PÂMELA MARINHO REZENDE

PIPEMIRSEQ: *PIPELINE* INTEGRATIVO PARA ANÁLISES DE EXPRESSÃO DIFERENCIAL EM DADOS DE *MIRNA-SEQ* DE PLANTAS

LAVRAS – MG
2017

PÂMELA MARINHO REZENDE

PIPEMIRSEQ: *PIPELINE* INTEGRATIVO PARA ANÁLISES DE EXPRESSÃO DIFERENCIAL EM DADOS DE *MIRNA-SEQ* DE PLANTAS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Biotecnologia Vegetal, para a obtenção do título de Mestre.

Antonio Chalfun Júnior, PhD.
Orientador
Dr. Matheus de Souza Gomes
Coorientador

LAVRAS – MG
2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Rezende, Pâmela Marinho.
pipeMIRSEQ : *pipeline* integrativo para análises de expressão diferencial em dados de *miRNA-seq* de plantas / Pâmela Marinho Rezende. - 2017.
90 p. : il.

Orientador(a): Antonio Chalfun Júnior.
Coorientador(a): Matheus de Souza Gomes.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2017.
Bibliografia.

1. Bioinformática. 2. MicroRNAs. 3. Expressão Gênica. I. Júnior, Antonio Chalfun. II. Gomes, Matheus de Souza. III. Título.

PÂMELA MARINHO REZENDE

PIPEMIRSEQ: *PIPELINE* INTEGRATIVO PARA ANÁLISES DE EXPRESSÃO DIFERENCIAL EM DADOS DE *MIRNA-SEQ* DE PLANTAS

PIPEMIRSEQ: AN INTEGRATIVE PIPELINE FOR DIFFERENTIAL EXPRESSION ANALYSES IN PLANT MIRNA-SEQ DATA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Biotecnologia Vegetal, para a obtenção do título de Mestre.

APROVADA em 24 de fevereiro de 2017.
Antonio Chalfun Júnior, PhD. UFLA
Dr. Matheus de Souza Gomes UFU
Dr. Laurence Rodrigues do Amaral UFU

Antonio Chalfun Júnior, PhD.
Orientador
Dr. Matheus de Souza Gomes
Coorientador

LAVRAS – MG
2017

Aos meus pais Regina e Sérgio que de forma incondicional me apoiam em toda minha vida, meus exemplos de amor e determinação.
Dedico

AGRADECIMENTOS

Agradeço a Deus, pelo dom da minha vida.

À Regina, minha mãe, pelo amor e apoio, ao meu pai Sérgio pelo exemplo de pessoa, a minha querida irmã Marcela, pelas palavras amigas e as minhas sobrinhas que mesmo não sabendo o que a Dindinha trabalha, sentem muito orgulho.

Ao meu namorado e amigo, Danilo, pelo amor, paciência, espera, renúncia e por sempre acreditar em mim.

A minha família, tias, primos, e avos, em especial Antônio Marinho pelo amor incondicional.

À Universidade Federal de Lavras, meu segundo lar, e ao programa de Biotecnologia Vegetal pela oportunidade.

Ao Antonio Chalfun Júnior, pela orientação e confiança em mim depositada.

Ao Matheus de Souza Gomes, pela coorientação e apoio aos meus trabalhos no mestrado.

Aos membros e ex-membros do Laboratório de Fisiologia Molecular de Plantas, pelo apoio, ensinamentos, pela formação de sinceras amizades e pelos momentos de alegrias.

A FAPEMIG, pela concessão da bolsa de mestrado. Ao INCT Café pelo apoio em inúmeros projetos.

Aos membros da banca de avaliação que dispuseram de seu tempo para poder acrescentar mais conhecimentos ao meu trabalho.

E a todos que de alguma forma contribuíram direto e indiretamente para a conclusão do meu mestrado.

Muito obrigada!

“Aqueles que são loucos o suficiente para acreditar que podem mudar o mundo são os que realmente mudam.” (Steve Jobs)

RESUMO

Técnicas computacionais para solucionar problemas biológicos vêm sendo altamente empregadas nos últimos anos. Atualmente, vários estudos de RNAs não-codantes utilizando a bioinformática estão sendo realizados. Os *miRNAs* são uma das grandes classes de *non-coding RNA* (*ncRNA*), e a utilização da bioinformática é um ponto crucial para a solução rápida, econômica e confiável para análises dos mesmos, como, predição de precursores, de *miRNAs* maduros, de alvos e análise de expressão. Para análise de expressão, a utilização de ferramentas, plataformas e pipelines computacionais é essencial para analisar os dados de sequenciamento em larga escala de *miRNAs*. Porém, as inciativas de análise de *miRNA-seq* de origem vegetal são poucas. O objetivo do presente trabalho foi desenvolver um *pipeline* (pipeMIRSEQ) que auxilie de forma eficiente e confiável todas as etapas da análise de dados provenientes do *miRNA-seq*. A implementação do mesmo foi separada em módulos: análise de qualidade, utilizando quatro tipos de ferramentas; mapeamento dos reads, comparando três tipos de diferentes alinhadores; quantificação, levando em consideração as famílias de *miRNAs* homólogos; diferença de expressão, utilizando dois tipos de pacotes para cálculos, e pôr fim a etapa de *review*, com o objetivo de sintetizar todas as informações retiradas em cada etapa. Para a validação do pipeline foi utilizado uma base de dados reduzida de *miRNA-seq* de café (*Coffea arábica* L.). Foi possível observar que o pipeMIRSEQ conseguiu analisar toda a base de dados, apresentando resultados condizentes com a literatura, e alcançando assim seus objetivos. Por fim, foram abordados temas de melhorias do pipeline tanto em questões computacionais quanto em questões biológicas.

Palavras-chave: Bioinformática. MicroRNAs. Expressão Gênica.

ABSTRACT

Computational techniques have been highly used in the recent years to mitigate biological issues. Currently, many studies of non-coding RNAs (ncRNAs) using bioinformatics have been presented. The miRNAs are one of the ncRNAs classes and the use of bioinformatics is crucial for a rapid, cost-effective and reliable solution for analysis of these molecules, for instance, prediction of the precursors, mature, and target sequences; and expression analysis. For expression analysis, the use of tools, platforms, and pipelines has been essential to analyze large-scale miRNA-seq data. However, there are few aiming at the analysis of plant miRNA-seq. The aim of this work was to develop a pipeline (pipeMIRSEQ) to assist in an efficient and reliable way to all the data analysis steps comprised within miRNA-seq. The implementation was separated in modules: quality analysis, using four kinds of tools; mapping of the reads, comparing three different aligners; quantification, considering the homolog miRNA families; differential expression, using two packages for calculation; and, at last, a review step, aiming at synthesizing all the information from each step. To validate the pipeline a reduced miRNA-seq dataset from coffee (*Coffea arabica L.)* was used. It was possible to observe that the pipeMIRSEQ could analyze all the dataset, presenting results comparable to the literature, and, therefore, achieving its aims. At last, improvements for the pipeline regarding both computational and biological aspects are discussed.

