Universidade Estadual de Londrina
Centro de Ciências Biológicas
Colegiado do Curso de Ciências Biológicas

Ciências Biológicas UEL

TRABALHO DE CONCLUSÃO DO CURSO DE GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS

MATHEUS HENRIQUE DE OLIVEIRA ROSA

ESTRATÉGIAS DE MONTAGEM DE TRANSCRIPTOMA EM *BOMBYX MORI* LINNAEUS 1758: ABORDAGENS GUIADAS POR GENOMA *VERSUS DE NOVO*

Londrina – Paraná

2024

TRABALHO DE CONCLUSÃO DO CURSO DE GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS

MATHEUS HENRIQUE DE OLIVEIRA ROSA

ESTRATÉGIAS DE MONTAGEM DE TRANSCRIPTOMA EM *BOMBYX MORI* LINNAEUS 1758: ABORDAGENS GUIADAS POR GENOMA *VERSUS DE NOVO*

Monografia apresentada ao Curso de Graduação em Ciências Biológicas da Universidade Estadual de Londrina como um dos requisitos à obtenção do título de Bacharel em Ciências Biológicas.

Orientador: Profa. Dra. Renata da Rosa

Coorientador: Prof. Dr. Rogério Fernandes de Souza

Londrina – Paraná

2024

Matheus Henrique de Oliveira

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

M427e Rosa, Matheus Henrique de Oliveira.
Estratégias de montagem de transcriptoma em Bombyx mori Linnaeus 1758 : Abordagens guiadas por genoma versus de novo / Matheus Henrique de Oliveira Rosa. - Londrina, 2024.
41 f. : il.

Orientador: Renata da Rosa.
Coorientador: Rogério Fernandes de Souza.
Trabalho de Conclusão de Curso (Graduação em Ciências Biológicas) – Universidade Estadual de Londrina, Centro de Ciências Biológicas, Graduação em Ciências Biológicas, 2024.
Inclui bibliografia.

1. Comparação entre montagem guiada por genoma de referência e montagem de novo - TCC. 2. Montagem de transcriptoma de Bombyx mori - TCC. 3. RNA-seq - TCC. 4. Bioinformática - TCC. I. Rosa, Renata da. II. Souza, Rogério Fernandes de . III. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Graduação em Ciências Biológicas. IV. Título.

CDU 574

BANCA EXAMINADORA

Prof(a). Dr(a). Renata da Rosa

Prof. Dr. Laurival Antônio Vilas Boas

Dr(a). Larissa Forim Pezenti

Londrina, 10 de maio de 2024

AGRADECIMENTOS

Gostaria de expressar minha mais profunda gratidão a todos que contribuíram para a realização deste trabalho, em especial à minha orientadora, Renata da Rosa. Sua orientação foi fundamental para que eu pudesse explorar um ramo tão fascinante da biologia, como a biologia molecular. Sou imensamente grato por sua dedicação e apoio ao longo deste percurso acadêmico.

Além disso, gostaria de agradecer ao meu coorientador, Rogério Fernandes de Souza, por compartilhar seu conhecimento e paixão pela bioinformática. Cada vez mais, me encanto com esse campo, e devo a você grande parte desse entusiasmo. Seus ensinamentos foram inestimáveis e contribuíram significativamente para o desenvolvimento deste trabalho.

Também agradeço aos membros dos laboratórios de Citogenética, Entomologia Molecular e Bioinformática da UEL. Vocês têm sido uma fonte inestimável de apoio ao longo deste período, sempre prontos para oferecer orientação, compartilhar conhecimento e estender uma mão amiga quando necessário. Sem dúvida, vocês não apenas me ajudaram a superar desafios, mas também me ensinaram lições valiosas que levarei para toda a vida.

Não posso deixar de mencionar a Universidade Estadual de Londrina, o CNPq, a SETI do Paraná e ao Fundo paraná, pelo apoio e financiamento que tornaram possível a realização desta pesquisa.

Por último, mas definitivamente não menos importante, quero expressar minha profunda gratidão aos meus pais, Claubeir da Silva Rosa e Sonia Rosane de Oliveira Rosa. Seu apoio incondicional e dedicação foram fundamentais em cada etapa da minha jornada acadêmica. Vocês sempre acreditaram em mim e fizeram tudo ao seu

alcance para me proporcionar as condições necessárias para perseguir meus sonhos. Sem vocês, nada disso seria possível.

A todos vocês, meu sincero muito obrigado. Este trabalho é fruto não apenas do meu esforço individual, mas também do apoio, orientação e incentivo de cada um de vocês. Que este seja apenas o início de uma jornada repleta de descobertas e realizações.

RESUMO

ROSA, Matheus Henrique de Oliveira. **ESTRATÉGIAS DE MONTAGEM DE TRANSCRIPTOMA EM *BOMBYX MORI* LINNAEUS 1758: ABORDAGENS GUIADAS POR GENOMA VERSUS *DE NOVO*.** 2024. 41. Trabalho de Conclusão de Curso (Graduação em Ciências Biológicas) – Universidade Estadual de Londrina, Londrina. 2024.

Para avaliar a eficácia na montagem de transcritos e as diferenças entre estratégias, este estudo comparou dois métodos de montagem de transcriptoma: a montagem guiada e a montagem *de novo*, utilizando bibliotecas de RNA-seq de lagartas de *Bombyx mori*, criadas sob diferentes temperaturas (26°C e 34°C). Para isso foi empregado os montadores StringTie e Trinity, respectivamente. Foi avaliada a qualidade da montagem, a anotação de transcritos e proteínas, e conduzida análises de expressão diferencial e enriquecimento de vias metabólicas. Os resultados mostraram diferenças entre os métodos. A montagem *de novo* produziu um número maior de transcritos (73.743) quando comparada a montagem guiada por genoma (29.157), bem como identificou um número maior de proteínas ao utilizar bancos de dados de Lepidoptera do NCBI (19.081 contra 15.629) e do UniProt (18.693 e 15.400). Por outro lado, a montagem guiada, utilizando um banco de dados de *B. mori* disponível no NCBI, identificou um número maior de proteínas (13.726 contra 13.702) e demonstrou transcritos mais longos e completos. Ambos os métodos identificaram proteínas exclusivas, ou seja, conseguiram construir sequências que o outro método não conseguiu. Os resultados permitiram concluir que a escolha do método depende dos objetivos da pesquisa. Enquanto a montagem *de novo* pode ser mais adequada para descobrir novos transcritos, a montagem guiada pode ser preferível para obter transcritos mais completos. Além disso, uma abordagem híbrida pode ser vantajosa para aumentar a anotação do transcriptoma, uma vez que nesse estudo, a junção dos dados obtidos pelas duas montagens, mostrou uma melhoria em relação a quantidade de proteínas anotadas. Esses resultados fornecem insights para pesquisadores interessados em otimizar suas estratégias de montagem de transcriptoma.

Palavras-chave: Comparação de montagem. Montagem *de novo*. Montagem guiada. RNA-seq

ABSTRACT

ROSA, Matheus Henrique de Oliveira. **TRANSCRYPTOME ASSEMBLY STRATEGIES IN BOMBYX MORI LINNAEUS 1758: GENOME-GUIDED APPROACHES VERSUS DE DEVO**. 2024. 41. Course Completion Work (Graduation in Biological Sciences) – State University of Londrina, Londrina. 2024.

To evaluate the effectiveness of transcript assembly and the differences between strategies, this study compared two transcriptome assembly methods: guided assembly and de novo assembly, using RNA-seq libraries from Bombyx mori caterpillars raised under different temperatures (26°C and 34°C). For this purpose, the StringTie and Trinity assemblers were used, respectively. The quality of the assembly, the annotation of transcripts and proteins were evaluated, and analyzes of differential expression and enrichment of metabolic pathways were conducted. The results showed differences between the methods. De novo assembly produced a greater number of transcripts (73,743) when compared to genome-guided assembly (29,157), as well as identifying a greater number of proteins when using Lepidoptera databases from NCBI (19,081 versus 15,629) and UniProt (18,693 and 15,400). On the other hand, guided assembly, using a B. mori database available at NCBI, identified a greater number of proteins (13,726 versus 13,702) and demonstrated longer and more complete transcripts. Both methods identified unique proteins, that is, they were able to construct sequences that the other method could not. The results allowed us to conclude that the choice of method depends on the research objectives. While de novo assembly may be more suitable for discovering new transcripts, guided assembly may be preferable for obtaining more complete transcripts. Furthermore, a hybrid approach can be advantageous to increase transcriptome annotation, since in this study, the combination of data obtained by the two assemblies showed an improvement in relation to the quantity of annotated proteins. These results provide insights for researchers interested in optimizing their transcriptome assembly strategies.

Keywords: Assembly comparison. Assembly again. Guided assembly. RNA-seq

SUMÁRIO

1 INTRODUÇÃO

Bombyx mori Linnaeus, 1758 (Lepidoptera: Bombycidae), conhecida popularmente como bicho-da-seda é um inseto holometábolo com um ciclo de vida composto por quatro estágios: ovo, larva, pupa e mariposa (Vieira, 2020). *B. mori* possui grande importância comercial devido à seda que produz para confecção de seus casulos. Atualmente, o Brasil é o sexto maior produtor de seda do mundo, sendo reconhecido pela alta qualidade de seu tecido (International Sericultural Commission, 2023). O estado do Paraná é o principal produtor de seda do país, e várias famílias dependem da criação desse inseto para sua renda (Cirio et al., 2021).

Devido ao interesse comercial relacionado ao *B. mori*, existe uma vasta literatura sobre sua biologia, sendo um dos insetos com o genoma mais bem descrito. Isso permite o desenvolvimento de diversos estudos. Por esse motivo, esse inseto pode ser uma excelente escolha para estudos relacionados à montagem de transcriptoma.

A montagem de um transcriptoma é um processo complexo e crítico na análise da expressão gênica. A escolha do método de montagem pode influenciar significativamente os resultados obtidos e, consequentemente, a interpretação dos dados biológicos. Portanto, este estudo teve como objetivo comparar as abordagens de montagem guiada por genoma de referência e a montagem *de novo*, oferecendo *insights* que auxiliem os pesquisadores na escolha da melhor abordagem para suas respectivas pesquisas.

