com a Resolução CoPGr 6018 de 2011. - - Piracicaba, 2023.

38 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. *Scaptotrigona postica* 2. Draft assembly 3. Genômica 4. Sequêncialonga 5. Sequência curta I. Título

AGRADECIMENOS

Agradeço primeiramente aos meus pais Raquel e Alexandre, e à minha irmã Stefane, que sempre estiveram ao meu lado, me amando e apoiando em toda minha jornada.

À minha orientadora, Professora Doutora Maria Imaculada Zucchi, pelo acolhimento, por me incentivar e guiar durante o mestrado.

Aos membros do Grupo de Genética e Genômica da Conservação, que além de grandes amigos muito me ensinaram e colaboraram no desenvolvimento deste projeto.

Ao Programa de Pós Graduação em Genética e Melhoramento de Plantas – ESALQ/USP, que me permitiu ter uma formação diferenciada e proporcionou um grande crescimento profissional.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que por dois anos e meio financiou o desenvolvimento deste projeto.

SUMÁRIO

RESUMO

Montagem e anotação do genoma de *Scaptotrigona postica*, uma importante abelhanativa sem ferrão

O declínio de abelhas, sejam elas nativas ou exóticas, é uma ameaça iminente aos serviços ecossistêmicos de polinização e, consequentemente, à segurança alimentar mundial e a preservação e conservação dos recursos naturais. Dada a importância desses polinizadores é necessário obter as informações sobre o genoma dessas espécies. A constante evolução de tecnologias, bem como as análises dos dados gerados, permite analisar um grande volume de dados com qualidade e precisão em um curto período e com menos custos financeiros. Neste contexto, a utilização de tecnologia de sequenciamento de leituras longas tem se apresentado de grande utilidade para geração de novos genomas. Ainda, destaca-se a abordagem híbrida de montagem de genoma com a união das tecnologias de sequenciamento de leituras curtas, a fim de diminuir as taxas de erro das leituras longas e aumentar a precisão da montagem. Além de uma montagem de genoma de boa qualidade, em estudos genômicos, a anotação funcionalé responsável por estabelecer uma correlação entre o sequenciamento e os estudos sobre a biologia dos organismos. O presente trabalho se propos a montar um *draft* e anotar o genoma da abelha *Scaptotrigona postica*, uma importante abelha nativa sem ferrão. Foram testados os montadores de genoma Flye e MaSuRCA, utilizando uma abordagem com apenas sequências longas (~26.000 pb) e uma híbrida, com sequências longas e curtas (50 pb). O tamanho do genoma foiestimado em 278.791.247 pb. Também foi feita a anotação de genes codificadores de proteínas. Usamos duas ferramentas de anotação funcional, EggNOG e Blast2GO, esta última auxiliada pelo programa Diamond, para verificar se há diferenças significativas entre elas com relação aos resultados, com dados gerado da predição de genes pelo software AUGUSTUS (15.168 genes preditos). A anotação funcional pelo blast2GO e Diamond apresentou aproximadamente 9.200 genes anotados funcionalmente e 15.168 de sequências proteicas anotadas pelo InterPro, enquanto o eggNOG gerou 9.906 genes anotados. Tais análisesbuscam fornecer o primeiro contato com o genoma desta espécie, a fim de fornecer importantes recursos para futuros estudos genéticos, assim como proporcionar um recurso valioso como referência para futuras pesquisas genéticas e evolutivas neste gênero.

Palavras-chave: *Scaptotrigona postica*, Draft assembly, Genômica, Sequência longa,Sequência curta

ABSTRACT

Genome assembly and annotation of *Scaptotrigona postica*, an important native stinglessbee

The decline of bees, whether native or exotic, is an imminent threat to pollination ecosystem services and consequently to world food security and the preservation and conservation of natural resources. Given the importance of these pollinators it is necessary to obtain information about the genome of these species. The constant evolution of technologies as well as the analysis of the generated data allows the analysis of a large volume of data with quality and precision in a short period and with less financial costs. In this context, the use of long read sequencing technology has proved to be very useful for generating new genomes. Also, the hybrid approach of genome assembly with the union of sequencing technologies for short reads stands out, in order to reduce the error rates of long *reads* and increase the accuracy of the assembly. In addition to good quality genome assembly, in genomic studies, functional annotation is responsible for establishing a correlation between sequencing and studies on the biology of organisms. The present work proposes to assemble a draft and annotate the genome of the bee *Scaptotrigona postica*, an important native stingless bee. Flye and MaSuRCA genome assemblers were tested, using both a long sequence approach and a hybrid approach with long (~26.000 pb) and short sequences (50 pb). Genome length was estimated of 278.791.247 bp. Annotation of protein-coding genes was also performed. We used two functional annotation tools, EggNOG and Blast2GO, the latter aided by the Diamond program, to verify if there are significant differences between them regarding the results, with data generated from gene prediction by the AUGUSTUS software (15.168 predicted genes). Functional annotation by blast2GO and Diamond presented approximately 9.200 functionally annotated genes and 15.168 protein sequences annotated by InterPro, while eggNOG generated 9.906 annotated genes. Such analyzes seek to provide the first contact with the genome of this species in order to provide important resources for future genetic studies, as well as providing a valuable resource as a reference for future genetic and evolutionary research in this genus.

Keywords: *Scaptotrigona postica*, Draft assembly, Genomics, Long read, Short read

1 INTRODUÇÃO

As abelhas são insetos pertencentes à Ordem Hymenoptera, a qual possui aproximadamente 130.000 espécies descritas e alguns dos insetos mais especializados no desenvolvimento de padrões e comportamentos complexos, relacionados ao comportamento social e fornecimento de nutrientes para a prole. São os principais agentes polinizadores presentes em grande parte do planeta, visto que, quando em busca de alimento realizam um papel de extrema importância na polinização cruzada. Sua importância para a manutenção e conservação do meio ambiente é imensurável, além de possuírem papel fundamental na segurança alimentar mundial de alimento. Os serviços ecossistêmicos de polinização relacionados à produção agrícola mundial foram valorados em aproximadamente US$235-577 bilhões. Sendo assim, os himenópteros são a ordem de insetos que mais beneficia os humanos (GILLOTT, 2005; POTTS *et al*. 2016)

Entretanto, o declínio dos polinizadores vem sendo relatado por diversos países, tendo como principais causas ações antrópicas, como aplicação intensiva de químicos em lavouras, abertura de novas áreas para plantio e crescimento de cidades sobre áreas verdes (MICHENER, 2013; FREITAS *et al.* 2016). Diante desse cenário, pesquisadores de diversos países têm se voltado ao estudo desses animais, envolvendo temas essenciais para o desenvolvimento de políticas públicas voltadas a conservação dessas espécies, como ecotoxicologia, morfofisiologia, estudos genéticos e taxonomia (SILVA *et al.* 2019). Esses estudos requerem um número significativo de recursos moleculares que inexistem atualmente,como estudos de expressão gênica diferencial e um genoma de referência de qualidade de diversas espécies de abelhas.

Em 2017 a agência federal dos Estados Unidos (*United States Environmental Protection Agency* - EPA) criou e implementou uma política paraproteger as abelhas de pulverização de inseticidas agrícolas, além de recomendar que os estados desenvolvam planos de proteção de polinizadores e melhores práticas de manejo. Também divulgou, em 2020, decisões provisórias sobre a utilização dos neonicotinóides acetamiprida, clotianidina, dinotefurano, imidacloprida e tiametoxam. O foco dos estudos recai sobre a *Apis mellifera*, espécie exótica.

A montagem de um genoma que possa ser utilizado como referência permitirá que inúmeros estudos, como o de expressão gênica diferencial, metabolômica e proteômica sejam realizados, possibilitando uma compreensão holística de as abelhas são influenciadas por ambientes estressantes, o que permitirá que planos de conservação possam ser melhor desenvolvidos para proteção dessa espécie e de outras abelhas. Diante disso, este trabalho teve como objetivo construir um *draft* do genoma *de novo* da abelha *S. postica* utilizando dois

métodos de montagem, com o software Flye, somente com as sequências longas e com o software MaSuRCA, montagem híbrida utilizando sequências longas e curtas. Também, usamos duas ferramentas de anotação funcional, egg-NOG e Blast2GO, esta última auxiliada pelo programa Diamond, para verificar se há diferenças significativas entre elas com relação aos resultados necessários para responder questionamentos biológicos de pesquisas futuras.

A falta de informação genética não permite entender processos biológicos importantes, como a tradução de proteínas em um ambiente contaminado com agrotóxico. Para a condução desse trabalho, a abelha *S. postica* foi escolhida como a espécie-alvo de estudo, considerando a importância que abelhas nativas apresentam para os ecossistemas brasileiros e como o seu declínio pode resultar em perdas de produtividade na área agrícola, tornando a produção de alimento, como frutas e vegetais, inferior à demanda de consumo mundial(GALLAI *et al*. 2009). Ademais, frente ao declínio das colônias que está ocorrendo em diversas partes do mundo, esse recurso também pode ser utilizado para uma melhor compreensão das ameaças à conservação das abelhas (LOZIER e ZAYED, 2017).