Keywords: Bioinformatics. microRNAs. Gene expression

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1
Introdução geral

RESUMO

Os *microRNAs* são pequenos RNAs, com tamanho de aproximadamente 19 a 24 nucleotídeos que apresentam um importante papel nos processos de regulação da expressão gênica em diversos organismos. Muitas técnicas de biologia molecular tradicionais já implementadas para *mRNAs* vêm sido utilizadas para análise de moléculas de *miRNAs*, das quais o sequenciamento em larga escala apresenta maior usabilidade, tendo em vista sua capacidade de quantificação de transcritos em larga escala. Esta técnica apresenta um alto desempenho e combina técnicas de descoberta e quantificação de transcritos em um único ensaio, diferindo assim de outras técnicas de análise de expressão. Entretanto, umas das etapas cruciais para se obter o resultado final do sequenciamento em larga escala de *miRNAs* (*miRNA-seq*) é a etapa de análise de dados, que consiste na utilização de ferramentas computacionais. A análise dos dados que tem como objetivo calcular a diferença de expressão dos *miRNAs* consiste em: análise de qualidade, mapeamento dos *reads*, quantificação dos *miRNAs* e pôr fim a determinação das diferenças de expressão. Muitos softwares são desenvolvidos com o intuito de facilitar as análises de dados do *miRNA-seq*, entretanto, poucos são aqueles que apresentam a capacidade de calcular a expressão diferencial dos genes.

Palavras-chave: RNA não-codificantes. *MiRNA-seq*. Diferença de expressão.

ABSTRACT

The *miRNAs* are small RNAs of about 19-24 nt that play an important role in the regulation of gene expression in several organisms. Many standard Molecular Biology techniques already used for mRNA analyses have been used for miRNAs analysis, from which the high throughput sequencing currently presents greater usability, considering the possibility of quantification of the transcripts in large-scale. It is a high throughput method that combines discovery and quantification of transcripts in a single essay, differing from the other techniques available for expression analyses. However, one of the crucial steps to get the results from miRNA-seq is the data analysis step, which comprises the use of computational tools. The data analysis for calculation of *miRNAs* differential expression consists in: quality analysis, reads mapping, quantification of the *miRNAs* and differential expression calculation. Many softwares have been developed for miRNA-seq data analysis, but a few present the feature of calculation of differential expression.

Keywords: non-coding RNA. miRNA-seq. Differential expression

1. INTRODUÇÃO

Moléculas de RNAs não codificantes de proteínas (*ncRNAs*) possuem importantes papéis no desenvolvimento de diferentes organismos como animais, plantas e vírus, devido à suas atuações na regulação da expressão gênica (DEBAT & DUCASSE, 2014; WANG & CHEKANOVA, 2016). Os microRNAs (*miRNAs*) pertencem a uma das classes de *ncRNAs* e tem o papel de silenciar a expressão de genes, sendo muito estudados em diferentes organismos (MENG et al., 2012; DE SOUSA CARDOSO et al., 2015; LAMONTAGNE et al., 2015).

Várias técnicas de análise de perfil de expressão de *miRNAs* já foram estabelecidas, como: *northern-blot*, PCR em Tempo Real (*RT-qPCR),* microarranjos e sequenciamento de RNA (*RNA-seq*) (MOREIRA, 2015). O *RNA-seq* é um dos métodos baseados em sequenciamento mais utilizados atualmente (CONESA et al., 2016), e se destaca por ser um método de sequenciamento de alto desempenho (PUNDHIR et al., 2015). Este método se distingue dos outros métodos de quantificação pois não necessita de uma predefinição de transcritos que se deseja identificar (AUER & DOERGE, 2010; GIT, A. et al., 2010), além de apresentar uma habilidade simultânea de detecção e quantificação da expressão gênica em larga escala (JOHNSON et al., 2016).

Uma vez que as classes de *miRNAs* envolvidas com regulação da expressão genica apresentou ser uma das classes mais promissoras de RNA a serem estudas, a tecnologia de sequenciamento de RNA em larga escala, passou a ser utilizada para análise de *miRNAs* (TAM et al., 2015). Porém, uma das etapas importantes para experimentos de *miRNA-seq* é a análise de dados *in silico* provenientes da técnica, a qual necessita de conhecimento computacional para solucionar a mesma (SUN et al., 2014; CONESA et al., 2016). Muitos softwares, pacotes e *pipelines* tem sido desenvolvidos a fim de facilitar e aprimorar as análises de dados provenientes do sequenciamento de *miRNAs* (TAM et al., 2015).

2. REFERENCIAL TEÓRICO

2.1. RNAs não-codantes

Grande parte do transcritoma não codifica proteínas, sendo essa fração de RNAs, denominados de RNAs não-codantes (*ncRNA*) (ZHAO et al., 2016). Moléculas de *ncRNAs* destacam-se por ser elementos chaves de vários processos celulares críticos (CONSORTIUM, 2017). Muitas classes recém descobertas de *ncRNA* desempenham papéis indispensáveis na expressão gênica em vários estádios de desenvolvimento dos organismos e em diferentes condições (WANG & CHEKANOVA, 2016).

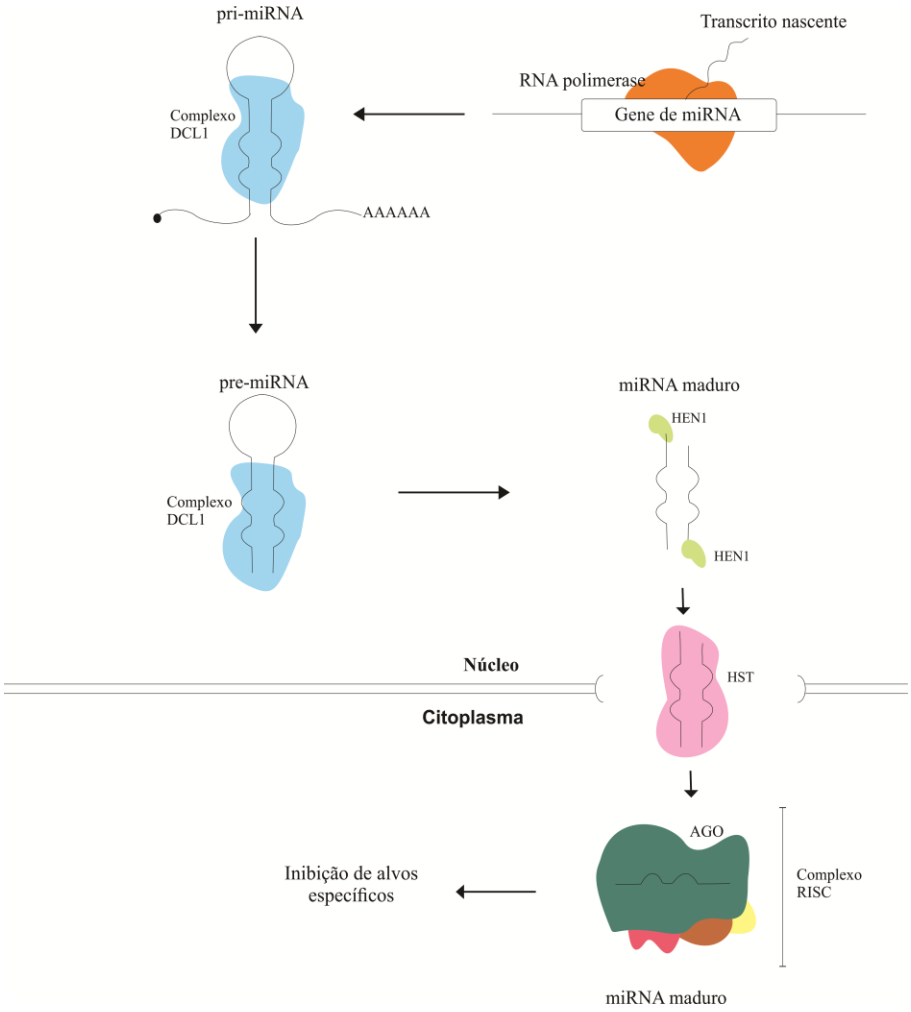
Dentre os vários tipos de *ncRNAs*, os microRNAs estão entre os mais estudados (AXTELL et al., 2011), sendo caracterizados por apresentarem cadeias nucleotídicas de aproximadamente 19 a 24 bases e sua principal função é interferir em diferentes etapas do processamento de *mRNAs* (RNAs mensageiros) (CARTHEW & SONTHEIMER, 2009; ZHAO et al., 2016). Em organismos eucariotos há três grandes classes de *ncRNAs*: *siRNAs* (Pequenos RNAs interferentes), *piRNAs* (RNA de interação piwi) e *miRNAs* (microRNAs). Os *miRNAs* promovem o silenciamento de genes, modulando a expressão através da complementariedade entre o *miRNAs* e o alvo, resultando assim na regulação pós-transcricional dos genes (WANG & CHEKANOVA, 2016).