Para atingir esse objetivo, foram empregados dois montadores de transcriptoma: Trinity e StringTie, para realizar a montagem *de novo* e a montagem guiada, respectivamente. As análises de bioinformática foram conduzidas a partir de dados de RNA extraídos de lagartas de *B. mori*, criadas sob diferentes condições de

temperatura. Foram investigadas a qualidade da montagem, a anotação de transcritos e proteínas, bem como a expressão diferencial e o enriquecimento em vias metabólicas.

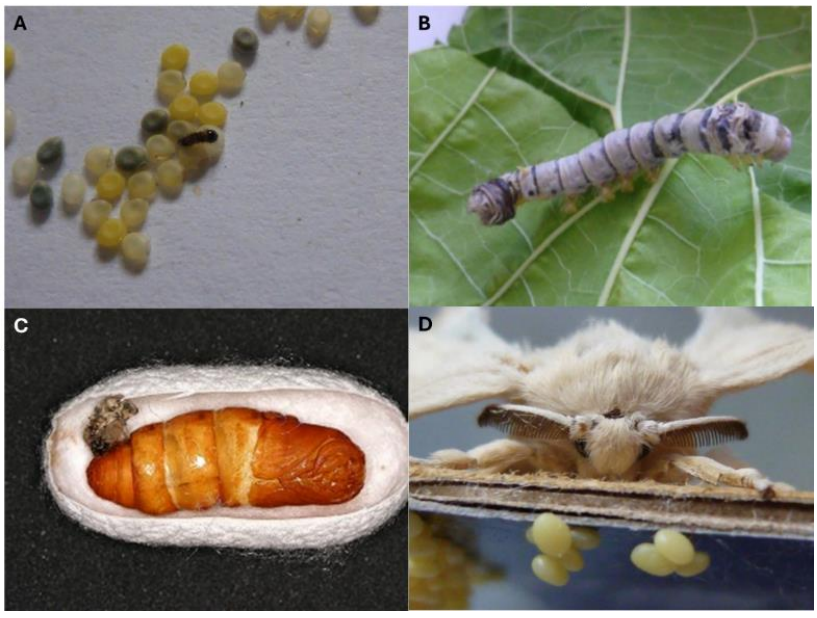
Ao longo deste estudo, foram empregadas métricas específicas para avaliar a completude, o número de isoformas, o tamanho dos transcritos e a eficiência do alinhamento dos *reads*. Além disso, foi explorada a expressão diferencial e o papel das proteínas identificadas em vias metabólicas, fornecendo uma visão abrangente das diferenças entre os métodos de montagem.

2 REVISÃO DE LITERATURA

2.1 *BOMBYX MORI*

Bombyx mori Linnaeus, 1758 (Lepidoptera: Bombycidae), conhecida popularmente como bicho-da-seda é um inseto holometábolo com um ciclo de vida composto por quatro estágios: ovo, larva, pupa e mariposa (Vieira, 2020) (Figura 1).

Figura 1 – Estágios de *Bombyx mori*: A) Ovos e larva no primeiro instar (Lopes, 2022); B) Larva (Lopes, 2022); C) Pupa dentro de casulo aberto por (Dvořák, 2006) D) Mariposa (Lopes, 2022).



Os ovos são geralmente ovóides e achatados, com tamanho de 1,0 a 1,3 mm de comprimento e 0,9 a 1,2 mm de largura. Quando depositados possuem a cor amarelo-clara. Caso estejam fecundados, se tornam amarelo escuro e durante o desenvolvimento vão escurecendo passando de laranja até o cinza Dependendo da raça existe uma variação na cor dos ovos (Vieira, 2020).

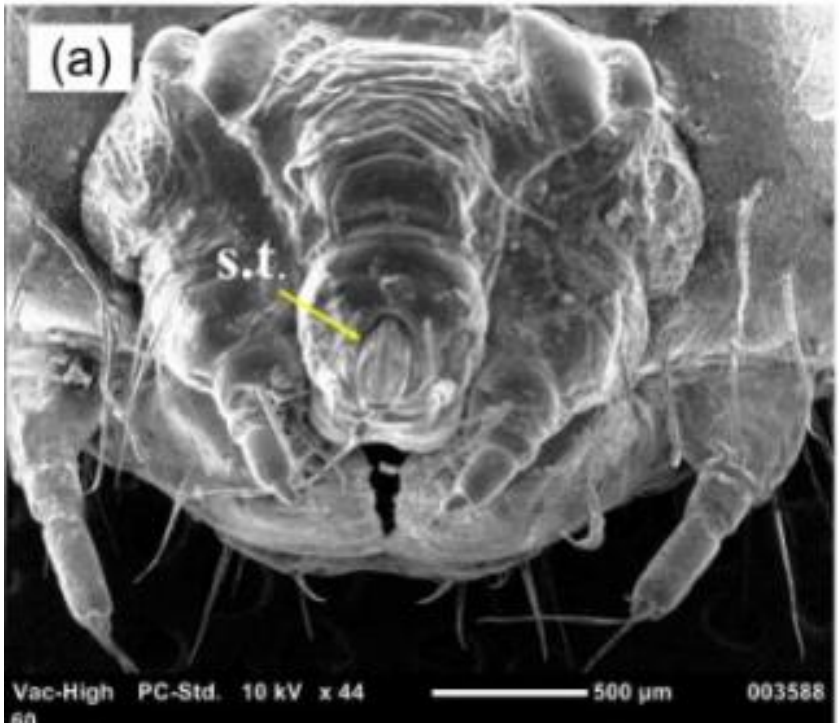
Do ovo eclode a larva, sendo a fase mais duradoura. Comumente chamada de lagarta, essa fase passa por cinco instares os quais são separados por quatro mudas (ecdises). No primeiro instar a lagarta apresenta cerca de 2 mm, com coloração escura e corpo recoberto de pelos escuros, sendo chamada popularmente de formiga. A partir do segundo instar a lagarta perde seus pelos e adquire uma coloração branca leitosa (Gallo et al., 2002; Vieira, 2020).

A cada ecdise as lagartas entram em um estado de repouso, cessando a alimentação e trocando o tegumento. A lagarta pode aumentar até setenta vezes seu tamanho original, sendo que no quinto instar ela pode atingir cerca de 5 cm, o que permite que esse animal seja facilmente manuseado (Munhoz, 2010 e Abdelli et al., 2018). No final do quinto instar, ela cessa completamente sua alimentação e inicia a produção de um casulo utilizando o fio de seda secretado pelas glândulas sericígenas localizadas na região látero-ventral do corpo (Gallo et al., 2002). As proteínas da seda são estocadas no lúmen da glândula durante cerca de oito dias no quinto instar. A principal proteína é a fibroína, sendo responsável por formar a fibra central, as outras são principalmente sericinas, que são proteínas adesivas e solúveis que revestem e cimentam a fibra de seda. Também foram identificados vários inibidores de proteases e enzimas na produção do fio (Dong et al., 2016).

Para a produção do casulo a lagarta secreta a seda por uma estrutura chamada de fiandeira, localizada entre os palpos labiais (Gallo et al., 2002) (Figura 2). Durante

a secreção do fio ela gira ao redor do corpo, se envolvendo com o fio. O casulo tem uma função protetora, protegendo o animal das condições ambientais adversas, como o ataque de animais, bactérias, alterações de temperatura e umidade, sendo essencial para a sobrevivência do inseto (Roy, et al., 2012).

Figura 2 – Posição da fiandeira na cabeça da lagarta (Gue et al., 2019)



Três dias após tecer o casulo, ela se transforma em pupa, também chamada de crisálida. Durante essa fase o hormônio do crescimento ecdisona, ativa a expressão de vários genes, promovendo a morte celular programada e a proliferação de novos tecidos, rearranjando todo o tecido do inseto para formar o adulto, que eclode do casulo após cerca de 13 dias (Tsuchida e Wells, 1988 e Vieira, 2020).

O inseto adulto chamado de mariposa possui corpo robusto coberto por pelos brancos onde os machos apresentam tamanho menor que as fêmeas. Ao longo de centenas de anos, devido ao processo de melhoramento genético para obtenção de

fios de altíssima qualidade, outras características desses animais foram consideradas secundárias, entre elas a capacidade de voar (apstesia). Por esse motivo, o animal tornou-se totalmente dependente do manejo humano, tanto para a alimentação quanto para a reprodução (Gallo et al., 2002).

De acordo com a origem geográfica, existem quatro raças para *Bombyx mori*: chinesa, japonesa, indiana e europeia (Krishnaswaml et al., 1979). Essas raças podem apresentar diferentes voltinismos: monovoltino, ou seja, com um ciclo anual; bivoltino, com dois ciclos anuais e polivoltinismo com vários ciclos anuais (Munhoz, 2010).

2.2 *BOMBYX MORI* E A SERICICULTURA

O bicho-da-seda possui uma grande importância econômica, científica e social, sendo um inseto utilizado há séculos pelos seres humanos. Hoje as principais áreas de interesse que utilizam o *B. mori* são a sericicultura para a produção de fio de seda, a ciência como organismo modelo, e a indústria de alimentos, como a produção de ração ou até alimentos de consumo humano.

A sericicultura é a prática agropecuária de criação do bicho-da-seda para retirada do fio encontrado nos casulos, sendo praticada pela humanidade há cerca de 5.000 anos (Nagaraju 2002). Ela consiste no cultivo da amoreira como alimento para as lagartas, criação das lagartas e retirada do casulo para a produção do fio. A atividade surgiu na China, que manteve segredo da produção por séculos (Oliveira et al., 2016). Devido a esse segredo muito bem guardado, povos que se encontravam com os chineses eram obrigados a comprar sua seda, se quisessem usufruir do produto. Isso acabou por desenvolver uma rota de comércio, chamada de rota da seda. Essa rota permitiu uma rede de comunicação entre a China e os portos do

Mediterrâneo, conectando a Ásia à Europa (Palazzo, 2009). Ela atravessava a plataforma continental euroasiática, permitindo o compartilhamento de culturas e conhecimento de povos que antes estavam isolados, o que levou ao desenvolvimento de várias regiões (Mendonça, 2016).