2 REVISÃO DE LITERATURA

2.1 Aspectos gerais sobre as abelhas

2.1.1 Importância das abelhas

Diversos cultivos de elevado valor econômico dependem da polinização para o aumento da produção de frutos e sementes ou para melhorar sua qualidade. De 1.330 cultivos tropicais polinizados por animais, 70% melhoraram e/ou aumentaram a produtividade quando polinizados corretamente (ROUBIK, 2018). Dentro desse contexto, no Brasil o valor econômico relacionado a polinização feita por insetos foi de US$12 bilhões em 2015, o que reforça o papel fundamental dos polinizadores na economia e no agronegócio (GIANNINI *et al*. 2015). Diversos fatores têm ocasionado o declínio das abelhas no Brasil, sejam elas exóticas ou nativas, como a redução e perda de habitat, poluentes, patógenos, competição por recursos e práticas agrícolas, como o uso de agrotóxicos (KLEIN *et al*. 2017; PIRES *et al*. 2016). Alguns desses fatores também podem afetar o seu sistema cognitivo, prejudicando a capacidade de aprendizagem, memória e navegação, o que acaba diminuindo o desempenho do forrageamento e impacta o desenvolvimento das abelhas e a sobrevivência da colônia (POTTS *et al*. 2010).

Estudos recentes demonstram uma relação direta entre o uso de agrotóxico e a perda de colônias e biodiversidade. Os princípios ativos mais utilizados nas lavouras do Brasil, como o fipronil e o imidacloprido, aparecem em diversos estudos como alguns dos maiores responsáveis pela morte das abelhas nos últimos anos (NOCELLI *et al*. 2012). Estima-se que do final de 2018 ao início de 2019 o fipronil foi responsável pela morte de mais de meio bilhão de abelhas no sul do Brasil (SILVA *et al*. 2021). Através da análise das mudanças na expressão gênica de espécimes de *Apis mellifera* em exposição ao fungicida piraclostrobina, foi observado que esse químico atua no organismo das abelhas desregulando o metabolismo da glicose e da respiração celular. Uma revisão meta-analítica com 154 conjuntos de dados, de 1951 a 2019, confirmou uma relação entre a morte de *A. mellifera* e o uso de agrotóxicos (SAMPAIO, 2020).

Abelhas são insetos pertencentes a ordem Hymenoptera e família Apidae. Dentro dessa família existe a tribo Meliponini, que constitui um grupo de abelhas eussociais. Fazem parte da história brasileira, junto aos povos originários das Américas, que desde antes da colonização, criam abelhas nativas usando potes de barro e cabaça como colmeia(CORTOPASSI *et al*. 2006). Conhecidas popularmente como abelhas sem ferrão (ASF), sãocaracterizadas por possuírem ferrão vestigial ou atrofiado, o qual não é usado para defesa. Como estratégia de proteção se enroscam nos pelos dos animais, também depositam resina na superfície da colmeia para impedir a entrada de invasores. Algumas espécies produzem uma substância cáustica que

é colocada sobre seu corpo. O tamanho populacional de suas colônias varia de centenas até milhares de indivíduos, dependendo da espécie. Possuem uma grande diversidade morfológica, com tamanho variando de 2 a 13mm. Existem cerca de 600 espécies no mundo, das quais 330 ocorrem no Brasil. Estas espécies apresentam uma grande complexidade de fenótipos, o que dificulta a delimitação de espécies. Estão presentes na América do Sul, América Central, Ásia, Ilhas do Pacífico, Austrália, Nova Guiné e África (ROUBIK, 2006). São responsáveis pela polinização de 30% das espécies da Caatinga e Pantanal e até 90% das espécies da Mata Atlântica (VENTURIERI *et al*. 2003).

O mel destas abelhas é muito apreciado e seu comércio regional traz um complemento financeiro importante para populações da região Norte e Nordeste do Brasil. As abelhas sem ferrão produzem mel com uma composição físico-química diferente do mel de *Apis mellifera*, apresentando diferentes características de sabor, cor e odor. Para a escolha da espécie, o produtor deve levar em consideração fatores como a ocorrência natural da região e adaptação ao manejo, sendo algumas de difícil domesticação. A disponibilidade de plantas produtoras de pólen e néctar que servirão de alimento deve ser levada em consideração na escolha do local de instalação do meliponário. As espécies mais recomendadas para criação segundo a Empresa Brasileira de Pesquisa e Agropecuária (EMBRAPA) são a jataí (*Tetragonisca angustula*), a uruçu (*Melipona scutellaris*), a tiúba (*Melipona fasciculata*), a jandaíra (*Melipona subnitida*), a uruçu-cinzenta (*Melipona manaosensis*), e a mandaçaia (*Melipona quadrifasciata anthidioides*) (PEREIRA *et al*. 2017).

2.1.2 Espécies do gênero *Scaptotrigona*

O gênero *Scaptotrigona* é formado por 8 espécies *S. affabra*, *S. bipunctata*, *S. depilis*, *S. fulvicutis*, *S. polysticta*, *S. tubiba*, *S. xanthotricha* e *S. postica* (MENEZES, 2014). Dentre elas, *S. postica* (Latreille, 1807), também conhecida como “mandaguari”, destaca-se por ser uma importante abelha nativa sem ferrão (Figura 2),polinizadora de diversas espécies de plantas exóticas e cultivares comerciais. A colônia desta espécie possui uma rainha-mãe e de 6 a 10 mil operárias, as quais vivem cerca de 40 dias e atingem um raio de voo de 0.9 Km (WILLE, 1983)

Figura 2. *Scaptotrigona postica* (Retirado de A.B.E.L.H.A., 2021)

Assim como as outras abelhas da tribo Meliponini, *S. postica* constrói os ninhos mais elaborados dentre as abelhas. Possuem o hábito de nidificar em ocos de árvores e produz com cera células de crias horizontais e sobrepostas, que são cercadas por camadas de resina. A entrada da colônia apresenta forma de tubo (Figura 3A) e é construído com cerúmen, mistura de cera e resina. As diversas camadas de cerúmen ao redor da câmara de criação dão origem ao invólucro (Figura 3B), onde estão os potes de armazenamento de alimentos. Para proteção essas abelhas colocam uma substância nas paredes do ninho chamada betume, geralmente formada de cera e barro (Figura 3 C) (WILLE, 1973; MICHENER, 2000).

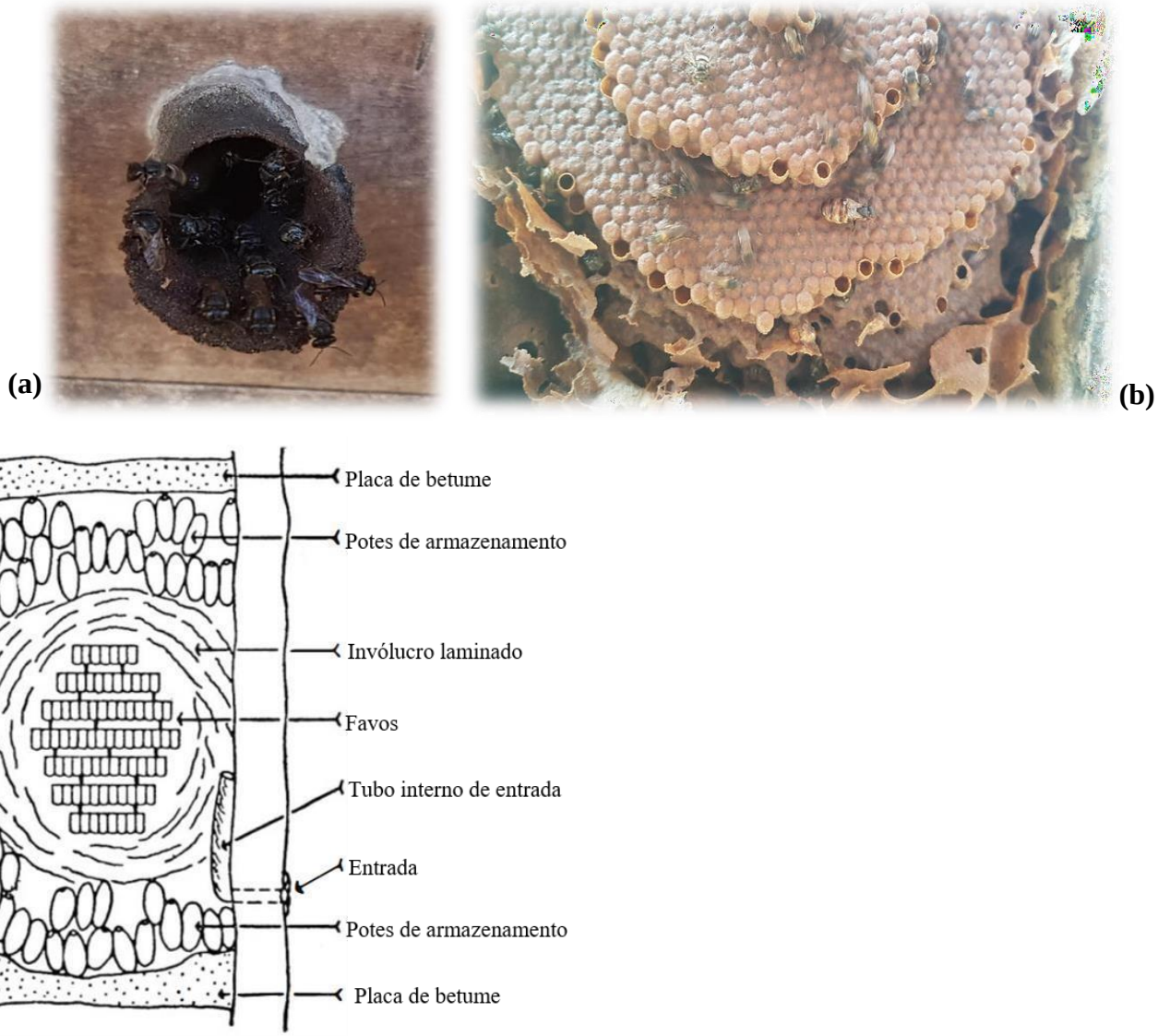

Figura 3. (a) Entrada da colônia da *Scaptotrigona postica*. Formato de cone, característica de algumas espécies de abelhas sem ferrão (Autoria própria, 2021)**. (b)** Abelha rainha da espécie *S. postica*, no interior da colônia, ao redor há cinco níveis de células, com operárias ao lado (Autoria própria, 2021). **(c)** Diagrama de um ninho de meliponídeos (Retirado de Wille, 1973).