Os *miRNAs* estão envolvidos em vários processos do crescimento e desenvolvimento de plantas e animais, como por exemplo, o *miR156* que controla o desenvolvimento dos órgãos reprodutivos feminino em Arabidopsis (*Arabidopsis Thaliana* L.*)* (DEBAT & DUCASSE, 2014). Outros *miRNAs* foram identificados em participar de estágios do desenvolvimento de células cancerígenas (SONG et al., 2012). Já o *miR146a*, está associado a inflamação em estudos com a epilepsia em humanos (ARONICA et al., 2010) e *miRNAs* que

regulam o processo de indução da embriogênese somática em *Arabidopsis* (SZYRAJEW et al., 2017).

Os *miRNAs* de plantas e animais são originários da enzima polimerase II. Em plantas, são produzidos a partir de precursores com estrutura secundária denominados de *miRNAs* primários longos (*pri-miRNAs*). Ainda no núcleo, o *pri-miRNA* é clivado por um complexo protéico, contendo as proteínas DCL1 (*Dicer*), DRB1 (*DsRNA BINDING PROTEIN1*), e SE (*SERRATE*), dentre outras, levando à formação de um *microRNA* precursor (*pre-miRNA*) que possui fita dupla e uma alça denominada hairpin (Figura 2.1). Posteriormente, este *pre-miRNA* é clivado pelo complexo da proteína DCL1, ainda no núcleo, e é metilado pela proteína HEN1 (*HUA- EHANCER 1*). O *miRNA* é então encaminhado ao citoplasma com ajuda da proteína HST (*HASTY*) e incorporado a outras proteínas formando um complexo denominado RISC (*RNA-Induced Silencing Complex*), do qual, também faz parte a proteína ARGONAUTA (Figura 2.1). O complexo RISC é responsável por escolher um lado da dupla fita de microRNAs e incorporá-la ao *miRNA*-alvo (CARTHEW & SONTHEIMER, 2009; DEBAT & DUCASSE, 2014).

Figura 2.1 – Sintetização da biogênese de microRNAs em plantas.



Legenda: Destacando as principais proteínas participantes do processo o pri-miRNA, pre-miRNA e miRNA maduro.

Análises moleculares são importantes para compreender o papel dos *miRNAs* na regulação de processos biológicos. Estas análises tradicionais, utilizadas para determinar a expressão gênica de *mRNAs* também foram adaptados para a utilização em *miRNAs* (GIT, ANNA et al., 2010). A expressão gênica é o

mecanismo no qual um fluxo de instruções provindas do código genético são usadas para o processamento de um produto gênico, geralmente resultando em uma proteína (GRIFFITH et al., 2015). A expressão gênica pode ser regulada em vários níveis e por vários fatores, logo, e através da sua análise é possível identificar quais, como e quando diferentes genes são regulados em diferentes condições fisiológicas (MOREIRA, 2015).

2.2. Técnicas de quantificação da expressão gênica

Inicialmente haviam poucas técnicas para a detecção da expressão gênica e detecção da expressão diferencial entre transcritos, como *Northern Blot* e Reação em Cadeia da Polimerase (PCR) (MOREIRA, 2015). Entretanto, nas duas últimas décadas, várias técnicas de determinação da variação da expressão gênica foram desenvolvidas (KUKURBA & MONTGOMERY, 2015; MOREIRA, 2015), ás quais podem ser classificadas em três categorias: métodos baseados em quantificação, tal como PCR em tempo real (*RT-qPCR*); métodos baseados em hibridação, tais como os microarranjos, e métodos baseados em sequenciamento, tal como sequenciamento em larga escala de *RNAs (RNA-seq*) (GIT, ANNA et al., 2010; KUKURBA & MONTGOMERY, 2015).

A técnica da *PCR* corresponde à amplificação de um fragmento de DNA pela enzima DNA Polimerase (PIERCE, 2013). A base desta técnica é a replicação de DNA em vários ciclos, todavia ela foi adaptada para moléculas de RNA para demonstrar o nível de expressão de um gene (MOREIRA, 2015). Entretanto, a técnica da *PCR* convencional é considerada demorada e de alto custo e de baixa sensibilidade (GIT, ANNA et al., 2010), surgindo assim uma evolução para a técnica de PCR, a qual é capaz fazer amplificações simultâneas e quantificar os transcritos ao mesmo tempo, denominada de PCR em tempo real (*qPCR* ou *RT-qPCR*) (HIGUCHI et al., 1992; PABINGER et al., 2014). Com a comercialização da técnica de *qPCR* em 1996 e o avanço da mesma, o número de publicações que

utilizam essa prática cresceu e continua a crescer exponencialmente (PABINGER et al., 2014). Uma desvantagem da utilização da técnica de *RT-qPCR* é o pequeno número de genes que podem ser analisados em cada ensaio da técnica comparado com a técnica de microarranjo e *miRNA-seq* (LIU et al., 2014), além do fato das análises de *RT-qPCR* de microRNAs necessitarem do aumento do tamanho das sequências de nucleotídeos, para que seja possível a detecção da amplificação dos *sRNAs* (*small RNAs*) (GIT, ANNA et al., 2010).

As técnicas de microarranjo utilizam pequenas sequências de DNA complementar (*cDNA*) chamadas de sondas, em superfície sólidas que são expostas a sequências de *cDNAs* marcados com corantes fluorescentes, os *quais* são complementares às sondas, se hibridizando às mesmas e, posteriormente, sendo detectados à partir da luminosidade produzida pela excitação por laser (GUL et al., 2016; SARWAT & YAMDAGNI, 2016).

Análises de microarranjo são utilizadas para estudos de transcritos em larga escala, podendo conter de centésimos a milhões de sondas, de aproximadamente 100 pares de bases (pb) (AGILENT, 2017), que permitem a detecção da expressão de vários genes em paralelo (SARWAT & YAMDAGNI, 2016), entretanto este método apresenta algumas limitações: utilização de *chips* específicos para cada análises, necessidade de conhecimento a priori das sequências que serão utilizadas para a análise, problema de hibridização-cruzada de sequências altamente semelhantes, ou seja, sondas que são complementares a mais de um tipo de gene e, pôr fim, a limitação de quantificar genes com expressão muito baixa (CASNEUF et al., 2007; SHENDURE, 2008). O tamanho das moléculas de *miRNAs* não apresentam desafios pela técnica de microarranjo, uma vez que utilizam sondas *LNAs* (*locked nucleic acid*), que também aumentam a especificidades de hibridização (CASTOLDI et al., 2006).

Em contraste aos métodos baseados em hibridização, os métodos baseados em sequenciamento não necessitam do conhecimento prévio da sequência dos

transcritos que se deseja identificar/quantificar (PAREEK et al., 2011). Um dos métodos baseados em sequenciamento altamente utilizado atualmente é o sequenciamento de RNA (CONESA et al., 2016), que se destaca por ser um uma técnica de sequenciamento de alto desempenho (PUNDHIR et al., 2015), combinando identificação e quantificação de transcritos em um único ensaio (CONESA et al., 2016). Diferente dos outros métodos baseados em quantificação, o sequenciamento de RNA apresenta uma maior sensibilidade de detecção, possibilitando estudos com novos transcritos, tal como estudos de quantificação da expressão de *microRNAs* em larga escala, detecção de novos *miRNAs* e alta distinção de *microRNAs* que pertencem à mesma família (TAM et al., 2015).