Essa atividade agrícola foi introduzida no Brasil com a chegada do Imperador Dom João VI, porém a produção só foi realmente intensificada mais de um século depois, por volta de 1920, graças a chegada dos imigrantes japoneses que já possuíam a expertise para a atividade (OLIVEIRA et al., 2016). No Brasil a sericicultura está concentrada no Vale da Seda paranaense. Atualmente o país é o sexto maior produtor de seda do mundo (International Sericultural Commission, 2023), tendo produzido em 2019 aproximadamente 3 mil toneladas (Cirio et al., 2021). A seda brasileira se destaca pela sua qualidade, seus fios podem chegar a 1,2 metros, sendo maiores e mais brancos do que os da China por exemplo (Cirio et al., 2021). O Paraná é responsável por 83,9% da produção de casulos do país e, em 2020, o valor bruto da produção gerada no estado foi de R$ 43.626.508,20 (Cirio et al., 2021).

A atividade sericícola gera uma boa renda por nove meses utilizando pequenas áreas de exploração, estimula a agricultura familiar, possibilita uma diversificação nas pequenas propriedades, introduz a mulher no campo, sendo economicamente viável, socialmente justa e ambientalmente correta (Yamaoka & Junior, 2019).

Devido à grande importância econômica do bicho-da-seda, esse inseto tem sido estudado exaustivamente pela humanidade durante séculos, o que acabou por gerar dados sobre sua biologia, tornando-o um candidato perfeito para a utilização como organismo modelo para diversas pesquisas.

2.3 GENÉTICA DE *BOMBYX MORI*

O *Bombyx mori* possui um genoma com tamanho de 475 Mb, sendo maior que o de *Drosophila*, *Anopheles* e abelhas de mel (Osanai-Futahashi et al., 2008). O número estimado de genes é de 14.623, sendo que desses genes 3.223 não são homólogos a nenhum outro inseto ou vertebrado (International Silkworm Genome Consortium, 2008). No entanto, cerca de metade do genoma, ou seja, 43,6%, é composto por sequências repetitivas (International Silkworm Genome Consortium, 2008). Além disso, grande parte dessas sequências repetitivas representa elementos transponíveis (ETs), sendo as principais classes os retrotransposons SINEs e os non-LTRs (Osanai-Futahashi et al., 2008).

O número haploide de cromossomos é 28, e seus tamanhos não se diferenciam muito, sendo que o maior é cerca de três vezes o tamanho do menor. A determinação do sexo é pelo sistema ZW, sendo que, para *B. mori*, a ausência do cromossomo W determina o macho (Z0 ou ZZ). Sendo assim a fêmea é o sexo heterogamético (ZW). Curiosamente, só foi relatado crossing over no sexo homogamético, característica essa só vista em *Drosophila* (Tanaka, 1957).

O bicho-da-seda também é capaz de realizar partenogênese. Isso acontece porque, em alguns ovos após a primeira divisão meiótica, os dois núcleos resultantes se unem novamente restabelecendo a diploidia. Esse fenômeno pode ser estimulado ao inserir ovos em uma solução quente com ácido clorídrico. Ambos os sexos podem ser partenogenéticos, porém as larvas serão sempre mais fracas que aquelas resultantes da reprodução sexuada (Tanaka, 1957).

Também é conhecida uma variedade de mutações para o bicho-da-seda. Mais de 400 mutações mendelianas já foram descritas para a espécie, sendo que a maioria surgiu espontaneamente durante a massiva criação desse animal. A maioria dos *loci*

descritos, possuem um ou alguns alelos conhecidos, porém há exceções, como o *locus p* que possui 10 alelos conhecidos e o *locus E* que possui 35 alelos conhecidos (Goldsmith et.al, 2005).

Vários experimentos de edição gênica já foram feitos com *B.mori* , para os diversos propósitos, como a descoberta da função de um gene, controle da expressão gênica e até o desenvolvimento de linhagens resistentes a vírus (Baci et al., 2021). Como exemplo temos o trabalho de Chen et al. (2017), que, utilizando a técnica CRISPR-Cas 9, desenvolveram lagartas transgênicas capazes de expressar a enzima Cas9 e RNAs guias que atacam dois genes (*ie -1 e me53*) relacionados a replicação do Baculuvírus, uma das principais ameaças ao bicho-da-seda. As lagartas transgênicas, além de apresentarem uma maior resistência ao vírus do que as selvagens, também foram capazes de transmitir essa resistência a outras gerações.

2.4 TRANSCRIPTOMA

Para a realização desse trabalho, o alvo de estudo foi o transcriptoma de *B. mori*. O transcriptoma é o conjunto completo de RNAs transcritos a partir de um genoma (Pierce, 2017).

O estudo do transcriptoma pode ser vantajoso pois ele pode fornecer informações que o genoma em si não oferece. A aplicação mais simples do transcriptoma nos estudos de biologia seria a caracterização de espécies. Como este representa o produto da transcrição de um genoma, frente a situações experimentais específicas, sendo principalmente determinado por genes codificadores de proteínas, isso pode fornecer uma grande quantidade de marcadores moleculares, os quais podem ser usados para diferenciar as espécies (Wolf, 2013). O transcriptoma também permite investigar diferenças nas expressões gênicas de diferentes populações, o que

possibilita estudos de vários tipos, como foi feito por Wolf e colaboradores em 2010, com o corvo-comum e a gralha cinzenta, que mediram o nível divergência dessas duas espécies. Nesse estudo eles mostraram que a divergência na expressão possivelmente evolui mais rápido do que a divergência nos nucleotídeos do DNA.Com o uso do transcriptoma também é possível estudar o comportamento de um grupo de células ou até mesmo de uma única célula frente a estímulos externos. Cada célula possui seu próprio transcriptoma mesmo quando compartilha o mesmo DNA, isso permite determinar as redes que regulam os genes e, através de nocaute de genes é possível estudar qual a função deles e como eles regulam a expressão gênica (Tang & Surani, 2011).

O DNA sequenciado, mais precisamente o cDNA para estudos transcriptômicos, chega na forma de pequenas sequências de leitura chamadas de *reads*. Os *reads* chegam embaralhados, sendo necessário organizá-los para que seja reconstruída a sequência real ou próxima do real do RNA expresso. Esse processo de organização dos *reads* é chamado de montagem, e consiste na identificação de sobreposições entre os *reads* e estabelecimento da sua ordem real (Florea & Salzberg, 2013). Atualmente, existem dois tipos de montagem de um transcriptoma a partir de sequências de RNAseq: a guiada por genoma de referência e a *de novo*.

A montagem guiada por genoma de referência, como o próprio nome sugere, utiliza um genoma já anotado de referência para a reconstrução do transcriptoma de interesse. Esse método pode ser dividido em três etapas: primeiramente, os *reads* são alinhados ao genoma de referência usando programas como Blat e TopHat, que são capazes de alinhar *reads* de cDNA a um genoma. Em seguida, os *reads* sobrepostos de um *locus*, são agrupados para formar um gráfico que representa todas as isoformas

possíveis. Por fim, o gráfico é percorrido, identificando os transcritos e suas diferentes isoformas (Martin & Wang, 2011).

Há diversos montadores que utilizam a estratégia guiada, como Cufflinks (Trapnell et al., 2010) e o StringTie (Pertea et al., 2015). Todos seguem o mesmo princípio, mas o método pelo qual ocorre a montam os gráficos e a união dos diferentes *reads* muda de acordo com o algoritmo que cada um utiliza, o que pode resultar em montagens diferentes para um mesmo transcriptoma.

Um dos modelos de gráfico utilizados pelos montadores é o *overlap graph*. Este modelo é encontrado, por exemplo, no Cufflinks, um dos montadores mais utilizados, onde seus nós representam os *reads* (Pertea et al., 2015). Dois nós são conectados por arestas, se os *reads* que eles representam se sobrepõem e se compartilham os mesmos padrões de *splicing* (fenômeno de retirada de íntrons durante a formação de um RNA) (Florea & Salzberg, 2013).

Outro modelo é o *splice graph*, em que os nós representam os éxons e as arestas significam os íntrons. Neste gráfico os nós são conectados de diferentes formas de acordo com a sua compatibilidade, permitindo a representação de todas as isoformas possíveis (Florea & Salzberg, 2013 e Pertea et al., 2015). Um montador guiado que utiliza esse gráfico é o StringTie. Este montador é um pouco diferente dos demais, pois além de utilizar a abordagem guiada, também pode empregar metodologias da montagem *de novo* (que serão abordadas adiante), permitindo que ele utilize os *reads* que não se alinharam ao genoma para melhorar a montagem. (Pertea et al., 2015). O StringTie primeiro reúne os *reads* em agrupamentos (*clusters*), e, em seguida, cria uma *splice graph*, unindo cada *cluster*. Isso permite a identificação dos transcritos e suas isoformas (Pertea et al., 2015).

A montagem guiada possui várias vantagens, por exemplo, não sofrer tanto com problemas de contaminação com RNAs de outras espécies, uma vez que *reads*, que não se alinharem ao genoma de referência, geralmente não são utilizados na montagem. Além disso, essa estratégia é muito sensível podendo montar transcritos com baixa abundância (Pertea et al., 2015). A principal desvantagem da montagem guiada é a dependência de um genoma de referência bem anotado (Pertea et al., 2015). Sendo assim, o estudo do transcriptoma de organismos que não são modelos, podem ser prejudicados utilizando essa estratégia, pois geralmente seus genomas estão parcialmente anotados.

A segunda estratégia de montagem é a *de novo*. Essa estratégia não utiliza genoma de referência, se baseando apenas nas sobreposições entre os *reads* para a construção dos transcritos (Pertea et al., 2015).

Para determinar quais são as melhores conexões entre os *reads*, os montadores *de novo* utilizam os gráficos de Brujin. Diferentemente dos outros gráficos que organizam os elementos de acordo com os *reads*, nesse gráfico, os elementos são organizados de acordo com k-mers (Zerbino & Birney, 2008). No caso dos estudos genômicos, os k-mers são sequências contendo k nucleotídeos. Por exemplo, a sequência ATCGAA, ela tem tamanho igual a 6 (n=6), se dividirmos ela com uma sobreposição de 3 (k=3), temos quatro sequências: ATC, TCG, CGA, GAA, ou seja 4 k-mers (Rosalind, 2013).