2.2 Recursos genômicos em abelhas

2.2.1 Genomas disponíveis

Manrique e Soares (2002) iniciaram um programa de seleção de abelhas, visando o aumento na produção de própolis e mel. Analisando 100, colônias observaram correlação entre a produção dos dois produtos, sendo possível selecionar abelhas para aumentar a produção de própolis e para melhorar a produtividade de mel. Um trabalho realizado com *Melipona quadrifasciata* e outras 10 espécies identificou assinaturas genômicas relacionadas à mudança do comportamento solitário para o social. Com sequenciamento de alto rendimento, foi

encontrado o RNA de um vírus potencialmente prejudicial presente em abelhas melíferas, fornecendo informações a respeito dos padrões evolutivos de dispersão e seleção nesses organismos (CORNMAN *et al*. 2013).

O sequenciamento do genoma nuclear e mitocondrial da *Frieseomelitta varia* permitiu a identificação de um alto grau de sintenia em um bloco de genes presente em *Apis mellifera*, relacionado ao comportamento social de estocagem de pólen, indicando um alto grau de ancestralidade entre essas espécies. O genoma mitocondrial de *F. Varia* compreende 15.144 pb, codificando 13 proteínas, 22 tRNAs e 2 rRNAs. Nesse trabalho, também foram identificados muitos componentes repetitivos do genoma e 1.946 RNAs longos não codificantes. Essas informações podem fornecer a base molecular para a plasticidade reguladora de genes, pois as sequências podem estar ligadas a morte celular programada dos ovários durante o desenvolvimento da pulpa, que torna as operárias estéreis (FREITAS *et al*. 2020).

Entre cerca de 120 mil espécies de himenópteros, existem cerca de 460 *assemblies* de 306 espécies depositados no *National Center for Biotechnology Information* (NCBI), sendo que apenas 81 espécies possuem montagens à nível cromossômico. Adicionalmente, dentro da família *Apidae*, existem genomas de somente 60 espécies sequenciados que abrangem montagens em nível de *contigs* e cromossomos. Quando consideramos apenas abelhas sem ferrão, a lacuna de informação é ainda maior, uma vez que no NCBI atualmente apenas nove espécies desse grupo com o genoma depositado no GenBank, sendo elas *Lepidotrigona ventralis*, *Heterotrigona itama*, *Tetragonula mellipes*, *Tetragonula hockingsi*, *Tetragonula davenporti*, *Tetragonula clypearis*, *Tetragonula carbonaria*, *Melipona quadrifasciata* e *Frieseomelitta varia*. Destas espécies, são nativas a *F. varia* e *M. quadrifasciata*. Na tabela 1 são mostrados alguns recursos genômicos disponíveis até o momento. Asinformações obtidas através desses trabalhos abrem caminhos para novos estudos genômicose até futuras pesquisas na área de melhoramento genético, visando a identificação de características de interesse econômico (FREITAS *et al*. 2020).

Tabela 1. Resumo dos recursos genômicos disponíveis para abelhas.

RECURSO GENÔMICO	DESCRIÇÃO E REFERÊNCIA
Genomas	▪ *Assembly* a nível de cromossomo, *Apis mellifera* (Uppsala University, 2018) ▪ Nível de cromossomo, *Bumbus terrestris* (Sanger Institute, 2022) ▪ *Draft assembly* a nível de *scaffold*, *Ceratina calcarata* (York University, 2022) ▪ Nível de *scaffold*, *Frieseomelitta varia* (University of São Paulo, 2020) ▪ Nível de *scaffold*, *Melipona quadrifasciata* (Beijing Genomics Institute, 2015) ▪ Nível de *contig*, *Nomada fucata* (Sanger Institute, 2023) ▪ Nível de *contig*, *Tetragonula carbonária* (The University of Queensland, 2020)
Genoma mitocondrial	▪ *Apis mellifera* (Crozier RH *et al*. 1993) ▪ *Apis cerana* (TAN *et al*. 2011) ▪ *Frieseomelitta varia genome* (Universidade de São Paulo, 2020)
Transcriptomas	▪ *Apis cerana cerana*, montados a partir de *short reads*, RNA-Seq (WANG, 2012)
Mapa genético	▪ *Apis mellifera*, mapa de ligação obtido a partir de marcadores moleculares RAPD (HUNT *et al*. 1995) ▪ *Apis mellifera*, mapa de ligação baseado em microssatélites de terceira geração (SOLIGNAC *et al*. 2007)
Bibliotecas de BACs	▪ *Apis mellifera* (TONKINS *et al*. 2002)

2.2.2 Sequenciamento e montagem de genomas

Os crescentes avanços na área da genômica e a diminuição dos custos envolvendo o sequenciamento de novos genomas permitiram o desenvolvimento de novas pesquisas, possibilitando um entendimento mais profundo a respeito da diversidade genética, elementos móveis no genoma, fluxo gênico, consanguinidade, gargalos populacionais, entre outras informações importantes para uma melhor compreensão das espécies (LOZIER; ZAYED, 2017). As informações obtidas através do uso dessas tecnologias permitiram expandir o conhecimento em diversas áreas da ciência. Estudos genômicos, como o sequenciamento

completo de genomas, transcriptomas e metagenomas, têm sido ferramentas úteis para aprofundar o conhecimento científico sobre a biologia das abelhas e corroborar os esforços para a conservação (FREITAS *et al*. 2020).

Além disso, esses avanços possibilitaram uma maior precisão dos dados gerados, reduzindo a taxa de erro associada. Os avanços tecnológicos na área das ômicas têm permitido que os mais diferentes organismos sejam estudados de forma holística, associando informações genéticas a características fenotípicas observadas quando submetidos a determinadas condições, possibilitando que um sistema complexo, como o organismo, possa ser entendido mais profundamente (SLATKO, 2018). Dentre as ômicas possíveis de serem abordadas está a genômica que se caracteriza por permitir o estudo do conteúdo completo do DNA, incluindo todos os seus genes, o que possibilita a predição de possíveis proteínas que podem estar sendo codificadas por esses genes, bem como o desenvolvimento de marcadores moleculares baseados em SNPs (polimorfismo de nucleotídeo único) e a possibilidade de identificar regiões candidatas para a condução de outros estudos (VAILATI, 2017). No entanto, a montagem de genomas, principalmente de organismos ainda não sequenciados, é uma tarefa desafiadora, por isso compreender como funcionam as ferramentas de sequenciamento, montagem e anotação desses genomas pode ser crucial para que dados mais precisos e confiáveis possam ser gerados (ZHANG *et al*. 2020).

Nos últimos anos, a utilização de sequenciamento de sequências longas tem sido essencial para gerar novos genomas, assim como a melhoria de genomas já sequenciados. O uso apenas de sequências curtas para a montagem de genomas complexos é um limitante, assim como o tempo para o processamento dos dados considerando as limitações fornecidas pelo uso de sequênciascurtas, dentre as quais o processamento de fragmentos repetitivos e o tempo para o processamento dos dados (ZHANG *et al*. 2020). No entanto, a utilização de sequências longas também fornece limitações, como taxas de erros relativamente altas, por conta dos erros de sequenciamento. Por isso, sequências curtas podem ser utilizadas em associação as sequências longas em montagens híbridas para fornecer um genoma mais preciso, com menos lacunas e menor taxa de erro (ZIMIN *et al*. 2017).

Duas classes de algoritmos amplamente utilizados são *Overlap–Layout–Consensus* (OLC) e os baseados em Grafos de de Bruijn (GdB). O OLC divide a montagem em três módulos: através de uma comparação par a par é feita a identificação de sobreposições entre as *reads* (O), seguido dos alinhamentos iniciais das sequências, chamados de *unitigs*, são a primeira etapa do montador. Em seguida é construído um layout (L) em um gráfico contendo as informações lidas e sobrepostas e *unitigs* são combinados em *contigs*, que por sua vez são

agrupados em *scaffolds*. Por fim, é montada a sequência de consenso (C), ou *assembly* (MYERS *et al*. 2000).