A utilização de um determinado método de quantificação depende do objetivo do estudo e do organismo de interesse (LIU et al., 2014). O uso de métodos baseados em quantificação, como *RT-qPCR*, é comumente utilizado para adicionar confiabilidade a estudos que utilizam Microarranjos e *miRNA-seq* (CONSORTIUM, 2014; RAJKUMAR et al., 2015; GUL et al., 2016), auxiliando na validação dos resultados obtidos por estas técnicas de quantificação da expressão. Por exemplo, vários estudos relatam que 85% de novos eventos de *splicing* e 88% de transcritos diferencialmente expressos, descobertos pela técnica de sequenciamento em larga escala são validados pela técnica de *RT-qPCR* (GRIFFITH et al., 2010).

2.3. Sequenciamento de larga escala de RNA

O sequenciamento genômico tem como objetivo identificar as sequências de nucleotídeos de DNA e RNA (MOREIRA, 2015). A metodologia de Sanger, foi uma das primeiras técnicas de sequenciamento (SANGER, 1975), a qual se baseia na utilização de nucleotídeos terminadores de cadeia (didesoxinucleotídeos) marcados radiotivamente e na técnica da PCR para a síntese da fita complementar à sequência sendo sequenciada (PAREEK et al., 2011).

Após a metodologia de Sanger ser automatizada, várias estratégias de sequenciamento foram surgindo, como o sequenciamento de RNA (MOREIRA, 2015). Ao invés de utilizarem o RNA mensageiro (*mRNA*) como material inicial para amplificação, é utilizado o DNA complementar (*cDNA*), devido à instabilidade da molécula de *mRNA* (AUER & DOERGE, 2010). Esta técnica é utilizada em vários estudos como: verificação do nível de transcritos em uma determinada situação, identificação de polimorfismos (*SNPs*) e descoberta de mutações (MOREIRA, 2015). Porém, com a aplicação de técnicas da computação na biologia molecular, esta abordagem de sequenciamento de RNA tornou-se atrasada em razão do seu baixo desempenho e baixo volume de dados (SANGER, 1975; KUKURBA & MONTGOMERY, 2015). Foram então desenvolvidas novas plataformas de sequenciamento, conhecidas como *NGS* (*Next-Generation Sequencing*) (MUIR et al., 2016).

As plataformas NGS apresentam tecnologias diferentes entre si, entretanto se assemelham por consistir em tecnologias de alto desempenho e aplicarem o sequenciamento em paralelo (KUKURBA & MONTGOMERY, 2015; MOREIRA, 2015), processando até bilhões de fragmentos ao mesmo tempo, enquanto o sequenciamento por Sanger processa poucos fragmentos (SANGER, 1975; PAREEK et al., 2011). O *RNA-seq* é um dos métodos que utiliza estas plataformas, capazes de determinar a sequência de nucleotídeos de diferentes moléculas de RNA, tais como: pequenos e longos RNAs, *mRNAs* ou transcritomas completos, conhecido também como RNA total (PUNDHIR et al., 2015).

Algumas etapas genéricas são definidas para se desenvolver um experimento de *RNA-seq*. Estas etapas consistem na preparação do RNA para o sequenciamento, preparo das bibliotecas do sequenciamento, sequenciamento, e por fim análise do produto final do sequenciamento (PUNDHIR et al., 2015; CONESA et al., 2016). A primeira etapa consiste em preparar o RNA para o sequenciamento, isto é, isolando o RNA a partir de uma amostra biológica e

convertendo-o em cDNA, a qualidade deste RNA tem impactos significativos no resultado final (GRIFFITH et al., 2015). É importante utilizar um protocolo de extração que possa evitar degradação da molécula de RNA por ser considerada uma molécula instável (KUKURBA & MONTGOMERY, 2015), assim como possa retirar a quantidade de RNA ribossomal (*rRNA*) que corresponde a 90% de todo o RNA encontrado na célula, caso o *rRNA* não seja o interesse do experimento (CONESA et al., 2016). Para o sequenciamento de pequenos RNAs, os fragmentos correspondentes são isolados do RNA total a partir de um gel desnaturante (TAM et al., 2015).

A escolha da plataforma de sequenciamento a ser utilizada em um estudo está diretamente relacionada aos objetivos do mesmo, sendo realizada antes do prepare das bibliotecas. As plataformas de sequenciamento podem ser definidas em plataformas baseadas em conjunto (sequenciamento de várias cópias idênticas de uma mesma molécula) ou baseada no sequenciamento de apenas uma molécula (KUKURBA & MONTGOMERY, 2015). A escolha da plataforma de sequenciamento altera as etapas de análise e interpretação dos dados, como por exemplo, para o fechamento de genomas utiliza-se plataformas baseadas no sequenciamento de apenas uma molécula, como a tecnologia PacBio (RHOADS & AU, 2015).

A etapa de preparação das bibliotecas para o sequenciamento envolve a fragmentação e/ou amplificação das moléculas de RNA, esta etapa pode variar de acordo com o tipo de RNA e a plataforma de NGS utilizada para o sequenciamento (KUKURBA & MONTGOMERY, 2015).

2.3.1. Análise de dados de sequenciamento de RNA

A última etapa de um experimento de sequenciamento de RNA são as análises dos produtos final do sequenciamento (CONESA et al., 2016), a qual é realizada inteiramente por meio de ferramentas *in silico* (SCHORDERET, 2016). As

metodologias e ferramentas utilizadas por esta etapa variam de acordo com o objetivo final do projeto, tais como: nível de expressão de transcritos, descobertas de novas estruturas de genes, isoformas de *splicing* alternativo, dentre outras aplicações (KUKURBA & MONTGOMERY, 2015). A análise dos dados brutos de saída do sequenciamento de RNA é feita pelo controle de qualidade, seguido do alinhamento dos *reads*, montagem dos transcritos, quantificação dos transcritos e análise da expressão diferencial (GRIFFITH et al., 2015; CONESA et al., 2016).

O controle de qualidade dos produtos do sequenciamento assegura que erros de sequenciamento, artefatos da PCR ou contaminações não sejam passados adiante nas análises, evitando assim a ocorrência de falsos positivos (CONESA et al., 2016). Utilizando ferramentas como FastQC (ANDREWS, 2010) é possível visualizar a qualidade das bases dos *reads*, assim como a ferramenta NGSQC (DAI et al., 2010) que também é aplicada para análise de qualidade (CONESA et al., 2016; JOHNSON et al., 2016). Como a qualidade das sequências vão diminuindo na exterminada 3' dos *reads*, ferramentas como FASTX-Toolkit (HANNON, 2009), Trimmomatic (BOLGER et al., 2014) e Prinseq Lite (SCHMIEDER & EDWARDS, 2011) aparam os *reads* retirando os adaptadores (CONESA et al., 2016; JOHNSON et al., 2016).

O alinhamento de dados de *miRNA-seq* é considerado um grande desafio para a computação, pois os *reads* são curtos (~50 bases) devido ao pequeno tamanho tais *reads* tem maior probabilidade de alinhar em vários *loci* (KUKURBA & MONTGOMERY, 2015). Porém, os *reads* de *miRNAs* não são mapeados em locais de junção, pois os microRNAs maduros não são fragmentados pelo sequenciamento, o que diminui muitos dos problemas na etapa de mapeamento (GARBER et al., 2011; KUKURBA & MONTGOMERY, 2015). A porcentagem de *reads* mapeados e o acúmulo de *reads* na extremidade 3' também são parâmetros importantes do alinhamento, pois é um indicativo de sequências

contaminantes e baixa qualidade do RNA extraído, respectivamente (CONESA et al., 2016).