No gráfico de Brujin, cada nó representa uma série de k-mers sobrepostos. As sequências desses nós são representadas pelos últimos nucleotídeos de cada k-mer que os compõem. Além disso, cada nó está junto com um gêmeo representado pela sequência complementar a dele, essa união é chamada de bloco. Dois nós diferentes são conectados por arestas, se houver uma sobreposição do último k-mer de um nó

de origem com o primeiro k-mer do outro nó de destino (Zerbino & Birney, 2008). Dessa maneira os montadores *de novo*, podem enumerar todos os caminhos possíveis que ligam os *reads* e, então, através de um sistema de pontuação, determinar quais são os transcritos mais plausíveis (Grabherr et. al, 2011).

Existem vários montadores *de novo*, como The Rnnotator, Multiple-k, Trans-ABySS e Trinity (Pertea et al., 2015). Dentre esses montadores, o Trinity possui uma característica especial. Ele recebe esse nome pois é a junção de três softwares, *Inchworm*, *Chrysalis* e *Butterfly*. O *Inchworm*, responsável pela primeira etapa, é baseado em k-mers, montando os *reads* em sequências chamadas de *contigs*. Ele recupera apenas os *contigs* que possuem a melhor representação dentro do conjunto de possíveis variáveis. Em seguida, o *Chrysalis* agrupa os *contigs* que estão relacionados e monta um gráfico de Brujin para cada agrupamento. Por fim, o *Butterfly* analisa as conexões de cada gráfico de Brujin e determina quais são os transcritos mais plausíveis, bem como as suas isoformas (Grabherr et. al., 2011).

A principal vantagem da montagem *de novo*, independe de um genoma de referência, permitindo o estudo do transcriptoma de organismos que possuem poucas ou nenhuma informação genética. Porém, esse método requer mais recursos computacionais e uma profundidade de sequenciamento maior com relação a montagem guiada. Além disso, está mais suscetível a erros, porém, atualizações e novos montadores estão constantemente sendo desenvolvidos para corrigir tais erros (Martin & Wang, 2011).

Outra forma de reconstruir um transcriptoma é combinar as duas montagens. É interessante juntar as duas montagens, pois é possível que as deficiências de uma estratégia podem ser supridas pela outra estratégia. Por exemplo, uma montagem

híbrida pode se beneficiar da alta sensibilidade da montagem guiada e da capacidade da montagem *de novo* para detectar novos transcritos (Martin & Wang, 2011).

Alguns autores já demonstraram que a montagem híbrida pode proporcionar melhorias. Um exemplo é o estudo de Lu et al. (2013), no qual demonstrou que a montagem híbrida demandou menos tempo e utilizou menos memória que a montagem *de novo*. Além disso, apresentou maior precisão, com relação a outras montagens em um transcriptoma de cérebro humano.

Podemos realizar a montagem híbrida de duas formas distintas. Uma delas é o alinhamento seguido de montagem (*Align-then-assemble*), que implica iniciar com a montagem guiada e, posteriormente, realizar uma montagem *de novo* com os *reads* que não se alinharam. A outra abordagem seria montagem seguida do alinhamento (*Assemble-then-align*), que envolve realizar primeiramente uma montagem *de novo* e, com os *contigs* formados, efetuar uma montagem guiada. Essa estratégia é mais recomendada quando o genoma de referência é pouco anotado, ou quando a única referência é o genoma de uma espécie próxima à de interesse (Martin & Wang, 2011).

Com relação ao estudo do transcriptoma de *Bombyx mori*, temos como exemplo o trabalho de Lopes et al. (2023), que comparou o transcriptoma de duas populações de lagartas criadas a diferentes temperaturas (26°C e 34°C). As lagartas expostas a maior temperatura tiveram uma diminuição da expressão de vários genes, como por exemplo aqueles relacionados a produção de fibroína e com a resposta imune, como inibidores de serina protease, inibidores de apoptose e resposta a defesa contra bactérias. Esse resultado mostrou o quanto a temperatura de criação da lagarta pode afetar sua saúde e, consequentemente, a produção de seda.

Outro exemplo de estudo utilizando o transcriptoma de *Bombyx mori* é o de Zheng et al. (2022), que utilizaram o bicho-da-seda como animal modelo para

investigar a influência do pesticida DMT (Dimetoato) em insetos que não são seu alvo. Essa questão é importante, uma vez que esse veneno permanece no meio ambiente, podendo afetar espécies que não causam prejuízo. O estudo utilizou uma análise de expressão genica digital (DGE) e RT-qPCR, para avaliar a expressão genica de ovos de *B.mori* que foram expostos ao DMT. O estudo apresentou uma série de genes diferencialmente expressos, principalmente relacionados coma promoção de transportadores de Trealose, a resposta ao estresse, proteínas dedo de zinco e proteínas ligadas à via 5-HT. De acordo com os autores, resultados como esse podem fornecer informações valiosas para os estudos de toxicologia de insetos e estudos sobre os efeitos do DMT no bicho-da-seda.

2.7 ANÁLISE DE EXPRESSÃO DIFERENCIAL

Análise de expressão diferencial (DGE, *differential gene expression.*), é uma das diversas análises possíveis a partir de um transcriptoma montado, sendo a principal aplicação do RNA-seq (Stark et. al, 2019). Com esse tipo de análise, é possível formular hipóteses sobre a relação entre expressão gênica e fenótipo (Muzellec et. al, 2023).

A DGE permite determinar quantitativamente as diferenças na expressão dos genes entre grupos experimentais (Stark et. al, 2019). Essa análise tem início no laboratório, com a extração de RNA e preparação das bibliotecas de cDNA dos grupos a serem comparados, como a realizada neste trabalho, com lagartas *Bombyx mori* expostas a diferentes temperaturas. As bibliotecas então devem ser sequenciadas e posteriormente reconstruídas por meio da montagem do transcriptoma. Por fim, utilizando softwares específicos, realiza-se a análise dos níveis de expressões dos genes em cada grupo avaliado, a fim de determinar quais genes tiveram sua

expressão alterada de acordo com as condições que cada grupo do experimento foi exposto (Stark et. al, 2019).

Existem diversas ferramentas que possibilitam esse tipo de análise e, assim como os montadores, cada uma utiliza uma estratégia diferente de acordo com seu algoritmo. Algumas baseiam-se no nível de expressão do gene, enquanto outras dependem de estimativas da transcrição (Stark et. al, 2019). Algumas delas são programadas em phyton como a PyDESeq2 (Muzellec et. al, 2023) e outras são programas em R como o DEseq2 (Huber, 2017) utilizado neste trabalho.

3 OBJETIVOS

O objetivo específico desta pesquisa é comparar dois métodos de montagem de transcriptoma, guiada e de novo, em relação à eficácia na produção total de transcritos, quantidade de proteínas anotadas e diferencialmente expressas em condições experimentais específicas. Para alcançar esse objetivo, os seguintes objetivos específicos foram delineados: avaliar a qualidade de cada montagem realizada, analisar a produção total de transcritos gerados em cada método de montagem, avaliar a quantidade de proteínas anotadas em cada método de montagem e analisar as proteínas diferencialmente expressas entre as condições experimentais utilizando ambos os métodos de montagem de transcriptoma.

4 MATERIAL E MÉTODOS

4.1 Criação das lagartas e Extração de RNA

O RNA foi extraído de seis lagartas *B. mori*, sendo três delas criadas a 26ºC, as quais constituíram o grupo controle, e três criadas a 34ºC. As bibliotecas de cDNA foram preparadas utilizado o protocolo *dUTP*-stranded e um mínimo de 20 milhões

de paired-end reads (2x125-bp) por amostra foram sequenciados por Illumina HiSeq 2500, essa etapa foi realizada por Lopes et al. (2023). Portanto, o presente trabalho concentrou-se exclusivamente na análise de bioinformática, relacionada à comparação das montagens.

4.2 Montagem e Anotação dos Transcritos

As etapas de montagem das bibliotecas de RNAseq e anotação dos transcritos foram realizadas em um Servidor Linux. A qualidade das bibliotecas obtidas no sequenciamento foram analisadas usando o programa FastQC v0.11.7 (ANDREWS, 2017). As bibliotecas de RNASeq foram limpas com programa TrimGalore v0.6.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), sendo removidas as sequências adaptadoras, bem como os *reads* menores que 25 pares de bases e com índice de qualidade inferior a 20 na escala de Phred.Para a comparação dos métodos de montagem, foi empregados dois montadores: Trinity v2.8.5, responsável pela montagem *de novo*, e o StringTie v2.1.1, que realiza montagem guiada. Neste último caso, foi utilizado somente a função guiada do StringTie e o genoma de *B. mori* (GCF_014905235.1) de referência obtido junto ao NCBI (National Center for Biotechnology Information).

A anotação dos transcritos foi conduzida por meio da ferramenta Blastx em servidor com sistema operacional Linux. Com o objetivo de avaliar possíveis diferenças na anotação dos transcritos em função do banco de dados utilizados no alinhamento, foram realizados três alinhamentos diferentes para cada montagem: um deles utilizou o banco de dados de *B.mori* (GCF_014905235.1) do NCBI com 27.309 peptídeos, outro utilizou o banco de dados de Lepidoptera do UniProt com 1.037.030

peptídeos e o terceiro utilizou o banco de dados de Lepidoptera do NCBI com 2.476.778 peptídeos.

Além dos três alinhamentos Blastx, para determinar a se os transcritos montados pertenciam a famílias proteicas, domínios proteicos ou a outro tipo de sequência proteica, foi realizado um alinhamento ao banco de dados Pfam.

4.3 Qualidade da montagem

Para avaliar a qualidade da montagem foi utilizado três parâmetros com base no trabalho de Lu e colaboradores em 2012, sendo estes: Completude, Número Médio de Isoformas e Tamanho Médio de Nucleotídeos. Para a determinação da Completude ou alinhamento total, foi utilizada a ferramenta BUSCO (Benchmarking Universal Single-Copy Orthologs) com um banco de dados de Insecta. Com a ferramenta BUSCO também foi comparado o número de transcritos alinhados completos e cópia única, completos e duplicados, fragmentados e não encontrados.

Para determinar o número médio de isoformas, utilizou-se a identificação de cada transcrito. Cada montador atribui uma identificação ao transcrito; se ele tiver uma isoforma, o montador repete a identificação e adiciona um indicativo de isoforma. Assim, para calcular a média de isoformas, primeiro verificou-se quantas vezes o nome de cada transcrito se repete. Em seguida, somou-se o número de repetições e dividiu-se pelo número total de transcritos, sem considerar suas isoformas.