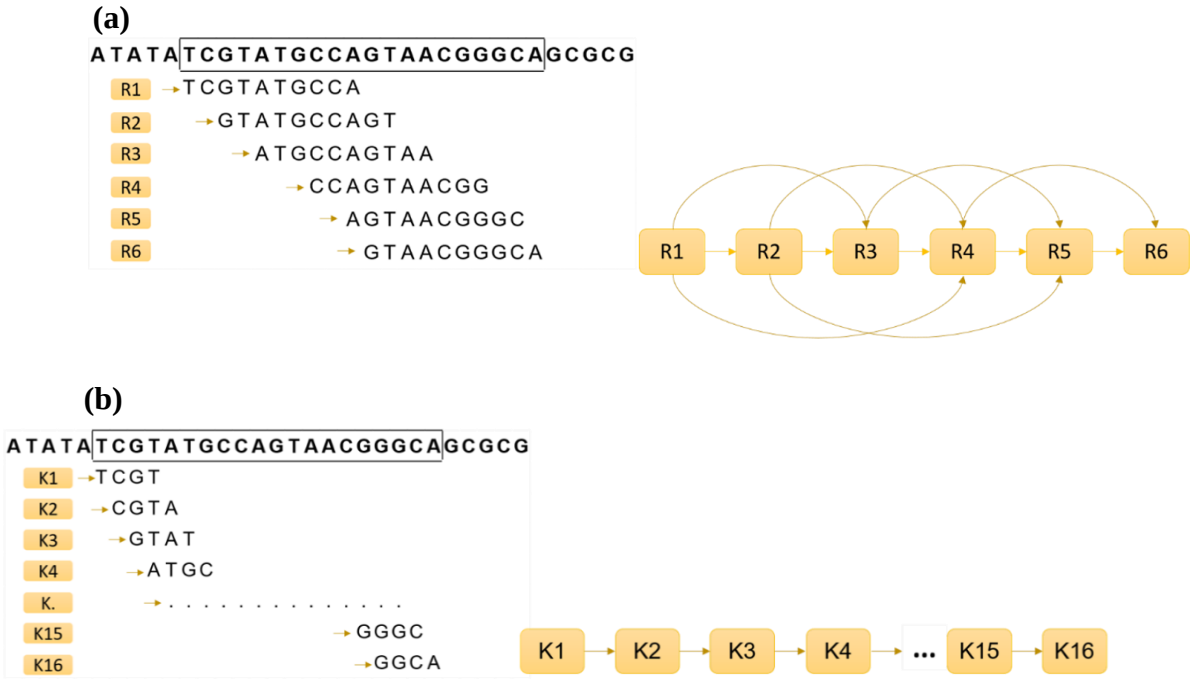

Figura 5: Abordagens OLC e GbB aplicadas em algoritmos de montagem de genoma. **(a)** Para a região delimita foram gerados seis fragmentos de leitura com tamanho de 10 pb. Estão ordenados de acordo com o grafo OLC ilustrado. Neste paradigma vários nós apresentam mais de uma aresta de saída e de chegada. **(b)** GdB montado com 16 *k-mers* de comprimento 4 pb. A maior parte dos nós possuem apenas uma aresta de saída (Adaptado de: LI, 2011)

Diversas ferramentas para montagem de sequências oriundas de métodos de sequenciamento de segunda geração (NGS) baseiam-se em GdB (COMPEAU, 2011). Algoritmos que usam essa abordagem dividem as leituras em fragmentos de tamanho muito menor com o mesmo comprimento (*k-mers*). Com os *k-mers* é construido um GdB que é usado para inferir o genoma. Para a escolha do melhor algoritmo a ser usado é preciso levar em consideração o consumo computacional de tempo e memória, uma vez que realizar todasas comparações possíveis em uma grande quantidade de dados terá um alto custo. Abordagens com o OLC são mais adequadas para leituras longas de baixa cobertura, enquanto com GbD possuem melhor desempenho com leituras curtas de alta cobertura (LI, 2011).

Outra etapa importante em estudos genômicos é a anotação funcional de genes, responsável por estabelecer uma ponte entre a sequência e a biologia do organismo, capaz de identificar características importantes do genoma, como os genes e seus produtos, permitindo discussões e conclusões biológicas relevantes para a ciência. É importante destacar que as

ferramentas e recursos disponíveis para anotação do genoma estão se desenvolvendo e se difundindo rapidamente, permitindo anotações cada vez de maior qualidade (STEIN, 2001; YANDELL, 2012). Existem várias ferramentas disponíveis para realizar a anotação funcional de genes e compreender o que essas ferramentas proporcionam com relação aos dados gerados é essencial para escolher a que melhor responderá as perguntas biológicas de cada pesquisa.

3 MATERIAL E MÉTODOS

3.1 Sequenciamento e controle de qualidade

Para a obtenção das sequências longas foi utilizado 0.5 g de um *pool* genético de cinco machos de *S. postica* provenientes de uma colônia saudável foi submetido ao procedimento de extração de DNA. Foi usado o kit blood and tissue da Quiagen, adaptado de um protocolo para Drosophila. Foram selecionados os fragmentos acima de 10 kpb, com o Circulomics. O controle de qualidade foi realizado com o NanoDrope, Qubit 4 e gel de agarose a 0.7%. Os fragmentos extraídos foram de alto peso molecular, necessário para a montagem das bibliotecas de DNA de sequências longas, que foram construídas usando o kit SQK-LSK109 (Oxford-Nanopore - ONT), de acordo com instruções do fabricante. O sequenciamento foi feito em *flow cell* R.9.4.1 do tipo FLO-MIN106 em sequenciador MinION. Foram realizadas quatro corridas com a mesma *flow cell*. O trabalho foi realizado no Departamento de Genética da Esalq/USP.

Para a obtenção das sequências curtas foi feita a extração de DNA de 1 indivíduo macho utilizando protocolo CTAB. O controle de qualidade foi realizado com o NanoDrope, Qubit 4 e gel de agarose a 1%. As bibliotecas de DNA para a obtenção de sequências curtas foram construídas conforme recomendações do fabricante e utilizando-se de kit DNA Prep, sendo posteriormente encaminhadas para o laboratório Central de Tecnologia de Alto Desempenho da Universidade de Campinas (LaCTAD), onde foram sequenciadas em plataforma Illumina NextSeq1000/2000.

A qualidade das *short reads* foi aferida com o software FastQC v.0.11.8 e para avaliar a qualidade das ONT *long reads* foi utilizado o programa LongQC -1.2.0c (ANDREWS *et al*. 2010; FUKASAWA *et al*. 2020). A métrica *Phred quality score* (*Q score*) foi utilizada para avaliar a acurácia do sequenciamento e os critérios avaliados foram a qualidade da sequência por base e conteúdo da sequência por base. As *reads* não passaram por filtro, uma vez que a qualidade dos dados foi considerada aceitável.

A qualidade das montagens realizadas tanto pelo montador Flye quanto pelo MaSuRCA foi realizada com o auxílio do software Quast v.5.2.0, o qual avaliou os dados com base em diferentes parâmetros, dos quais adotamos para este trabalho: a) tamanho total da montagem, b) número total de *contigs*, c) *contigs* > 25 k, d) *contigs* > 50 k, e) tamanho do maior *contig* (pb), f) %GC, g) N50 dos *contigs* e h) L50 dos *contigs*. As montagens também foram submetidas a análises no software BUSCO, que emprega conjuntos de *Benchmarking Universal Single-Copy Orthologs* para fornecer medidas quantitativas da completude dos genomas, no qual nós utilizamos o banco de dados de Hymenoptera do OrthoDB como referência.

O tamanho do genoma foi estimado pelo GenomeScope v.2.0, o qual utiliza o

programa de linha de comando Jellyfish para estabelecer um modelo matemático de como as frequências de *k-mer* serão distribuídas no genoma. Um *k-mer* é uma sequência decomprimento *k,* e contar as ocorrências de todas essas sequências é uma das etapas centrais em muitas análises da sequência de DNA (VURTURE *et al*. 2017).

3.2 Montagem do genoma

- **Montagem do genoma de forma híbrida**

Leituras longas e curtas podem ser usadas em conjunto numa estratégia de montagem híbrida (Figura 6), em projetos de sequenciamento. Softwares que usam sequências homogêneas estão mais propensos a ter um desempenho ruim em dados híbridos, sendo necessário realizar ajustes nos montadores (MILLER *et al.* 2008). Abordagens de sequenciamento desse tipo aproveitam os pontos fortes de duas ou mais plataformas de sequenciamento (HALL, 2007).

Para a montagem do genoma *de novo* de *S. postica* usando da combinação de sequências curtas obtidas pelo sequenciamento Illumina e longas pela tecnologia Nanopore foi utilizado o conjunto de ferramentas de análise e montagem do genoma denominado de *Maryland Super-Read Celera Assembler* (MaSuRCA) v.4.1.0. (ZIMIN *et al*. 2013). O MaSuRCA utiliza o software Celera, que combina os benefícios do grafo de Brujin e abordagens de montagem *Overlap-layout-Consensus*, permitindo montagem híbrida com *long* e *short reads* minimizando os erros associados às leituras longas e produzindo uma montagemmais concisa e com menos lacunas.

- **Montagem do genoma usando sequências longas**

Para a montagem do genoma *de novo* de *S. postica* a partir somente de sequências longas obtidas por sequenciamento usando tecnologia Nanopore foi utilizado o software Flye v.3.10.6 (LIN *et al*. 2016), o qual foi desenvolvido para manuseio de leituras de sequenciamento de moléculas únicas como as produzidas pela PacBio e Oxford Nanopore (KOLMOGOROV *et al*. 2019) e gera como arquivos de saída: 1) Montagem final em formato .fasta, contendo *contigs* e possivelmente *scaffolds;* 2) Grafo de repetição final em formato.gva e 3) informações extras sobre *contigs* como comprimento e cobertura em formato .txt.