Alinhadores convencionais como Bowtie 1 (LANGMEAD et al., 2009), Bowtie 2 (LANGMEAD & SALZBERG, 2012) e BWA (LI & DURBIN, 2009), utilizam algoritmos conhecidos como *unspliced read aligners*, que são algoritmos que alinham sem permitir grandes lacunas entre os *reads* (GARBER et al., 2011), logo são recomendados para mapeamentos de dados derivados de *miRNA-seq*, pois não são necessários o tratamento de junções entre regiões exons-exons (KUKURBA & MONTGOMERY, 2015; TAM et al., 2015). Para pequenos RNAs alguns alinhadores como PatMaN (PRÜFER et al., 2008) e MicroRazerS (EMDE et al., 2010) foram projetados para mapear sequências curtas (CONESA et al., 2016). Por outro lado devido à grande utilização de dados de *RNA-seq*, outros alinhadores foram desenvolvidos para alinhar transcritomas, conhecidos como *Spliced aligners* (GARBER et al., 2011; KUKURBA & MONTGOMERY, 2015). Este procedimento pode reconhecer *reads* com grandes lacunas exon-íntron e com lacunas menores (KUKURBA & MONTGOMERY, 2015), algumas ferramentas que utilizam este modelo de algoritmo são TopHat (TRAPNELL et al., 2012) e STAR (DOBIN et al., 2013).

Com os *reads* alinhados, por meio do processo de montagem é possível a identificação de transcritos e outras regiões do genoma (KUKURBA & MONTGOMERY, 2015; CONESA et al., 2016), entretanto esta etapa não é necessária para análise de *miRNA-seq*, pois os microRNAs já possuem o seu tamanho total. Este processo de identificação dos transcritos coletivamente é conhecido como reconstrução de transcritoma (GARBER et al., 2011). A reconstrução de um transcritoma é considerado um processo de alto custo computacional devido à alguns fatores, tais como, a diferença de magnitude de expressão entre genes, alguns genes contêm baixa quantidade de *reads* associados

a ele; além da difícil identificação de RNA maduros, pois estes não possuem íntrons. (GARBER et al., 2011; CONESA et al., 2016).

A montagem dos transcritos pode ser feita a partir de duas abordagens, por meio do alinhamento com um genoma de referência (TRAPNELL et al., 2010) ou pela montagem *de novo* (KUKURBA & MONTGOMERY, 2015). Quando não se tem um genoma de referência é preciso montar o genoma *de novo* pela primeira vez, processo este que pode ter a sua qualidade influenciada por vários fatores, como a natureza do transcritoma (complexidade do mesmo, grau de polimorfismo, *splicing* alternativo, magnitude da expressão gênica) e fatores inerentes às tecnologias NGS (erros de sequenciamento, *gaps* no sequenciamento) (KUKURBA & MONTGOMERY, 2015; CONESA et al., 2016). Entre os montadores *de novo* mais conhecidos pode-se citar o Trinity (GRABHERR et al., 2011) que monta dados da plataforma Illumina, e o MIRA (CHEVREUX et al., 2004) que monta genomas bacterianos provenientes da plataforma IonTorrent PGM.

Uma análise comum a ser feita com dados de sequenciamento em larga escala é estimar o nível de expressão de genes em diferentes condições, conhecida como quantificação de transcritos (KUKURBA & MONTGOMERY, 2015; CONESA et al., 2016). A princípio, umas das tecnologias mais utilizadas para análise de expressão em larga escala eram os microarranjos, os quais a partir de 2007 se tornaram menos utilizados com o avanço das tecnologias NGS (SHENDURE, 2008; KUKURBA & MONTGOMERY, 2015). Embora estas tecnologias tenham o mesmo objetivo, há diferenças estatísticas entre os cálculos de expressão. Microarranjos utilizam a intensidade das sondas que podem se aproximar de uma distribuição normal, enquanto que para dados de sequenciamento de alta performance, a contagem de *reads* não pode ser aplicado a uma distribuição normal (KUKURBA & MONTGOMERY, 2015). As medidas mais utilizadas em quantificação de transcritos são: *count-based*, *Reads* por Kilobase por Milhões de *Reads* Mapeados

(*RPKM*), Fragmentos por Kilobase de Transcritos por Milhões de Fragmentos Mapeados (*FPKM*) e Transcritos por Milhão (*TPM*).

A medida *count-based* é baseada na contagem de *reads*, entretanto, vários estudos relatam que essa distribuição não leva em consideração fatores de variabilidade biológica, tal como comprimento dos transcrito e o total do número de *reads* (ROBINSON & SMYTH, 2007; LANGMEAD et al., 2010). Esta contagem de *reads* deve ser normalizada para corrigir variabilidades, sendo que o *RPKM* é uma medida que pode ser utilizada para manter uniformização dos dados, correspondendo ao número de *reads* mapeados em uma determinada região pelo tamanho da região em mil pares de bases multiplicado pelo número total de milhões de *reads* mapeados (GARBER et al., 2011). Entretanto, quando o sequenciamento é feito utilizando a metodologia *paired-end* (fragmentos são sequenciados nas duas direções de suas extremidades (OZSOLAK & MILOS, 2011) deve-se utilizar a medida Fragmentos por Kilobase de Transcritos por Milhões de Fragmentos Mapeados (*FPKM*), a qual difere da utilização da *RPKM* pois trabalha com pares de *reads* de um mesmo fragmento (Equação *(2. 1)*) (TRAPNELL et al., 2010; GARBER et al., 2011; PIMENTEL, 2014).

$$FPKM_i = \frac{X_i}{\left(\frac{\widetilde{l_\iota}}{10^3}\right)\left(\frac{\widetilde{N}}{10^6}\right)} = \frac{X_i}{\widetilde{l_\iota}\,N} \cdot 10^9 \qquad (2.\ 1)$$

Espera-se que a média da expressão relativa de todos os genes entre amostras diferentes sejam iguais, entretanto utilizando a medida FPKM/RPKM isto não acontece, pois é levado em consideração o tamanho dos fragmentos/*reads*, e sabe-se que este tamanho é variável dentro de uma mesma amostra, logo foi criada a medida *TPM* que normaliza a medida *FPKM* pela soma total da medida *FPKM* de todos os fragmentos de uma mesma amostra (Equação *(2. 2)*) (LI et al., 2010;

PIMENTEL, 2014). Ou seja, as medidas *FPKM* e *RPKM* tendem a ser ineficazes em amostras em que o tamanho de fragmentos sejam heterogêneos (CONESA et al., 2016).

$$TPM_i = \left(\frac{FPKM_i}{\sum_i FPKM_i}\right) . 10^6 \qquad (2.\ 2)$$

Algumas ferramentas usadas para quantificar transcritos baseada na medida *count-beased*: HTSeq-count não leva em consideração variabilidades biológicas (KUKURBA & MONTGOMERY, 2015) e a ferramenta Samtools (LI et al., 2009). O programa Kallisto (BRAY et al., 2016) quantifica abundância de transcritos usando pseudo-alinhamento, que não alinha base por base de uma sequência e sim *k-mers* (tamanho determinado de uma palavra), possibilitando um pseudo-alinhamento altamente rápido comparado com outros quantificadores (PACHTER, 2015).