O Tamanho Médio de Nucleotídeos foi calculado somando o tamanho de nucleotídeos de cada transcrito montado e dividindo pelo número total de transcritos, incluindo suas isoformas. Cada montador fornece o número de nucleotídeos em cada transcrito nos resultados de sua montagem. Uma planilha eletrônica foi utilizada para processar os dados de montagem de cada programa.

Além dessas métricas mencionadas, também foi avaliada a porcentagem de *reads* sequenciados que se alinharam aos transcritos montados, a porcentagem de *reads* que se alinharam uma vez aos transcritos montados e porcentagem de *reads* que se alinharam mais de uma vez aos transcritos montados. Esses dados foram obtidos por meio do alinhamento dos *reads* ao transcriptoma montado de cada ensaio, utilizando as ferramentas BowTie2 (Trinity) e Hisat2 (StringTie).

4.4 Análise de expressão diferencial e enriquecimento

A análise de expressão diferencial para cada montagem foi realizada como o programa DESeq2 v1.34 (Bioconductor, 2003), utilizando os seguintes parâmetros para identificação dos transcritos diferencialmente expressos: padj < 0,05 e logFoldChange ≥ 2,0 e ≤ -2,0.

Além da avaliação de expressão diferencial, também foi investigado em quais vias metabólicas os transcritos diferencialmente foram associados. Essa caracterização foi realizada através do enriquecimento gênico, utilizando a ferramenta ShinyGo v0.77 (Steven Xijin Ge et al., 2020).

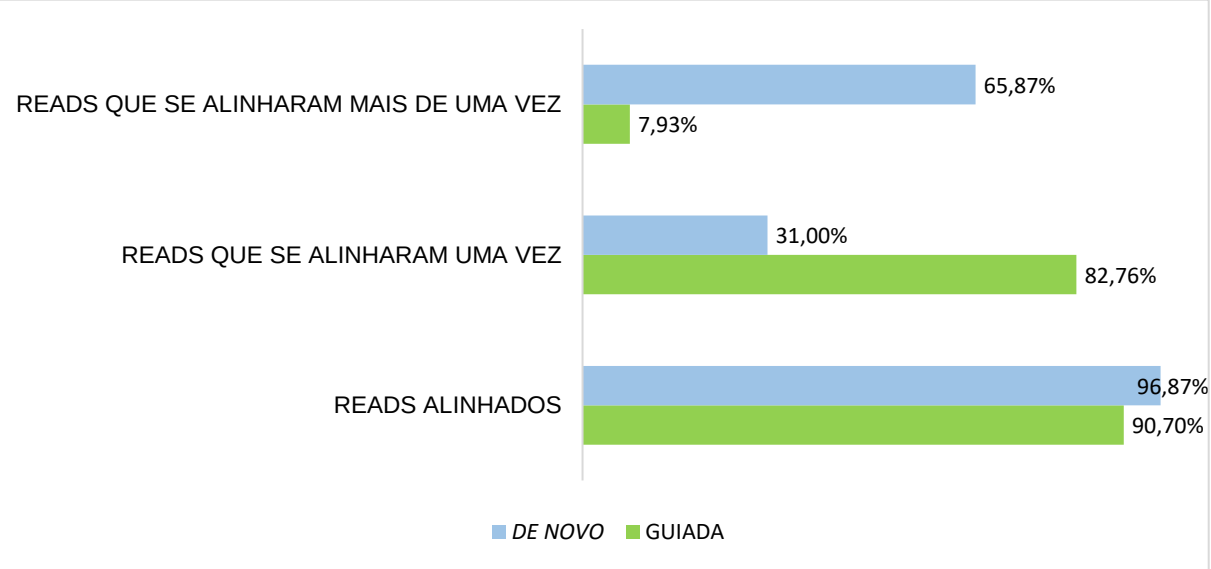
5 RESULTADOS

5.1 Qualidade da montagem

Com relação ao alinhamento dos *reads* nos transcritos montados, ambas as montagens obtiveram a maioria dos *reads* alinhados. De um total de 181.870.662 *reads*, 96,87% se alinharam ao transcriptoma construído na montagem guiada, e 90,70% ao transcriptoma construindo na montagem *de novo*. Isso revela que os dois montadores, foram capazes de utilizar a maior parte dos *reads* obtidos da extração de RNA (Figura 3).

82, 76% dos *reads* se alinharam apenas uma vez na montagem guiada, enquanto 31% se alinharam apenas uma vez na montagem *de novo.* Por outro lado, 7,93% dos *reads* se alinharam mais de uma vez na montagem guiada e 65,97% dos *reads* se alinharam mais de uma vez na montagem *de novo.* Essa diferença pode ter ocorrido, pois talvez a montagem *de novo* utilize mais vezes os *reads*, para montar seus transcritos, justificando a maior quantidade de *reads* se alinhando mais de uma vez (Figura 3) .

Figura 3 - Total de *reads* que se alinharam em cada montagem



Levando em consideração o banco de dados de Insecta do BUSCO com um total de 1367 grupos procurados, a montagem guiada resultou 92,80% de alinhamento total (Completude), sendo 56,50% completo e cópia única, 36,30% completo e duplicada, 2,6% fragmentada e 4,6% não encontradas. Para a montagem *de novo*, esses valores foram 89,70% de alinhamento total, sendo 48,50% completo e cópia única, 41,20% completo e duplicada, 4,50% fragmentada e 5,80% não encontradas. Ambas as montagens obtiveram uma completude semelhante, porém a guiada foi

superior nesse parâmetro. Isso possivelmente tem relação ao uso de um genoma de referência, que pode favorecer a montagem de transcritos completos (Quadro 1).

O número médio de isoformas foi semelhante nas duas montagens, sendo uma média de 1,57 isoformas por transcrito na montagem guiada, e uma média de 1,52 isoformas por transcrito na montagem *de novo*. Essas isoformas podem ser transcritos verdadeiros ou artefatos das montagens, necessitando mais análises para diferenciá-los. Era esperado que na montagem *de novo*, a média de isoformas fosse superior devido a não utilização de um genoma de referência e a quantidade superior de transcritos montados (Quadro 1).

Com relação ao tamanho médio de nucleotídeos por transcrito, a montagem guiada montou maiores sequências, com uma média de 1.902,53 nucleotídeos, enquanto a montagem *de novo* obteve uma média de 938,21 nucleotídeos. Novamente o tamanho superior das sequências da montagem guiada, possivelmente está relacionada a utilização de um genoma de referência (Quadro 1).

Quadro 1 - Qualidade das montagens

Montagem	**Guiada**	***De novo***
Alinhamento total ou Completude (%)	92,8	89,7
Completo e cópia única (%)	56,5	48,5
Completo e duplicado (%)	36,3	41,2
Fragmentado (%)	2,6	4,5
Não encontrado (%)	4,6	5,8
Número médio de isoformas por transcrito	1,57	1,52
Tamanho médio de nucleotídeos por transcrito	1902,53	938,21

5.2 Quantidade de transcritos montados e proteínas anotadas

A montagem guiada resultou em 29.157 transcritos, enquanto a montagem *de novo*, 73.743. Do total de transcritos obtidos em cada montagem, foi possível anotar com o banco de dados de *B. mori* do NCBI, 13.726 peptídeos na montagem guiada e 13.702 peptídeos na montagem *de novo*, dois quais 11.793 (75,43%) foram comuns a ambas as montagens, 1.933 (12,36%) foram anotados somente na montagem guiada e 1.909 (12,21%) na montagem *de novo*, totalizando 15.635 peptídeos anotados (Figura 4 A).

Com o banco de dados de Lepidoptera do NCBI, foi possível anotar 15.629 peptídeos na montagem guiada e 19.081 peptídeos na montagem *de novo*, dos quais 10.844 (45,44%) foram anotados em ambas as montagens, 4.785 (20,05%) foram anotados somente na montagem guiada e 8.237 (34,51%) foram anotados na montagem *de novo*, totalizando 23.866 peptídeos anotados (Figura 4 B).

Com o banco de dados de Lepidoptera do Uniprot, foi possível anotar 15.400 peptídeos na montagem de guiada e 18.693 peptídeos na montagem *de novo*, dos quais, 10.922 (47,14%) foram anotados em ambas as montagens, 4.478 (19,32%) anotados somente na montagem guiada e 7.771 (33,54%) anotados somente na montagem *de novo* totalizando 23.171 peptídeos anotados (Figura 4 C).

Quanto ao alinhamento ao banco de dados do Pfam, foram anotadas 4.257 sequências na montagem guiada e 4.228 na montagem de novo. Dessas 4.108 (93,85%) foram anotadas em ambas as montagens, 149 (3,38%) foram anotadas na montagem guiada e 120 (2,71%) na montagem de novo, totalizando 4.377 sequências anotadas (Figura 4 D). Notavelmente 4.031 (92,10%) dessas sequências representavam famílias ou domínios proteicos, sendo o restante compostos por sequencias repetitivas, desordenadas, *motif* ou *coiled-coil*, evidenciando que ambos

os montadores possivelmente, foram capazes de gerar uma quantidade significativa de transcritos que verdadeiramente representam proteínas (Figura 5).

Figura 4 Diagramas de Venn mostrando a quantidade de proteínas exclusivas e compartilhadas anotadas em cada montagem: A) Proteínas anotadas usando o banco de dados de B.mori do NCBI. B) Proteínas anotadas usando o banco de dados de Lepidoptera do NCBI. C) Proteínas anotadas usando o banco de dados de Lepidoptera do Uniprot. D) Proteínas anotadas usando o banco de dados Pfam.