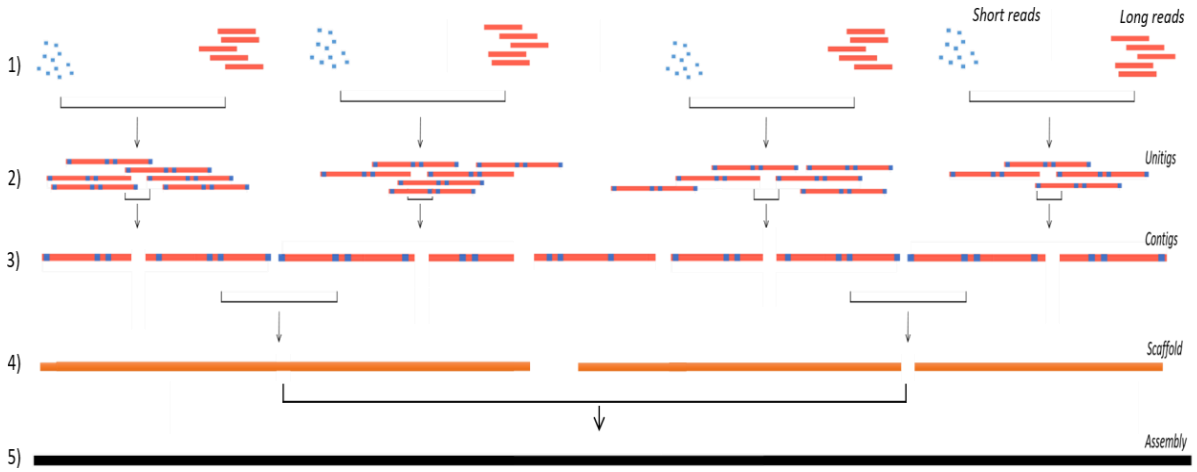

Figura 6: Esquema do processo de construção de um *assembly*. 1) *Short reads* estão representadas em azul e *long reads* em vermelho. 2) Construção de *unitigs* a partir de sequências híbridas e identificação de sobreposição entre elas. 3) Montagem de *contigs* para *scaffolds*. 4) Tentativa de preenchimento das lacunas (*gap*). 5) Por fim,é montada a sequência consenso.

3.3 Análise BUSCO e predição de genes

A predição dos genes foi realizada pelo software Augustus v3.2.2 (STANKE, 2006), que prevê genes em sequências genômicas eucarióticas. Para predição de *S. postica*, executamos Augustus com os parâmetros: "*--codingseq*", "*--protein*" e "--cds" e usamos *Apis mellifera* ("honeybee1") como 'espécie de referência', gerando dados de saída contendo sequências de proteínas e CDS e informações relacionadas a predição. A predição gênica de sequências proteicas de *S. postica* foram selecionadas para a anotação funcional por ontologias gênicas.

3.4 Anotação funcional

Para a anotação funcional de genes foram utilizadas duas ferramentas para comparação dos resultados fornecidos: a) eggNOG-mapper, b) blast2GO auxiliado pelo *diamond*.

a. *eggNOG-mapper:* As distribuições dos termos do grupo ortólogo eucariótico (KOG) foram atribuídas aos conjuntos de proteínas usando a ferramenta on-line *eggNOG- mapper* contra o banco de dados *eggNOG 5.0* (HUERTA *et al*. 2019).

b. *diamond* e *blast2GO:* o arquivo contendo as sequências de proteínas foi disponibilizado em linha de comando ao programa *diamond*, bem como foi utilizado o banco de dados *blastp* como referência para a anotação, possibilitando a geração de um arquivo xml, contendo dados associados ao nome das sequências, descrição,tamanho das sequências, *hits* e valores de *e-value*.

Esse arquivo foi carregado na interface gráfica desse programa, que possibilitou verificar a ontologia gênica por meio de buscas locais ou globais por sequências semelhantes as sequências de entrada, fornecendo assim, termos GO associados a cada um dos hits obtidos, o que retornou uma anotação GO avaliada para as sequências em estudo. Também foram obtidos códigos enzimáticos pelo mapeamento de GOs equivalentes e os motivos InterPro.

4 RESULTADOS

4.1 Sequenciamento e controle de qualidade

O DNA de *Scaptotrigona postica* foi sequenciado pela plataforma Illumina NextSeq1000/2000, com leituras 2x100 *paired-end* e MinION Nanopore com sequências de 10 kpb até 50 kpb. Foi gerado um total de 46.5 Gb de dados, sendo 155 milhões de leituras curtas, com cobertura de aproximadamente 96 vezes, e 5.7 milhões de leituras longas, com cobertura de 16 vezes. O tamanho do genoma haploide foi medido usando *k-mers* (*k*=21). Por meio do GenomeScope foi estimado o tamanho do genoma em 310 Mb (Figura 7).

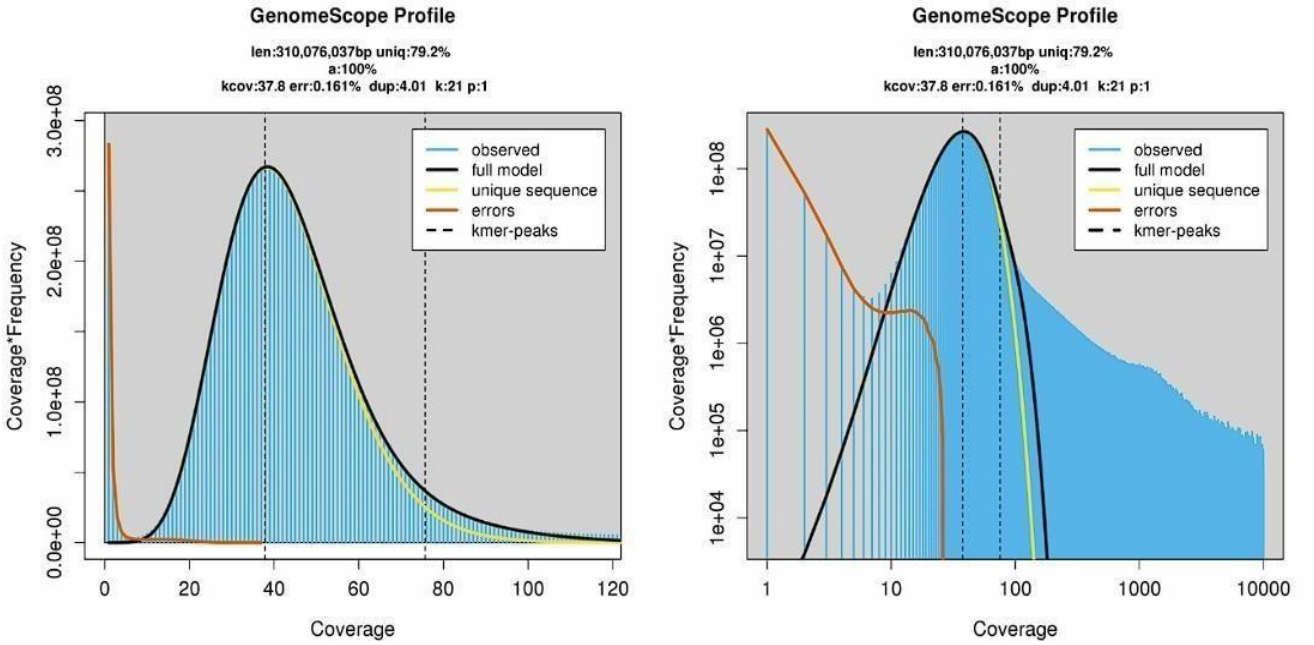

Figura 7: Resultado GenomeScope do genoma de *S. postica*. Os valores evidenciados na legenda correspondem ao comprimento do genoma total inferido (len), porcentagem de comprimento único do genoma (uniq), cobertura média de *k-mer* (kcov), taxa de erro de leitura (err), taxa média de duplicações de leitura (dup) e tamanho de *k-mer* (k)

4.2 Montagem do genoma

A montagem com o software Flye resultou em um genoma de 277 Mb com 37.48% de conteúdo GC. O *assembly* foi composto por 243 *contigs*, com cobertura média de 15x e valor de N50 de 13.2 Mb. A montagem híbrida, realizada pelo MaSuRCA, apresentou valores semelhantes, com um tamanho total de 278 Mb e 37.39% de conteúdo GC. O *draft assembly* foi composto por 227 *contigs*, cobertura média de 16x e valor de N50 de 3.7 Mb (tabela 2). As estimativas de ambos *assemblies* se encaixam muito bem com as estimativas para abelhas solitárias e sociais. A qualidade da montagem de *S. postica* é similar à de genomas publicados para outras espécies da família Apidae, como por exemplo *Frieseomelitta varia* e *Apis dorsata* (FREITAS *et al*. 2020; OPPENHEIM *et al*. 2020).

Tabela 2. Dados obtidos para as montagens através do software Flye e MaSuRCA gerados pelo QUAST.

	Softwares utilizados na montagem	
	Flye	MaSuRCA
Tamanho total da montagem (pb)	277.728.880	278.791.247
Número total de *contigs*	243	227
Contigs > 25k nucleotídeos	119	217
Contigs > 50k nucleotídeos	99	199
Tamanho do maior *contig* (pb)	18.200.319	13.172.936
%GC	37.48	37.39
N50 dos *contigs*	13.267.537	3.749.223
L50 dos *contigs*	9	20

4.3 Análise busco e predição de genes

A fim de identificar a completude das montagens foram feitas análises BUSCO contra o banco de dados de *Hymenoptera*. A montagem gerada pelo Flye apresentou 95.2% de genes completos, sendo 95.0% de cópia única e 0.5% duplicados, também foram identificados 1.7% de genes fragmentados e 3.1% dos genes não obtiveram nenhuma correspondência. A montagem produzida pelo MaSuRCA apresentou 96.3% de genes completos, sendo 96.3% de cópia única e 0.7% duplicados, também foi identificado 0.5% de genes fragmentados e 3.2% dos genes não obtiveram nenhuma correspondência. Os valores absolutos de cada categoria são apresentados na tabela 3.

Tabela 3. Valores absolutos das categorias avaliadas pelo BUSCO para ambos os softwares utilizados namontagem do genoma.

	Softwares utilizados na montagem	
	Flye	MaSuRCA
Complete BUSCOs (C)	5704	5770
Complete and single-copy BUSCOs (S)	5694	5727
Complete and duplicated BUSCOs (D)	10	43
Fragmented BUSCOs (F)	99	28
Missing BUSCOs (M)	188	193
Total BUSCO groups searched	5991	5991

A predição dos genes realizada através do software Augustus v3.2.2 (STANKE, 2006), a partir da montagem híbrida, gerou dois arquivos *fasta*, sendo um relativo à predição de CDS (*coding sequences*) e o segundo de sequências de proteínas e gerou um total de 15.168 genes preditos.