Através da etapa de quantificação é possível fazer o cálculo das expressões diferenciais entre transcritos. A etapa de cálculo da expressão diferencial de genes tem como objetivo identificar os transcritos que são significativamente diferencialmente expressos por meio de testes estatísticos em larga escala (KHANG & LAU, 2015). Uma das questões cruciais para a diferença de expressão é considerar a variabilidade biológica das amostras, que corresponde à resposta individual aos diferentes estímulos a que um organismo é submetido resultando em uma variação entre os indivíduos (KROLL, 2005).

Dados provenientes de variabilidade biológica segue a distribuição gaussiana, onde no centro é definido a média das expressões dos transcritos, e os transcritos são dispostos de acordo com sua variação em relação à média (SEO et al., 2016). A primeira etapa da diferença de expressão é calcular a medida *Fold-change*, para avaliar a mudança entre o nível de expressão entre o tratamento e

controle (LOVE et al., 2014). Após, para verificar se um transcrito é considerado diferencialmente expresso, usa-se o índice Z (Equação (*2.3*) que retribui a um valor-p, definindo assim como a probabilidade do transcrito x_i rejeitar a hipótese nula, ou seja, ser diferencialmente expresso (DE) (ANDERS & HUBER, 2010; CONESA et al., 2016). Onde na equação *(2.3*, x_i é o valor da expressão (*fold-change*), μ é a média da expressão e σ é o valor da variância em relação à média (ZHANG et al., 2015).

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{2.3}$$

False discovery rate (FDR) é uma medida definida para tratar a taxa de erros cometidos pela falsa rejeição da hipótese nula (BENJAMINI & HOCHBERG, 1995), ou seja, é uma medida utilizada em diferença de expressão para tratar falsos positivos. Vários pacotes e ferramentas foram desenvolvidos para calcular transcritos DE em dados provenientes de sequenciamento em larga escala: DESeq (WANG et al., 2010), edgeR (ROBINSON et al., 2010) e DEseq2 (LOVE et al., 2014) utilizam a distribuição binomial negativa, que trata repetições biológicas e replicatas técnicas; GFOLD (FENG et al., 2012) e PoissonSeq (LI et al., 2012) empregam a distribuição de Poisson onde trata apenas de replicatas técnicas.

A distribuição de Poisson consiste na probabilidade de um número de eventos acontecer em um intervalo de tempo, ela é utilizada para eventos raros, entretanto, ao se tratar de dados provenientes de *RNA-seq* ela não é muito utilizada, pois não leva em consideração a variação entre a média e a variância de uma população, tratando assim apenas de replicatas técnicas (FURTADO, 2005; KHANG & LAU, 2015). Em contrapartida, é recomendado a utilização da distribuição binomial negativa para experimentos com repetições biológicas, pois não considera a média

igual a variância e calcula o número de tentativas necessárias para se obter uma quantidade definida de sucessos (RAJKUMAR et al., 2015).

Além das análises de expressão, outras análises também podem ser feitas, tais como: *splicing* alternativo, que detecta alterações na transição de isoformas do mesmo gene, e a ferramenta Cuffdiff2 (TRAPNELL et al., 2012) que estima primeiro a expressão da isoforma e depois compara suas diferenças (CONESA et al., 2016).

2.3.2. Integração de ferramentas computacionais para análise de *miRNA-seq*

A popularidade da utilização das técnicas de NGS vem crescendo desde de 2007 (SHENDURE, 2008; MUIR et al., 2016), devido a custos menores de sequenciamento (METZKER, 2010; NIELSEN et al., 2011) e à possibilidade da multiplexação de várias bibliotecas em paralelo, obtendo assim um grande volume de dados (AUER & DOERGE, 2010). Em muitos experimentos, o volume de dados varia na grandeza de gigabytes (HODKINSON & GRICE, 2015), como em banco de dados colaborativos de dados provenientes de NGS como o *Sequence Read Archive* (SRA) (NAKAMURA et al., 2013) os quais, em 2016 apresentou aproximadamente 6 petabases (6×10^{15}) de dados totais (NCBI, 2016). O decréscimo do custo de sequenciamento e o número de dados gerados pelo mesmo está colocando maior esforço computacional e conhecimento para lidar com a análise destes dados (MUIR et al., 2016). Devido a estes fatores o desenvolvimento de plataformas, *pipelines* e *softwares* que analisam dados de NGS vem crescendo (NEVADO & PEREZ-ENCISO, 2015).

A integração de ferramentas computacionais para análises de bioinformática são comumente conhecidas como *pipelines* (NEVADO & PEREZ-ENCISO, 2015). Para análise de sequenciamento de microRNAs várias ferramentas estão disponíveis, entretanto a maioria delas apresentam o objetivo

de identificar e caracterizar *miRNAs*/*pre-miRNAs*, tais como: miRDeep* (AN et al., 2013), miRDeep2 (FRIEDLANDER et al., 2012), miRDeep-P (YANG & LI, 2011), miREvo (WEN et al., 2012) e miRPlant (AN et al., 2014). Poucas ferramentas, como o miRExpress (WANG et al., 2009), tem o objetivo de analisar o perfil de expressão, entretanto, este é gerado seguindo a contagem de *miRNAs* encontrados na etapa de alinhamento, levando em consideração cálculos estatísticos para diferença de expressão entre os *miRNAs*.

REFERÊNCIAS

AGILENT, T. Custom ChIP-on-chip Microarrays. 2017. Disponível em: < http://www.genomics.agilent.com >. Acesso em: 10/01.

AN, J. et al. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. **Nucleic Acids Res,** v. 41, 2013.

AN, J. et al. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. **BMC Bioinformatics,** v. 15, n. 1, p. 275, 2014. ISSN 1471-2105.

ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. **Genome Biol,** v. 11, 2010.

ANDREWS, S. A quality control tool for high throughput sequence data., 2010. Disponível em: < http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. >. Acesso em: 12/07.

ARONICA, E. et al. Expression pattern of miR-146a, an inflammation-associated microRNA, in experimental and human temporal lobe epilepsy. **European Journal of Neuroscience,** v. 31, n. 6, p. 1100-1107, 2010. ISSN 1460-9568.

AUER, P. L.; DOERGE, R. W. Statistical Design and Analysis of RNA Sequencing Data. **Genetics,** v. 185, n. 2, p. 405-416, 2010.

AXTELL, M. J.; WESTHOLM, J. O.; LAI, E. C. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. **Genome Biology,** v. 12, n. 4, p. 221, 2011. ISSN 1474-760X.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society. Series B (Methodological),** v. 57, n. 1, p. 289-300, 1995. ISSN 00359246.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114-20, 2014. ISSN 1367-4811 (Electronic) 1367-4803 (Linking).

BRAY, N. L. et al. Near-optimal probabilistic RNA-seq quantification. **Nat Biotech,** v. 34, n. 5, p. 525-527, 2016. ISSN 1087-0156.

CARTHEW, R. W.; SONTHEIMER, E. J. Origins and Mechanisms of *miRNAs* and siRNAs. **Cell,** v. 136, n. 4, p. 642-655, 2009. ISSN 0092-8674.

CASNEUF, T.; VAN DE PEER, Y.; HUBER, W. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. **BMC Bioinformatics,** v. 8, n. 1, p. 1-13, 2007. ISSN 1471-2105.

CASTOLDI, M. et al. A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). **RNA,** v. 12, n. 5, p. 913-920, 2006. ISSN 1355-8382 1469-9001.

CHEVREUX, B. et al. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. **Genome Research,** v. 14, n. 6, p. 1147-1159, 2004. ISSN 1088-9051.

CONESA, A. et al. A survey of best practices for RNA-seq data analysis. **Genome Biol,** v. 17, n. 1, p. 13, 2016. ISSN 1474-760X (Electronic) 1474-7596 (Linking).

CONSORTIUM, S. M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. **Nat Biotech,** v. 32, n. 9, p. 903-914, 2014. ISSN 1087-0156.