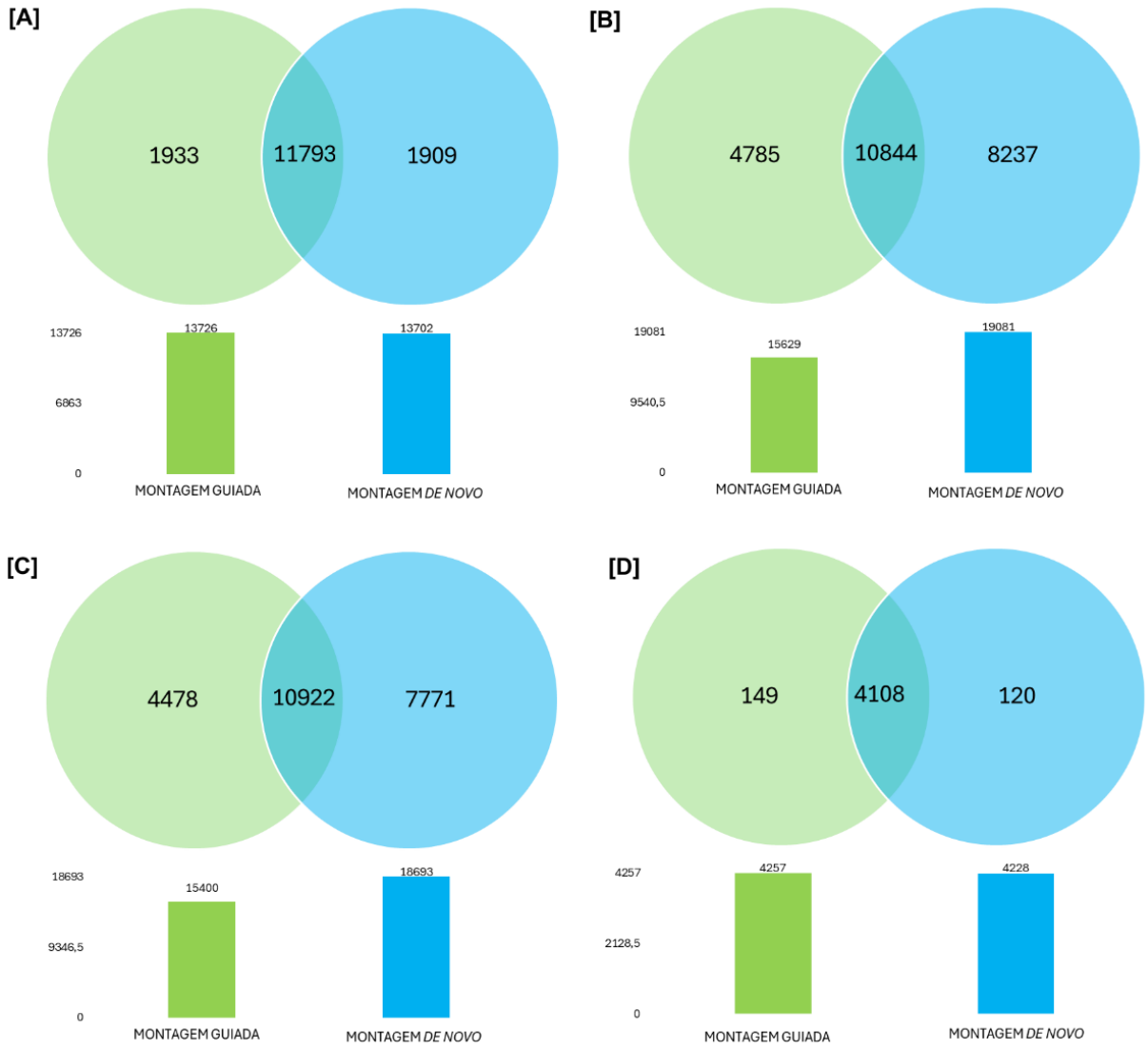
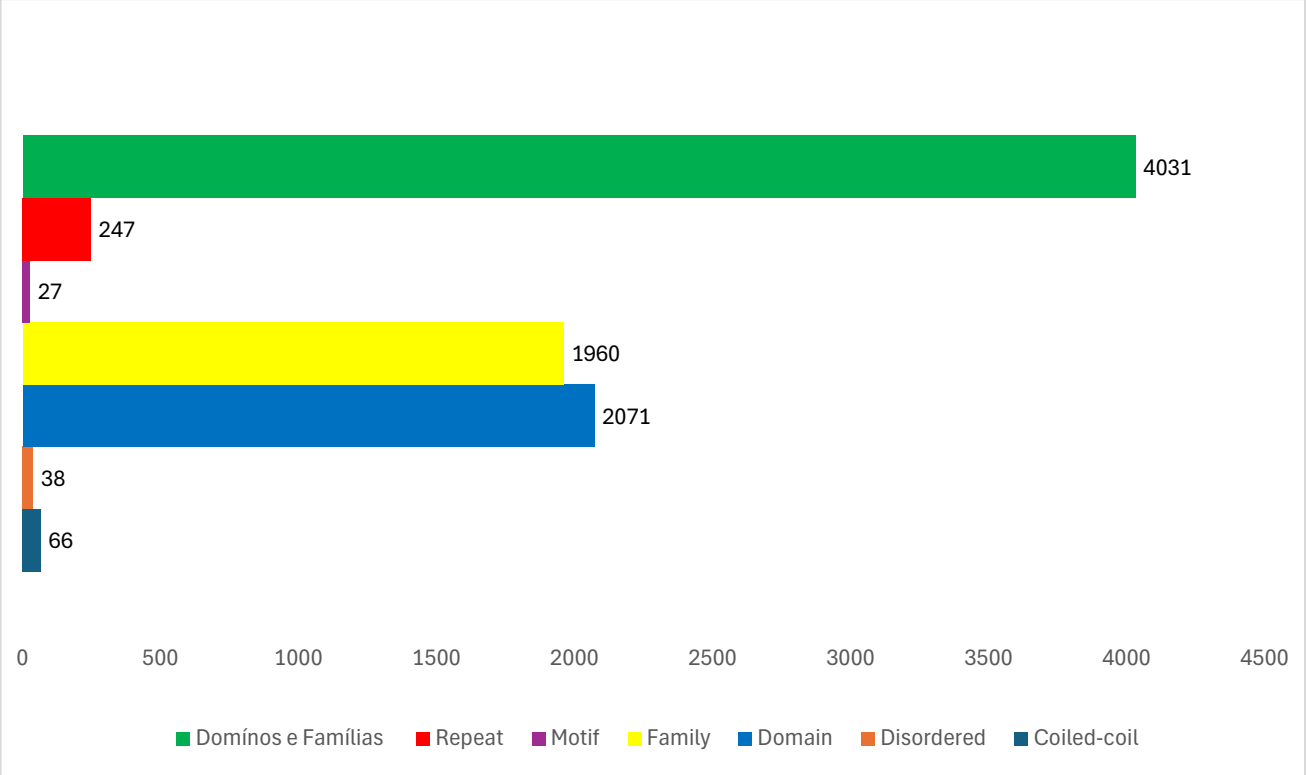


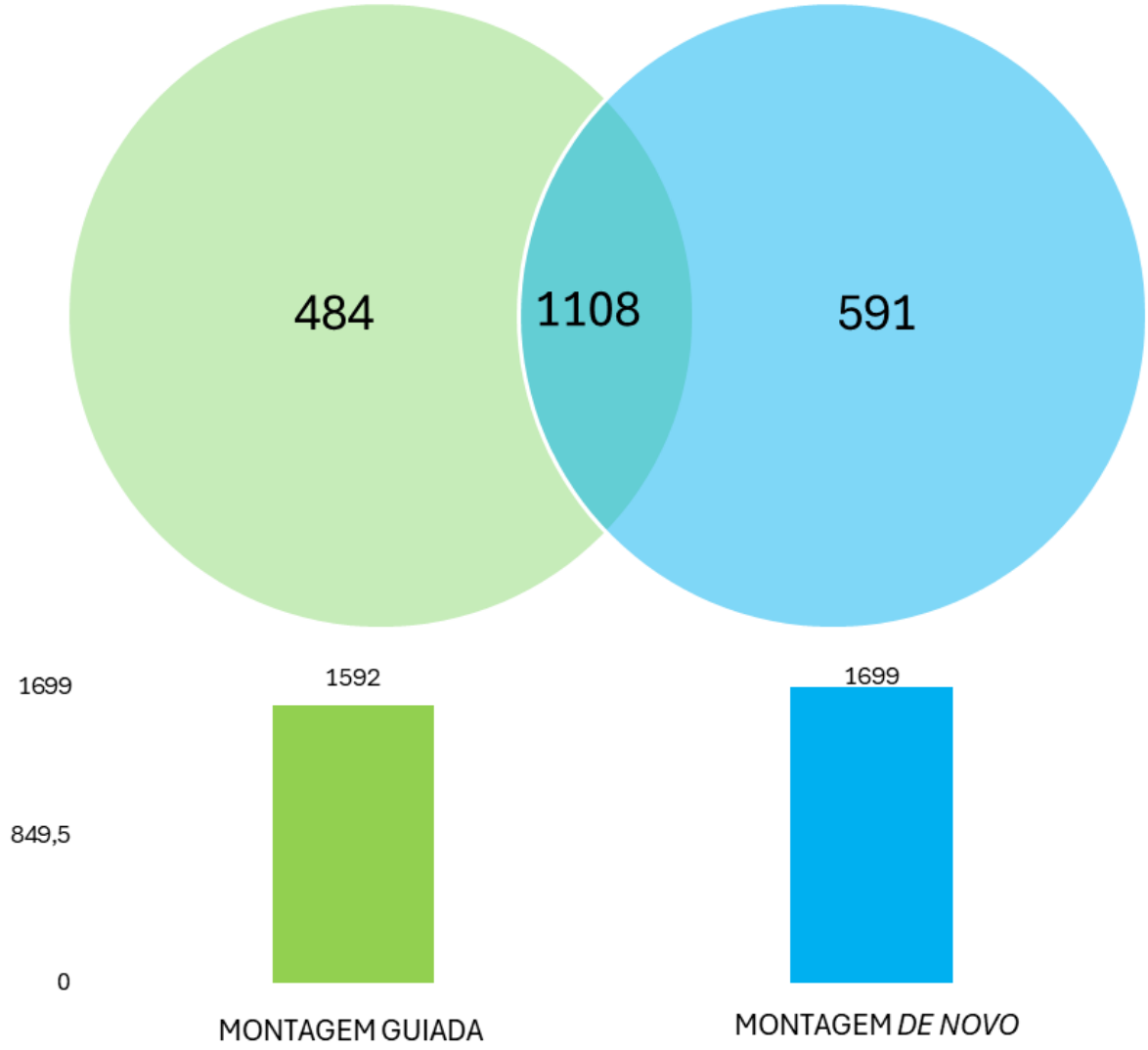
Figura 5 - Quantidade e tipos de sequencias anotadas com o banco de dados do Pfam para as duas montagens.