4.4 Anotação funcional

A anotação funcional foi realizada pela ferramenta online eggNOG-mapper v.2.1.8 contra o banco de dados eggNOG 5.0. Foi utilizado *euKaryotic Orthologous Group terms* (KOG), a partir da sequência de proteínas obtidas na predição de genes. Para esta ferramenta, o total de 15.168 sequências proteicas resultou em 9.906 sequências classificadas funcionalmente. As sequências de proteínas foram classificadas funcionalmente em 25 categorias (Figura 8), incluindo uma categoria de funções desconhecidas e de função não encontrada a partir do eggNOG. Nesta ferramenta uma mesma sequência pode ser classificada em mais de uma categoria, sendo que uma mesma sequência foi classificada em até 5 categorias diferentes. A maior porcentagem de categorias identificadas para as 9.906 sequências foram: i) mecanismos de transdução de sinal (1.334); ii) não encontrados em

eggNOG (865); iii) Chaperonas, renovação de proteínas e modificações pós-traducionais(838) e iv) transcrição (748).

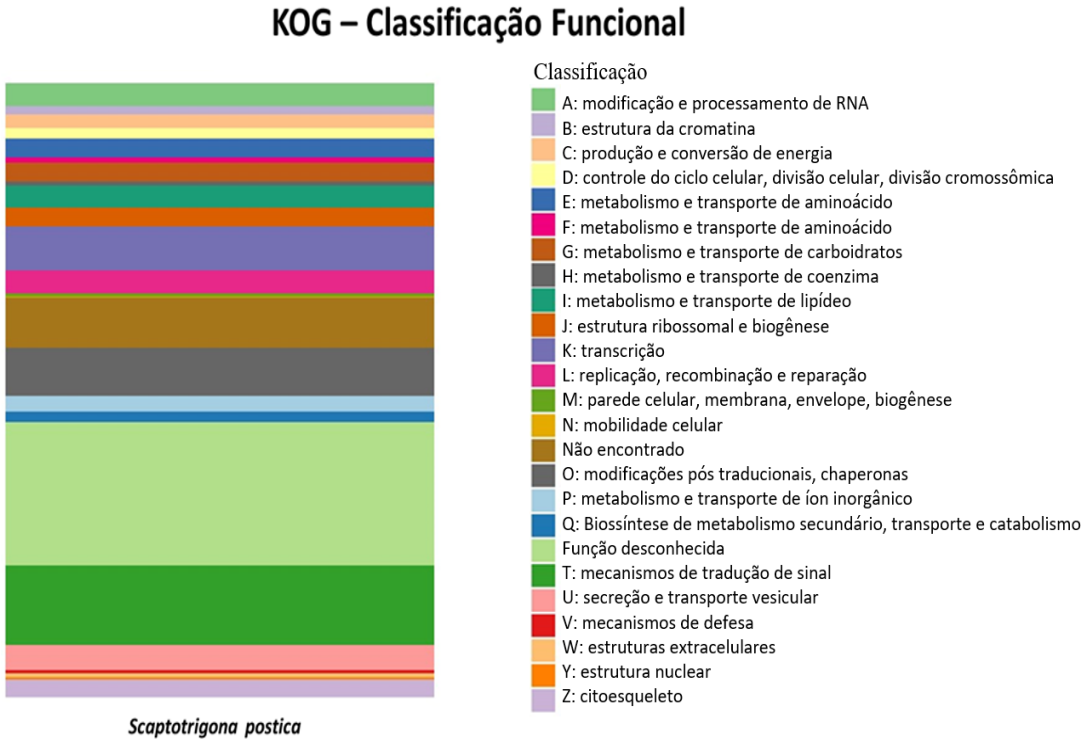

Figura 8. Ilustração da distribuição de categorias funcionais KOG entre genes codificadores de proteínas realizado pelo RStudio. Os nomes de cada definição de categoria KOG (A–Z) são fornecidos à direita. Códigos de cores conforme indicado.

A outra estratégia utilizada para realizar a anotação funcional das sequências de proteínas geradas durante a etapa de predição, o blast2GO associado ao *diamond*, resultou na anotação de 100% das proteínas preditas usando o banco de dados de InterPro e 12.022 sequências foram anotadas utilizando apenas a ferramenta blastp (Figura 9). Os dados de proteínas atribuídos a *S. postica* apresentaram *top hits* significativos com outras espécies de abelhas como *Frieseomelitta varia* com mais de 7000 *top hits* e *Melipona quadrifasciata* com mais de 1000 *top hits* (Figura 10).

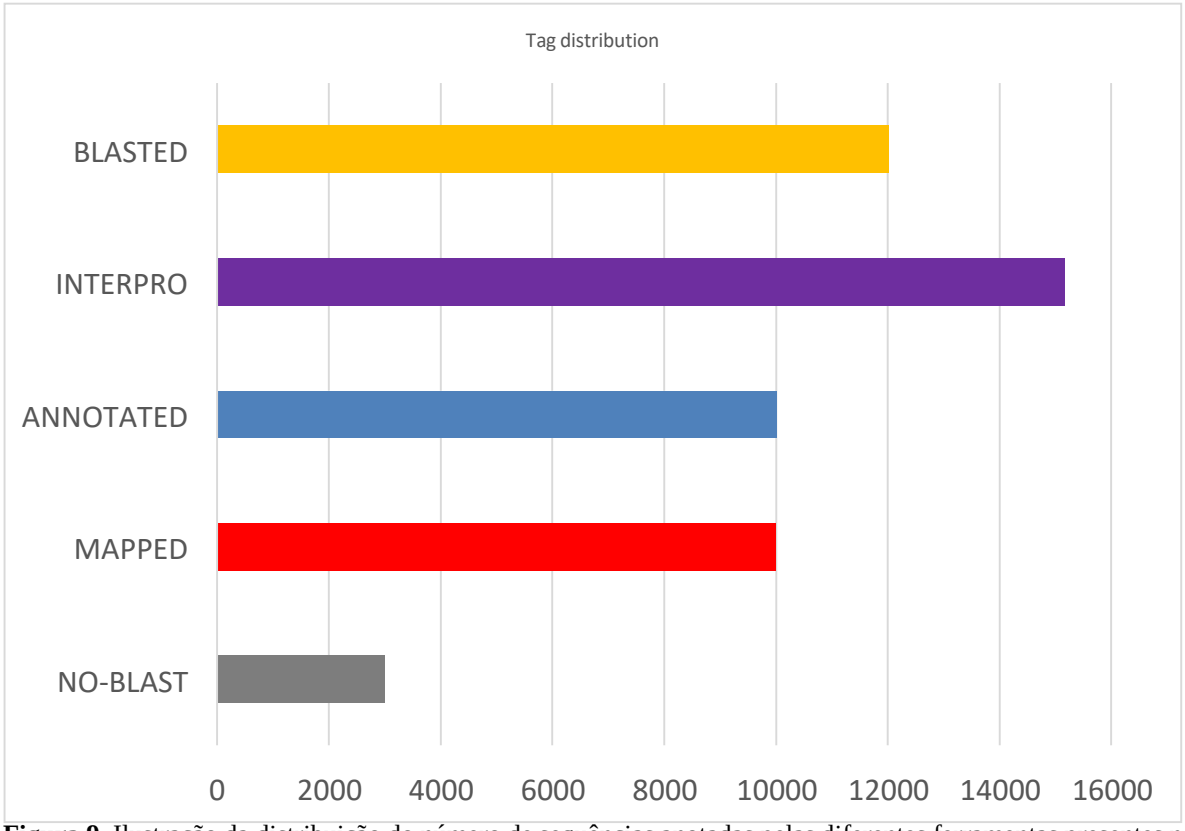

Figura 9. Ilustração da distribuição do número de sequências anotadas pelas diferentes ferramentas presentes no programa blast2GO, como blast, interpro, anotado e mapeado.

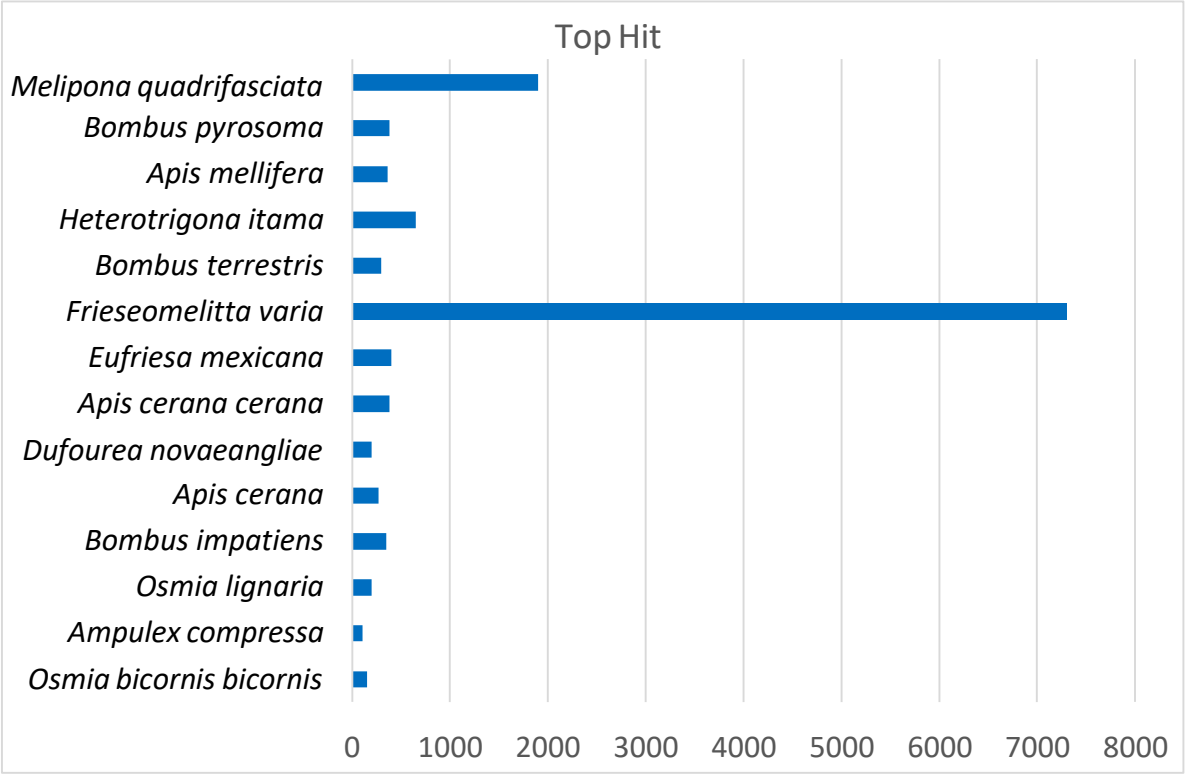

Figura 10. Distribuição do número dos principais *top-hit* entre espécies descritas.