CONSORTIUM, T. R. RNAcentral: a comprehensive database of non-coding RNA sequences. **Nucleic Acids Research,** v. 45, n. D1, p. D128-D134, 2017. ISSN 0305-1048.

DAI, M. et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. **BMC Genomics,** v. 11, n. 4, p. 1-9, 2010. ISSN 1471-2164.

DE SOUSA CARDOSO, T. C. et al. Genome-wide identification and in silico characterisation of microRNAs, their targets and processing pathway genes in Phaseolus vulgaris L. **Plant Biol (Stuttg)**, 2015. ISSN 1438-8677 (Electronic) 1435-8603 (Linking).

DEBAT, H. J.; DUCASSE, D. A. Plant microRNAs: Recent Advances and Future Challenges. **Plant Molecular Biology Reporter,** v. 32, n. 6, p. 1257-1269, 2014. ISSN 0735-9640 1572-9818.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics,** v. 29, n. 1, p. 15-21, 2013. ISSN 1367-4811 (Electronic)1367-4803 (Linking).

EMDE, A.-K. et al. MicroRazerS: rapid alignment of small RNA reads. **Bioinformatics,** v. 26, n. 1, p. 123-124, 2010.

FENG, J. et al. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. **Bioinformatics,** v. 28, n. 21, p. 2782-2788, 2012. ISSN 1367-4803.

FRIEDLANDER, M. R. et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. **Nucleic Acids Res,** v. 40, 2012.

FURTADO, D. **Estatística Básica**. 2005. ISBN 9788587692719.

GARBER, M. et al. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nat Meth,** v. 8, n. 6, p. 469-477, 2011. ISSN 1548-7091.

GIT, A. et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. **Rna,** v. 16, 2010.

GIT, A. et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. **RNA,** v. 16, n. 5, p. 991-1006, 2010. ISSN 1355-8382 1469-9001.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nat Biotechnol,** v. 29, n. 7, p. 644-52, 2011. ISSN 1546-1696 (Electronic) 1087-0156 (Linking).

GRIFFITH, M. et al. Alternative expression analysis by RNA sequencing. **Nat Meth,** v. 7, n. 10, p. 843-847, 2010. ISSN 1548-7091.

GRIFFITH, M. et al. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. **Plos Computational Biology,** v. 11, n. 8, p. 20, 2015. ISSN 1553-734X.

GUL, A. et al. Microarray: gateway to unravel the mystery of abiotic stresses in plants. **Biotechnology Letters,** v. 38, n. 4, p. 527-543, 2016. ISSN 0141-5492.

HANNON, G. J. FASTX-Toolkit. 2009. Disponível em: < http://hannonlab.cshl.edu/fastx_toolkit/ >. Acesso em: 12/07.

HIGUCHI, R. et al. Simultaneous Amplification and Detection of Specific DNA Sequences. **Nat Biotech,** v. 10, n. 4, p. 413-417, 1992.

HODKINSON, B. P.; GRICE, E. A. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. **Advances in Wound Care,** v. 4, n. 1, p. 50-58, 2015. ISSN 2162-1918 2162-1934.

JOHNSON, B. K. et al. SPARTA: Simple Program for Automated reference-based bacterial RNA-seq Transcriptome Analysis. **Bmc Bioinformatics,** v. 17, p. 4, 2016. ISSN 1471-2105.

KHANG, T. F.; LAU, C. Y. Getting the most out of RNA-seq data analysis. **PeerJ,** v. 3, p. e1360, 2015.

KROLL, M. H. Evaluating Sequential Values Using Time-adjusted Biological Variation. **Clinical Chemistry and Laboratory Medicine,** v. 40, n. 5, p. 499-504, 2005. ISSN 1434-6621.

KUKURBA, K. R.; MONTGOMERY, S. B. RNA Sequencing and Analysis. **Cold Spring Harb Protoc,** v. 2015, n. 11, p. pdb top084970, 2015. ISSN 1559-6095 (Electronic) 1559-6095 (Linking).

LAMONTAGNE, J.; STEEL, L. F.; BOUCHARD, M. J. Hepatitis B virus and microRNAs: Complex interactions affecting hepatitis B virus replication and hepatitis B virus-associated diseases. **World Journal of Gastroenterology : WJG,** v. 21, n. 24, p. 7375-7399, 2015. ISSN 1007-9327 2219-2840.

LANGMEAD, B.; HANSEN, K. D.; LEEK, J. T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. **Genome Biology,** v. 11, n. 8, p. 1-11, 2010. ISSN 1474-760X.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Meth,** v. 9, n. 4, p. 357-359, 2012. ISSN 1548-7091.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology,** v. 10, n. 3, p. 1-10, 2009. ISSN 1474-760X.

LI, B. et al. RNA-Seq gene expression estimation with read mapping uncertainty. **Bioinformatics,** v. 26, n. 4, p. 493-500, 2010.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics,** v. 25, n. 14, p. 1754-1760, 2009.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, 2009. ISSN 1367-4803 1460-2059.

LI, J. et al. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. **Biostatistics,** v. 13, n. 3, p. 523-538, 2012. ISSN 1465-4644.

LIU, Y.; ZHOU, J.; WHITE, K. P. RNA-seq differential expression studies: more sequence or more replication? **Bioinformatics,** v. 30, n. 3, p. 301-304, 2014. ISSN 1367-4803 1367-4811.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology,** v. 15, n. 12, p. 550, 2014. ISSN 1465-6906 1465-6914.

MENG, Y. et al. Expression-Based Functional Investigation of the Organ-Specific MicroRNAs in Arabidopsis. **PLoS ONE,** v. 7, n. 11, p. e50870, 2012. ISSN 1932-6203.

METZKER, M. L. Sequencing technologies [mdash] the next generation. **Nat Rev Genet,** v. 11, n. 1, p. 31-46, 2010. ISSN 1471-0056.

MOREIRA, L. M. **Ciências genômicas : fundamentos e aplicações**. Ribeirão Preto: Sociedade Brasileira de Genética, 2015. 403 ISBN 978-85-89265-22-5.

MUIR, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. **Genome Biology,** v. 17, n. 1, 2016. ISSN 1474-760X.

NAKAMURA, Y.; COCHRANE, G.; KARSCH-MIZRACHI, I. The International Nucleotide Sequence Database Collaboration. **Nucleic Acids Research,** v. 41, n. Database issue, p. D21-D24, 2013. ISSN 0305-1048 1362-4962.

NCBI. Sequence Read Archive.NCBI/NLM/NIH. 2016. Disponível em: < http://www.ncbi.nlm.nih.gov/Traces/sra/ >. Acesso em: 19/07.

NEVADO, B.; PEREZ-ENCISO, M. Pipeliner: software to evaluate the performance of bioinformatics pipelines for next-generation resequencing. **Molecular Ecology Resources,** v. 15, n. 1, p. 99-106, 2015. ISSN 1755-098X.

NIELSEN, R. et al. Genotype and SNP calling from next-generation sequencing data. **Nat Rev Genet,** v. 12, n. 6, p. 443-451, 2011. ISSN 1471-0056.

OZSOLAK, F.; MILOS, P. M. RNA sequencing: advances, challenges and opportunities. **Nature Reviews Genetics,** v. 12, n. 2, p. 87-98, 2011. ISSN 1471-0056.

PABINGER, S. et al. A survey of tools for the analysis of quantitative PCR (qPCR) data. **Biomolecular Detection and Quantification,** v. 1, n. 1, p. 23-33, 2014. ISSN 22147535.

PACHTER, L. Near-optimal RNA-Seq quantification with kallisto. 2015. Disponível em: < https://liorpachter.wordpress.com/2015/05/10/near-optimal-rna-seq-quantification-with-kallisto/ >. Acesso em: 19/07.