5.3 Análise de expressão diferencial

A análise de expressão diferencial realizada com os dados do alinhamento blastx do banco de dados de B. mori do NCBI com 27.309 sequencias peptídicas revelou que, na montagem guiada, 1.592 proteínas foram diferencialmente expressas e, na montagem de novo, 1.699 proteínas foram diferencialmente expressas. Dessas 1.108 (50,73%) foram comuns às duas montagens, 484 (22,16%) foram identificadas somente na montagem guiada e 591 (27,06%) foram identificadas somente na montagem de novo, totalizando 2.184 proteínas diferencialmente expressas identificadas (Figura 6).

Figura 6 - Diagrama de Venn de proteínas exclusivas e em comum identificadas em cada montagem, após a análise expressão diferencial



A montagem guiada obteve 18 vias metabólicas diferentes, enquanto a montagem *de novo* obteve 15. Dessas 12 (57,14%) vias metabólicas foram comum as duas montagens, 6 (28,57%) foram exclusivas da montagem guiada, e 3 (14,28%) exclusivas da montagem *de novo*, totalizando 21 vias metabólicas identificadas (Quadro 2). Ao analisar as vias metabólicas em que as proteínas diferencialmente expressas atuam, observa-se que a maioria delas está relacionada à resposta imune e à defesa contra outros organismos (Quadro 2 e Figura 7, 8)

Quadro 2 – Quantidade de vias metabólicas identificadas em cada montagem, após a análise expressão diferencial.

Tipo de via metabólica	Exclusivas na montagem *de novo*	Exclusivas na montagem guiada	Ambas as montagens
Atividades enzimáticas	1	1	3
Atividade de transporte	-	3	-
Biogênese e metabolismo	2	1	-
Região extracelular	-	-	1
Respostas imunes e defesa contra organismos	-	1	8
Total	3	6	12

Figura 7 - Vias metabólicas identificadas na montagem guiada

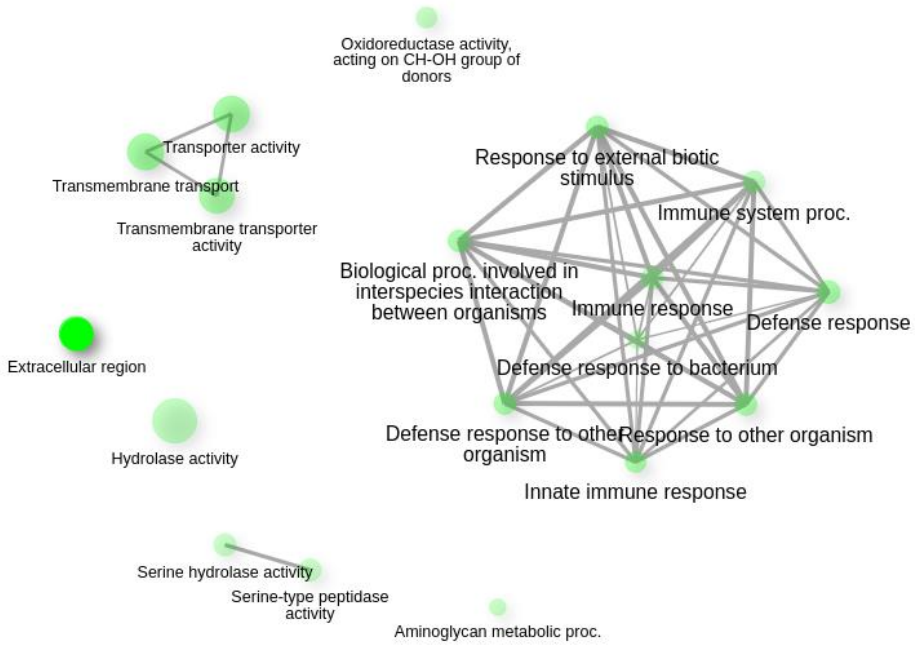
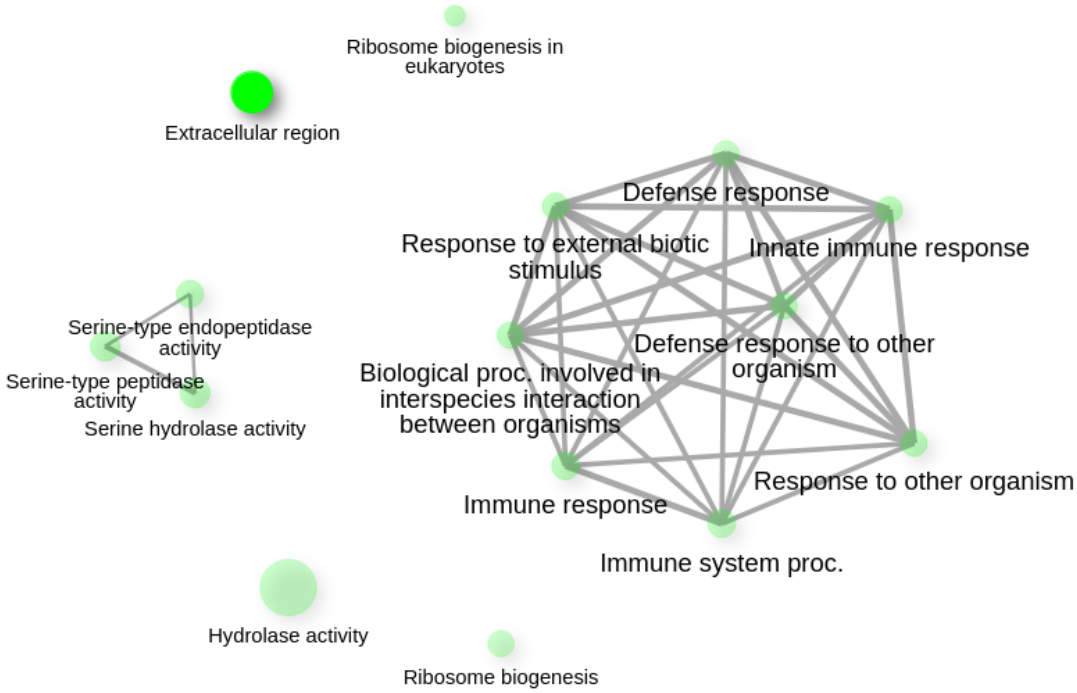


Figura 8 - Vias metabólicas identificadas na montagem *de novo*



6 DISCUSSÃO

As duas estratégias mostraram diversas semelhanças na reconstrução do transcriptoma. Elas apresentaram uma alta completude, um número médio de isoformas baixo, utilizaram a grande maioria dos *reads* fornecidos e a maioria de seus transcritos representavam domínios ou famílias proteicas. Porém algumas diferenças observadas podem ser úteis para novas pesquisas com transcriptoma.

A montagem *de novo* foi capaz de construir mais transcritos que a montagem guiada. Isso está de acordo com os resultados de Huang et al. (2016), que em um estudo comparativo entre estratégias de montagens, os montadores obtiveram um maior número de *contigs* em comparação com os guiados**.**

Quando esses transcritos foram submetidos a banco de dados mais abrangentes, ou seja, os de Lepidoptera, foi possível notar um aumento maior na anotação de transcritos na montagem *de novo* do que na montagem guiada. A montagem guiada teve um aumento de 1.903 transcritos anotados no banco de dados

de Lepidoptera do NCBI, enquanto a montagem *de novo* teve um aumento de 5.379. Com o banco de dados de Lepidoptera do Uniprot, a montagem guiada teve um aumento de 1.674 proteínas anotadas, enquanto a montagem *de novo* teve um aumento de 4.991. Isso pode significar que existem proteínas que ainda não foram catalogadas no banco de *B. mori* e que a utilização da montagem *de novo* aliado a banco de dados mais abrangentes pode ser uma boa alternativa para identificação de novas sequências proteicas.

Com relação a montagem guiada, apesar de ter resultado em menos transcritos, eles foram maiores e mais completos, além disso, essa estratégia teve uma maior porcentagem de *reads* alinhados uma única vez. Esses resultados também estão de acordo com o estudo Huang e colaboradores em 2016, em que os *contigs*, das montagens guiadas foram maiores.

Por outro lado, os dados mostram que ambos os montadores são eficientes, e possuem diferenças. Sendo assim, escolher qual estratégia utilizar em uma pesquisa depende, em primeiro lugar, da existência de genomas ou transcriptomas de referência disponíveis para a espécie analisada. Por outro lado, mesmo nessa situação, a utilização da montagem *de novo* pode ser uma estratégia que pode levar à descoberta de novos transcritos. Inclusive, como observado nesse trabalho, quando os dados obtidos por cada montagem são somados, é possível obter uma melhor anotação, uma vez que em todos os alinhamentos existiram proteínas exclusivamente anotadas, isto é, transcritos que apenas um dos montadores conseguiu construir. Isso aconteceu até mesmo na análise de expressão diferencial, em que juntando os dados de cada montador foi possível identificar mais vias metabólicas diferencialmente expressas. Isso pode ser um indicativo que a utilização de duas estratégias de montagens, ou uma estratégia híbrida pode ser vantajoso para obter um melhor

transcriptoma. Nesse sentido, Voshall e Moriyama (2018) sugerem que as limitações de uma ferramenta de montagem podem ser superadas pelas combinações dos resultados de múltiplas montagens. De acordo com eles, selecionando os *contigs*, montados corretamente por cada ferr

amenta e ignorando aqueles redundantes montados incorretamente, seria possível aproveitar os pontos fortes de cada método sem ser afetado pelos pontos fracos, efetuando assim uma montagem híbrida que utiliza o melhor de cada estratégia.

Lu et al. (2013), compararam a eficiência de uma montagem hibrida em relação a montagem guiada e *de novo*. Através do software Inchworm eles foram capazes de mesclar as duas estratégias, onde a montagem hibrida foi mais rápida que as montagens *de novo*; utilizou menos memória do que as outras montagens, apresentou baixo quimerismo e uma alta precisão, no entanto, apresentou baixa completude e detectou mais isoformas por gene. Para os autores, o Inchworm teve um desempenho desejado, porém ainda precisa de otimização.