Outra representação visual disponibilizada pelo programa blast2GO é o grafo das vias correspondentes aos processos biológicos, moleculares e celulares, no qual a principal atividade foi atribuída à processos celulares, a qual compreendeu 98,19% das sequências atribuídas a GOs associados a componentes celulares. Dentre osprocessos moleculares o mais destacado foi o de construção das vias metabólicas (72% das sequências) e processos celulares (89,13%).

5 DISCUSSÃO

A estimativa do tamanho do genoma foi de 310 Mb pelo GenomeScope 2.0. As duas montagens atingiram 278 e 277 Mb pelo MaSuRCA e Flye, respectivamente. As estatísticas são bastante similares ao da espécie próxima filogeneticamente *Friesomielitta varia*, que possui um *assembly* com 275 Mb (FREITAS *et al*. 2020). O *draft assembly* apresentou o tamanho esperado para uma representação haploide do genoma. Os softwares testados usam grafos de Bruijn como estratégia para montar os *contigs*, mas possuem diferentes características. O MaSuRCA possui maior flexibilidade com as sobreposições entre as leituras e permite o uso de sequências longas e curtas. Cada montador possui diferentes métodos para corrigir erros de sequenciamento e manejar sequências repetitivas. O montador MaSuRCA possui particularidades que permitiram gerar *contigs* grandes e em menor quantidade. Apesar disso, o Flye apresentou maior valor de N50 e o maior *contig* com valor de pares de bases maior do que o obtido pelo MaSuRCA.

A profundidade do *draft assembly* para *S. postica* foi de 16x, sendo 12 vezes menor que o de *A. mellifera*. O valor de N50 representa a contiguidade do *assembly*, quanto maior o seu valor e menor o L50, maior é o tamanho dos *contigs* e menor é a quantidade necessária de *contigs* para cobrir o genoma, resultando num *assemblie* de maior qualidade. Essas métricas devem ser interpretadas em conjunto para uma melhor compreensão dos dados. Conseguimos gerar 227 *contigs*, obtivemos valores de N50 dos *contigs* maiores e valores de L50 menores que as demais montagens, com exceção do L50 dos *contigs* para o genoma da *A. melífera*, indicando uma boa qualidade do nosso *draft assembly*.

Tabela 4. Estatísticas das montagens do genoma das espécies *Scaptotrigona postica*, *Frieseomelitta varia*, *Apisdorsata* e *Apis melífera*.

	Softwares utilizados na montagem					
	Flye *S. postica*	MaSuRCA *S. postica*	ABySS v. AUG-2019 *F. var* 1	SPAdes v. 3.9.0 *F. var* 1.2	SOAPdenovo v. 1.05 *A. dorsata*	FALCON 0.5.0 v *A. mellifera*
Tamanho total da montagem (pb)	277.728.880	278.791.247	301.350.145	275.412.029	230.340.171	225.250.884
Número total de *contigs*	243	227	9.755	6.729	45,204	228
%GC	37.48	37.39	36.5	36.72	31.9	32.5
N50 dos *contigs* (pb)	13.267.537	3.749.223	83.201	98.566	30.868	5.382.476
L50 dos *contigs*	9	20	946	790	1.978	13
Profundidade	15x	16x	10x	93x	60x	192x

Em termos de completude de regiões gênicas, nossa montagem de *S. postica* apresentou um número significativo de genes de cópia única de himenóptera conservados, com 96.3% de genes completos e 0,5% de duplicação. O nível de duplicação pode serexplicado por um conjunto quimérico de haplótipos ou por eventos recentes de duplicação. Em comparação com *assemblies* de outras espécies de abelhas os resultados estão dentro do esperado para a montagem de um *draft assembly*.

Tabela 5. Análise BUSCO da montagem do genoma das espécies *Scaptotrigona postica*, *Frieseomelitta varia*,*Apis dorsata* e *Apis melífera*.

	S. postica (1)	*S. postica* (2)	*F. var* 1.2	*A. dorsata*	*A. mellifera*
Complete BUSCOs (C)	95.2%	96.3%	98.1%	96.8%	98.8%
Complete and single-copy BUSCOs (S)	95.0%	96.3%	97.9%	96.7%	98.6%
Complete and duplicated BUSCOs (D)	0.5%	0.7%	0.2%	0.1%	0.2%
Fragmented BUSCOs (F)	1.7%	0.5%	0.9%	1.3%	0.3%
Missing BUSCOs (M)	3.1%	3.2%	1.0%	1.9%	0.9%
Total BUSCO *groups searched*	5991	5991	5991	5991	5991

A escolha da pipeline de bioinformática para a etapa de predição dos genes interfere diretamente na anotação do genoma, a predição dos genes foi realizada a partir do software Augustus que é um localizador de genes baseado em Modelo oculto de Markov (GHMM) muito utilizado e referenciado pela comunidade científica (STANKE *et al*. 2006). Os resultados obtidos a partir desta predição podem ser extraídos em formato de texto ou gráfico, a saída de texto (txt) apresenta os limites de éxons, íntrons, transcrição e genes no Formato "General Feature" - (GFF), além das sequências de CDS e de proteínas no formato FASTA.

A partir da predição, a anotação funcional foi realizada pela ferramenta online eggNOG-mapper, utilizando o banco de dados eggNOG 5.0 que é um banco de dados público de relações ortológicas, histórias evolutivas de genes e anotações funcionais. A qualidade das atribuições de ortologia e anotações funcionais apresentam em média 80% de cobertura, além de fornecer uma anotação funcional rápida e previsão de ortologia de dados de genomas através da ferramenta online (HUERTA-CEPAS *et al*. 2019). Apesar da facilidade e agilidade da anotação, como mencionado, a atribuição de ortologias para o genoma de *S. postica* foi de 65%. Já a anotação funcional realizada pelo blast2GO auxiliado por Diamond utiliza o banco de dados não redundantes (nr) de peptídeos do NCBI e utiliza como ferramentas de anotação

o blast (*basic local alignment search tool*), que se trata de uma ferramenta de busca de similaridade entre sequências biológicas, sendo capaz de identificar relação entre sequências que compartilham similaridade (AMARAL *et al*. 2007). O programa blast2GO gera representações gráficas de diferentes parâmetros associados a anotação por blast e interpro, bem como o número de sequências de proteínas anotadas pelo interpro e pelo blast foram superiores às obtidas pelo *eggNOG*, porém o *eggNOG* permite que as sequências anotadas sejam classificadas em diferentes categorias funcionais obtidas pela utilização do banco de dados KOG, que não está presente no blast2GO, bem como permite que interações entre ontologias sejam analisadas por meio do banco de dados KEGG. Dessa forma, ambas as ferramentas de anotação utilizadas nesse trabalho apresentam pontos positivos e negativos e a escolha do seu uso dependerá dos objetivos propostos pelo pesquisador.

6 CONCLUSÃO

Com os dados obtidos foi possível construir um *draft assembly* para espécie *Scaptotrigona postica* e sua caracterização quanto ao tamanho do genoma, predição e anotação funcional de genes. A montagem do genoma utilizando a abordagem híbrida apresentou melhores estatísticas, sendo considerada a melhor montagem. O *draft assembly* representa 278 Mb, o que é similar com valores encontrados na literatura para genomas de espécies próximas na filogenia, como *F. varia* e *A. dorsata*.

A anotação funcional gerada pelo blast2GO apresentou o maior número de genes anotados funcionalmente. Foram preditos 15.168 genes, baseado em uma predição *ab initio* e 15.168 sequências anotadas. Este estudo abre perspectivas para a montagem do genoma completo, identificaçãoe quantificação de elementos repetivos e sequencias microssatélites além da utilização destas sequencias obtidas para a montagem de um genoma mitocondrial.

Estes resultados são de extrema valia pois apresentam um primeiro passo para os estudos populacionais e de descobertas de genes de interesse para preservação desta importante espécie de abelha nativa. Este é o primeiro trabalho com uma abordagem híbrida para montagem de genoma de uma espécie de abelha eussocial, a qual está indicada para avaliação de risco a agrotóxicos. Dessa forma, o trabalho gerou importantes subsídeos para pesquisas futuras na área de genômica e conservação de espécies nativas.