Sendo assim levando em consideração os trabalhos citados acima e os resultados deste estudo, a estratégia hibrida pode ser uma boa alternativa para a melhoria da anotação dos genomas. Porém, ainda isso precisa ser mais bem pesquisado, permitindo o desenvolvimento de novos softwares que permitam realizá-la de forma mais eficaz e segura.

7 CONCLUSÃO

Esse estudo demonstra que as montagens guiadas e *de novo* para um mesmo conjunto de transcriptoma pode resultar em transcritos diferentes e complementares, algo que pode ser fundamental quando procuramos explicar os mecanismos de resposta dos organismos desafiados com diferentes condições experimentais.

REFERÊNCIAS

ABDELLI, N.; PENG, L.; KEPING, C. Silkworm, Bombyx mori, as an alternative model organism in toxicological research. **Environmental science and pollution research international**, v. 25, n. 35, p. 35048–35054, 2018.

ANDREWS, S. **FastQC: A quality control analysis tool for high throughput sequencing data**. [s.l: s.n.].

ARUGA, H. Silkworm and its Strains. **Oxford: CRC Press**, p. 367, 1994.

BACI, G.-M. et al. Advances in editing silkworms (Bombyx mori) genome by using the CRISPR-Cas system. **Insects**, v. 13, n. 1, p. 28, 2021.

CHEN, S. et al. Transgenic clustered regularly interspaced short palindromic repeat/Cas9-mediated viral gene targeting for antiviral therapy of Bombyx mori nucleopolyhedrovirus. **Journal of virology**, v. 91, n. 8, 2017.

CHENG, L. et al. Characterization of silkworm larvae growth and properties of silk fibres after direct feeding of copper or silver nanoparticles. **Materials & design**, v. 129, p. 125–134, 2017.

CIRIO, G. M.; PEREIRA, J. R.; DE PAULA, L. C. **PROGNÓSTICO AGROPECUÁRIO S E R I C I C U L T U R A RELATÓRIO ANUAL**. [s.l: s.n.].

DE WIT, P. et al. The simple fool's guideto population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. **Molecular Ecology Resources**, v. 12, p. 1058–1067, 2012.

DONG, Z. et al. Analysis of proteome dynamics inside the silk gland lumen of Bombyx mori. **Scientific reports**, v. 6, n. 1, 2016.

DURST, P. B. **Forest Insects as Food: Humans Bite Back : Proceedings of a workshop on Asia-Pacific resources and their potential for development, 19-21 February 2008, Chiang Mai, Thailand**. [s.l: s.n.].

DVOŘÁK, J. **Photo #79456: Bombyx mori**. Disponível em: <https://insecta.pro/gallery/79456>. Acesso em: 19 mar. 2024.

FLOREA, L. D.; SALZBERG, S. L. Genome-guided transcriptome assembly in the age of next-generation sequencing. **IEEE/ACM transactions on computational biology and bioinformatics**, v. 10, n. 5, p. 1234–1240, 2013.

GALLO, D. et al. **Entomologia Agrícula**. Pirasicaba, SP, Brasil: FALQ, 2002.

GOLAN, D. E. et al. **Principios de Farmacologia: Bases Fisiopatologicas del Tratamiento Farmacologico**. 3. ed. [s.l.] Lippincott Williams & Wilkins, 2013.

GOLDSMITH, M. R.; SHIMADA, T.; ABE, H. The genetics and genomics of the silkworm, Bombyx mori. **Annual review of entomology**, v. 50, n. 1, p. 71–100, 2005.

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

GUO, N. et al. Structure analysis of the spinneret from Bombyx mori and its influence on silk qualities. **International journal of biological macromolecules**, v. 126, p. 1282–1287, 2019.

HAMAMOTO, H. et al. Silkworm as a model animal to evaluate drug candidate toxicity and metabolism. **Comparative biochemistry and physiology. Toxicology & pharmacology: CBP**, v. 149, n. 3, p. 334–339, 2009.

HUANG, X.; CHEN, X.-G.; ARMBRUSTER, P. A. Comparative performance of transcriptome assembly methods for non-model organisms. **BMC genomics**, v. 17, n. 1, 2016.

HUBER, M. L. S. A. **DESeq2**. [s.l.] Bioconductor, 2017.

INTERNATIONAL SERICULTURAL COMISSION. **Statistics**. Disponível em: <https://inserco.org/en/statistics>. Acesso em: 19 mar. 2024.

INTERNATIONAL SILKWORM GENOME CONSORTIUM. The genome of a lepidopteran model insect, the silkworm Bombyx mori.
Insect biochemistry and molecular biology, v. 38, n. 12, p. 1036–1045, 2008.

KIM, M.-J. et al. Development of detection method for edible silkworm (Bombyx mori) using real-time PCR. **Food control**, v. 94, p. 295–299, 2018.

KRISHNASWAMI, S. et al. **Sericulture manual 2- silkworm rearing**. [s.l.] Food and Agriculture Organization of the United Nations, 1979.

LOPES, T. B. F. et al. Influence of temperature variation on gene expression and cocoon production in Bombyx mori Linnaeus, 1758 (Lepidoptera: Bombycidae). **Comparative biochemistry and physiology. Part D, Genomics & proteomics**, v. 47, n. 101111, p. 101111, 2023.

LOPES, M.; ESCOLA AUGUSTO GIL - AGRUPAMENTO DE ESCOLAS AURÉLIA DE SOUSA. Bombyx mori. **Revista de ciência elementar**, v. 10, n. 1, 2022.

LU, B.; ZENG, Z.; SHI, T. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. **Science China. Life sciences**, v. 56, n. 2, p. 143–155, 2013.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature reviews. Genetics**, v. 12, n. 10, p. 671–682, 2011.

MENDONÇA, S. Rota da Seda, velha(s) e nova(s). **Janus anuário**, p. 124–125, 2016.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature methods**, v. 5, n. 7, p. 621–628, 2008.

MUNHOZ, R. E. F. **Variabilidade genética de raças e híbridos simples de Bombyx mori L. do banco de germoplasma da Universidade Estadual de Maringá**. Maringá – PR., Brasil: Biblioteca Central - UEM, 2010.

MUZELLEC, B. et al. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. **Bioinformatics (Oxford, England)**, v. 39, n. 9, 2023.

NAGARAJU, J.; GOLDSMITH, M. R. Silkworm genomics – progress and prospects. **Current science**, v. 83, n. 4, p. 415–425, 2002.

OLIVEIRA, R. A.; SANTOS, J. A.; BOROVIECZ, S. Análise do custo de produção e do processo produtivo da sericicultura: um estudo de caso no Paraná / Cost analysis of production and the production process of sericulture: a case study in Paraná. **Redes**, v. 22, n. 1, p. 528, 2016.

OSANAI-FUTAHASHI, M. et al. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori. **Insect biochemistry and molecular biology**, v. 38, n. 12, p. 1046–1057, 2008.

PALAZZO, C. L. A CULTURA MATERIAL NA ROTA DA SEDA: FONTES PARA PESQUISA EM HISTÓRIA MEDIEVAL. **Revista Aedos**, v. 2, n. 2, p. 464, 2009.

PANTHEE, S. et al. Advantages of the silkworm as an animal model for developing novel antimicrobial agents. **Frontiers in microbiology**, v. 8, 2017.

PERTEA, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. **Nature biotechnology**, v. 33, n. 3, p. 290–295, 2015.

PIERCE, B. A. **Genética: Um enfoque Conceitual**. Rio de Janeiro, Brasil: Guanabara Koogan, 2017.

RAHEEM, D. et al. Traditional consumption of and rearing edible insects in Africa, Asia and Europe. **Critical reviews in food science and nutrition**, v. 59, n. 14, p. 2169–2188, 2019.

ROSALIND PLATFORM. **ROSALIND**. Disponível em: <https://rosalind.info/search/?q=k+mers>. Acesso em: 25 jan. 2024.

ROY, M. et al. Carbondioxide gating in silk cocoon. **Biointerphases**, v. 7, n. 1, 2012.
TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature biotechnology**, v. 28, n. 5, p. 511–515, 2010.

SNUSTAD, P.; SIMMONS, M. J. **Fundamentos de Genética**. Rio de Janeiro, Brasil: Guanabara Koogan, 2020.

STARK, R.; GRZELAK, M.; HADFIELD, J. RNA sequencing: the teenage years. **Nature reviews. Genetics**, v. 20, n. 11, p. 631–656, 2019.

SULTAN, M. et al. Influence of RNA extraction methods and library selection schemes on RNA-seq data. **BMC genomics**, v. 15, p. 675, 2014.

TANAKA, Y. Genetics of the silkworm, Bombyx mori. Em: **Advances in Genetics**. [s.l.] Elsevier, 1953. p. 239–317.

TANG, F.; LAO, K.; SURANI, M. A. Development and applications of single-cell transcriptome analysis. **Nature methods**, v. 8, n. 4 Suppl, p. S6-11, 2011.

TSUCHIDA, K.; WELLS, M. A. Digestion, absorption, transport and storage of fat during the last larval stadium of Manduca sexta. Changes in the role of lipophorin in the delivery of dietary lipid to the fat body. **Insect biochemistry**, v. 18, n. 3, p. 263–268, 1988.

VIEIRA, M. I. A.; MONTEIRO, V. R. S. **Ciclo de Vida**. Disponível em: <https://www.unioeste.br/portal/bichodaseda/ciclo-de-vida/ciclo-de-vida>. Acesso em: 25 jan. 2024.

VOSHALL, A.; MORIYAMA, E. N. Next-generation transcriptome assembly: Strategies and performance analaysis. In: **Bioinformatics in the Era of Post Genomics and Big Data**. [s.l.] InTech, 2018.WOLF, J. B. W. et al. Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. **Molecular ecology**, v. 19, n. s1, p. 162–175, 2010.

YAMAOKA, R. S.; JÚNIOR, D. S. **Sericultura**. Disponível em: <https://www.idrparana.pr.gov.br/Pagina/Sericicultura>. Acesso em: 25 jan. 2024.

ZHENG, X. et al. Transcriptome analysis of the reproduction of silkworm (Bombyx mori) under dimethoate stress. **Pesticide biochemistry and physiology**, v. 183, n. 105081, p. 105081, 2022.