REFERÊNCIAS

AMARAL AM *et al* (2007) O programa BLAST: guia prático de utilização. **Embrapa Recursos Genéticos e Biotecnologia**. Brasília, DF, p. 24

ANDREWS S *et al* (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics

COMPEAU PEC *et al* (2011) How to apply de Bruijn graphs to genome assembly. **Nature biotechnology**, v. 29, n. 11, p. 987-991

CORNMAN RS *et al* (2013) Population-genomic variation within RNA viruses of the Western honey bee, Apis mellifera, inferred from deep sequencing. **BMC Genomics**, v. 14, n. 1

CORTOPASSI LM *et al* (2006) Global meliponiculture: challenges and opportunities. **Apidologie**, v. 37, n. 2, p. 275-292

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

CRUZ DO *et al* (2005) Pollination efficiency of the stingless bee Melipona subnitida on greenhouse sweet pepper. **Pesquisa Agropecuária Brasileira**, v. 40, p. 1197-1201

FREITAS FCP *et al* (2020) The nuclear and mitochondrial genomes of *Frieseomelitta varia* - A highly eusocial stingless bee (Meliponini) with a permanently sterile worker caste.**BMC Genomics**, v. 21, n. 1

FREITAS PV *et al* (2016) Declínio populacional das abelhas polinizadoras: Revisão. **Pubvet**, v. 11, p. 1-102

FUKASAWA Y, ERMINI L, WANG H, CARTY K, CHEUNG MS (2020) LongQC: A Quality Control Tool for Third Generation Sequencing Long *Read* Data. **G3 Genes|Genomes|Genetics**, v. 10, n. 4, p. 1193–1196

GALLAI N, SALLES JM, SETTELE J, VAISSIÈRE BE (2009) Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. **Ecological Economics**, v. 68, n. 3, p. 810–821

GIANNINI TC, CORDEIRO GD, FREITAS BM,SARAIVA AM, IMPERATRIZ-FONSECA VL (2015) The Dependence of Crops for Pollinators and the Economic Value of Pollination in Brazil. **Journal of Economic Entomology**, v. 108, n. 3, p. 849–857

GILLOTT C (2005) Entomology. **Springer Science** & Business Media

HALL N (2007) Advanced sequencing technologies and their wider impact in microbiology." **Journal of experimental biology** 210.9 p. 1518-1525

HEARD TA (1999) The role of stingless bees in crop pollination. **Annual review of entomology**, v. 44, n. 1, p. 183-206

HUERTA CJ *et al* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. **Nucleic Acids Research**, v. 47, n. D1, p. D309–D314

HUNT GJ *et al* (1995) Linkage map of the honey bee, Apis mellifera, based on RAPD markers. **Genetics**, v. 139, n. 3, p. 1371-1382

KAPHEIM KM, PAN H, LI C *et al* (2015) Genomic signatures of evolutionary transitions from solitary to group living. **Science**, v. 348, n. 6239, p. 1139–1143

KLEIN S, CABIROL A, DEVAUD JM, BARRON AB, LIHOREAU M (2017) Why Bees Are So Vulnerable to Environmental Stressors. **Trends in Ecology and Evolution**

KOLMOGOROV M *et al* (2019) Assembly of long, error-prone *reads* using repeat graphs. **Nature biotechnology**, v. 37, n. 5, p. 540-546

LI Z *et al* (2012) Comparison of the two major classes of assembly algorithms: overlap– layout–consensus and de-bruijn-graph. **Briefings in Functional Genomics**, v. 11, p. 25– 37

LIN Y, YUAN J, KOLMOGOROV M *et al* (2016) Assembly of long error-prone *reads* using de Bruijn graphs. **Proceedings of the National Academy of Sciences of the United States of America**, v. 113, n. 52, p. E8396–E8405

LOZIER JD, ZAYED A (2017) Bee conservation in the age of genomics. **Conservation Genetics**, v. 18, n. 3, p. 713–729

MANRIQUE AJ, SOARES EEA (2002) Início de um programa de seleção de abelhas africanizadas para a melhoria na produção de própolis e seu efeito na produção de mel. **Interciencia**, v. 27, n. 6, p. 312-316

MENEZES PS (2014) The stingless bee fauna in Brazil (Hymenoptera: Apidae). **Sociobiology**, v. 61, n. 4, p. 348–354

MICHENER CD (2013) The Meliponini, Pot-Honey. **Springer**, New York, NY. https://doi.org/10.1007/978-1-4614-4960-7_1

MICHENER CD (2000) **The bees of the world**. JHU press

MILLER JR *et al* (2008) Aggressive assembly of pyrosequencing *reads* with mates. **Bioinformatics**, v. 24, p. 2818–2824, https://doi.org/10.1093/bioinformatics/btn548

MILLER JR *et al* (2010) Assembly algorithms for next-generation sequencingdata. **Genomics**, v. 95, n. 6, p. 315-327

MYERS EW *et al* (2000) A whole-genome assembly of Drosophila. **Science**, v. 287, n. 5461,

p. 2196-2204

NOCELLI RC *et al* (2012) Riscos de pesticidas sobre as abelhas. **Semana dos Polinizadores**, v. 3, p. 196-212

OPPENHEIM S *et al* (2020) Whole Genome Sequencing and Assembly of the Asian Honey Bee *Apis dorsata*, **Genome Biology and Evolution**, v. 12, Issue 1, p. 3677–3683

PEREIRA FM et al (2017) Criação de abelhas-sem-ferrão. **Embrapa Meio-Norte**

PIRES CSS *et al* (2016) Weakness and collapse of bee colonies in Brazil: Are there cases of CCD? **Pesquisa Agropecuaria Brasileira**, v. 51, n. 5, p. 422–442

POTTS SG *et al* (2010) Global pollinator declines: trends, impacts and drivers. **Trends in ecology & evolution**, v. 25, n. 6, p. 345-353

POTT SG *et al* (2016) Summary for policymakers of the assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production. **Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services**, Bonn, Germany, v. 36

ROUBIK DW (1992) Ecology and natural history of tropical bees. Cambridge University Press

ROUBIK DW (2006) Stingless bee nesting biology. **Apidologie**.

ROUBIK DW (2018) 100 species of meliponines (Apidae: Meliponini) in a parcel of western Amazonian forest at Yasuní Biosphere Reserve, Ecuador. **Pot-Pollen in Stingless Bee Melittology**, p.189–206

SAMPAIO AR *et al* (2020) Análise de agrotóxicos e Acibenzolar-S-Metílico (ASM) em *Apis mellifera L*.(Hymenoptera: Apidae) e a ação de ASM na indução de resistência em soja. Dissertação de Mestrado. **Universidade Tecnológica Federal do Paraná**

SILVA CM, TORRE PA, DELLA, MATOS JDC (2021) O uso incorreto do inseticida fipronil e sua influência na morte das abelhas no sul do Brasil. **Revista Processando o Saber**, v. 13, p. 93–110

SILVA GR *et al* (2019) Pesquisas com abelhas-sem-ferrão (Hymenoptera: Meliponini) e aplicabilidade dos marcadores moleculares: uma revisão sistemática da literatura. **Pubvet**, v.13, n.1, a250, p.1-19

SLAA EJ *et al* (2006) Stingless bees in applied pollination: practice and perspectives. **Apidologie**, v. 37, n. 2, p. 293-315

SLATKO BE, GARDNER AF, AUSUBEL FM (2018) Overview of Next-Generation Sequencing Technologies. **Current Protocols in Molecular Biology**, v. 122, n. 1

SOLIGNAC M *et al* (2007) A third-generation microsatellite-based linkage map of the honey

bee, Apis mellifera, and its comparison with the sequence-based physical map. **Genome Biology**, v. 8, p. 1-14

STANKE M *et al* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. **Nucleic Acids Research**, v. 34, n. 2, p. 435-439

STEIN L (2001) Genome annotation: from sequence to biology. **Nature Reviews Genetics**, v. 2, n. 7, p. 493–503

TAN HW *et al* (2011) The complete mitochondrial genome of the Asiatic cavity-nesting honeybee *Apis cerana* (Hymenoptera: Apidae). **Plos one**, v. 6, n. 8

VAILATI RM, PALOMBO V, LOOR JJ (2017) What are omics sciences? **Periparturient Diseases of Dairy Cows: A Systems Biology Approach**, p. 1–7

VENTURIERI GC *et al* (2003) Avaliação da introdução da criação racional de Melipona fasciculata (Apidae: Meliponina), entre os agricultores familiares de Bragança-PA, Brasil. **Biota Neotropica**, v. 3, p. 1-7

VURTURE GW, SEDLAZECK FJ, NATTESTAD M *et al* (2017) GenomeScope: fast reference-free genome profiling from short *reads*. **Bioinformatics**, v. 33, n. 14, p. 2202–2204

WANG L, WANG S, LI W (2012) RSeQC: quality control of RNA-seq experiments. **Bioinformatics**, v. 28, n. 16, p. 2184-2185

WILLE A, MICHENER CD (1973) The nest architecture of stingless bees with special reference to those of Costa Rica (Hymenoptera, Apidae). **Revista de biologia tropical**

WILLE A (1983) Biology of the stingless bees. **Annual review of entomology**, v. 28, n. 1, p. 41-64

YANDELL M, ENCE D (2012) A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p. 329-342

ZHANG H, JAIN C, ALURU S (2020) A comprehensive evaluation of long *read* error correction methods. **BMC genomics**, v. 21, p. 1-15

ZIMIN AV *et al* (2013) The MaSuRCA genome assembler. **Bioinformatics**, v. 29, n. 21, p. 2669–2677

ZIMIN AV *et al* (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of b*read* wheat, with the MaSuRCA mega-*reads* algorithm. **Genome research**, v. 27, n. 5, p. 787-792