PAREEK, C. S.; SMOCZYNSKI, R.; TRETYN, A. Sequencing technologies and genome sequencing. **Journal of Applied Genetics,** v. 52, n. 4, p. 413-435, 2011. ISSN 1234-1983 2190-3883.

PIERCE, B. A. **Genetics: A Conceptual Approach**. 5. W. H. Freeman, 2013. ISBN 978-1464109461.

PIMENTEL, H. What the FPKM? A review of RNA-Seq expression units. 2014. Disponível em: < https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/ >. Acesso em: 19/07.

PRÜFER, K. et al. PatMaN: rapid alignment of short sequences to large databases. **Bioinformatics,** v. 24, n. 13, p. 1530-1531, 2008. ISSN 1367-4803 1460-2059.

PUNDHIR, S.; POIRAZI, P.; GORODKIN, J. Emerging applications of read profiles towards the functional annotation of the genome. **Frontiers in Genetics,** v. 6, p. 188, 2015. ISSN 1664-8021.

RAJKUMAR, A. P. et al. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. **BMC Genomics,** v. 16, p. 548, 2015. ISSN 1471-2164 (Electronic) 1471-2164 (Linking).

RHOADS, A.; AU, K. F. PacBio Sequencing and Its Applications. **Genomics, Proteomics & Bioinformatics,** v. 13, n. 5, p. 278-289, 2015. ISSN 1672-0229 2210-3244.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics,** v. 26, n. 1, p. 139-140, 2010. ISSN 1367-4803 1367-4811.

ROBINSON, M. D.; SMYTH, G. K. Moderated statistical tests for assessing differences in tag abundance. **Bioinformatics,** v. 23, n. 21, p. 2881-2887, 2007.

SANGER, F. The Croonian Lecture, 1975: Nucleotide Sequences in DNA. **Proceedings of the Royal Society of London B: Biological Sciences,** v. 191, n. 1104, p. 317-333, 1975.

SARWAT, M.; YAMDAGNI, M. M. DNA barcoding, microarrays and next generation sequencing: recent tools for genetic diversity estimation and authentication of medicinal plants. **Critical Reviews in Biotechnology,** v. 36, n. 2, p. 191-203, 2016. ISSN 0738-8551.

SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics,** v. 27, n. 6, p. 863-864, 2011. ISSN 1367-4803 1367-4811.

SCHORDERET, P. NEAT: a framework for building fully automated NGS pipelines and analyses. **Bmc Bioinformatics,** v. 17, 2016.

SEO, M. et al. RNA-seq analysis for detecting quantitative trait-associated genes. **Scientific Reports,** v. 6, p. 24375, 2016. ISSN 2045-2322.

SHENDURE, J. The beginning of the end for microarrays? **Nat Meth,** v. 5, n. 7, p. 585-587, 2008. ISSN 1548-7091.

SONG, J. et al. Identification of Suitable Reference Genes for qPCR Analysis of Serum microRNA in Gastric Cancer Patients. **Digestive Diseases and Sciences,** v. 57, n. 4, p. 897-904, 2012.

SUN, Z. et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. **BMC Genomics,** v. 15, n. 1, p. 423, 2014. ISSN 1471-2164.

SZYRAJEW, K. et al. MicroRNAs Are Intensively Regulated during Induction of Somatic Embryogenesis in Arabidopsis. **Frontiers in Plant Science,** v. 8, n. 18, 2017. ISSN 1664-462X.

TAM, S.; TSAO, M. S.; MCPHERSON, J. D. Optimization of miRNA-seq data preprocessing. **Brief Bioinform,** v. 16, n. 6, p. 950-63, 2015. ISSN 1477-4054 (Electronic) 1467-5463 (Linking).

TRAPNELL, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nat Protoc,** v. 7, n. 3, p. 562-78, 2012. ISSN 1750-2799 (Electronic) 1750-2799 (Linking).

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nat Biotech,** v. 28, n. 5, p. 511-515, 2010. ISSN 1087-0156.

WANG, H. L. V.; CHEKANOVA, J. A. Small RNAs: essential regulators of gene expression and defenses against environmental stresses in plants. **Wiley Interdisciplinary Reviews-Rna,** v. 7, n. 3, p. 356-381, 2016. ISSN 1757-7004.

WANG, K. et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. **Nucleic Acids Research,** v. 38, n. 18, p. e178-e178, 2010. ISSN 0305-1048 1362-4962.

WANG, W.-C. et al. miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression. **BMC Bioinformatics,** v. 10, n. 1, p. 328, 2009. ISSN 1471-2105.

WEN, M. et al. miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. **BMC Bioinformatics,** v. 13, p. 140-140, 2012. ISSN 1471-2105.

YANG, X.; LI, L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. **Bioinformatics,** v. 27, n. 18, p. 2614-2615, 2011.

ZHANG, L.; CHEN, S. C.; LIU, X. J. Detecting differential expression from RNA-seq data with expression measurement uncertainty. **Frontiers of Computer Science,** v. 9, n. 4, p. 652-663, 2015. ISSN 2095-2228.

ZHAO, Y. et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. **Nucleic Acids Research,** v. 44, n. D1, p. D203-D208, 2016. ISSN 0305-1048.

CAPÍTULO 2

pipeMIRSEQ: *pipeline* integrativo para análises de expressão diferencial em dados de *miRNA-seq* de plantas

RESUMO

Uma das técnicas mais utilizadas para identificação e quantificação em larga escala de RNAs é o *RNA-seq*. Entretanto para se analisar os dados provenientes da mesma, deve-se utilizar ferramentas computacionais. Com o objetivo de analisar o nível de expressão de microRNAs, algumas etapas devem ser implementadas, porém diferentes ferramentas são utilizadas para cada etapa. Logo, muitas plataformas e *pipelines* vêm sido desenvolvidos para integrar tais ferramentas, otimizando o tempo gasto com cada etapa. Entretanto a maioria dos *pipelines* não são desenvolvidos para analisar expressão diferencial em *miRNAs* de plantas. Com o intuito de avaliar o nível da expressão diferencial em plantas foi desenvolvido um *pipeline*¸ pipeMIRSEQ. Ele foi desenvolvido em cinco módulos: controle de qualidade, mapeamento dos reads, quantificação dos *miRNAs*, diferença de expressão e etapa de *review*. Na etapa de controle de qualidade, foi utilizado as ferramentas, FastQC, Minion, Trimommatic e Prinseq Lite. Na etapa de mapeamento foi comparado três alinhadores, Bowtie 1, Bowtie 2 e BWA. A etapa de quantificação foi elaborada levando em consideração os *miRNAs* homólogos, e na diferença de expressão foram utilizados dois pacotes para cálculos, edgeR e DESEq2 e pôr fim a etapa de *review* teve como objetivo sintetizar os resultados obtidos nas etapas anteriores. Foi utilizado para validação, uma biblioteca de microRNAs de café arábica no estádio G5 de desenvolvimento floral de duas cultivares diferentes. O *pipeline* apresentou um bom controle de qualidade dos *miRNAs*. Dois alinhadores foram escolhidos, Bowtie 1 e BWA em vista de diversos fatores apresentados. No geral, o *pipeline* conseguiu alcançar seus objetivos e executar os dados com uma performance de tempo razoável.

Palavras-chave: *RNA-seq*. Mapeamento. Nível de expressão.

ABSTRACT

One of the most used techniques for identification and quantification in large-scale of RNA is RNA-seq. However, to analyze the data, computational tools are needed. Aiming at analyzing the level of expression of miRNAs, some steps might be implemented, but different tools are used in each step. Therefore, many platforms and pipelines have been used to integrate these tools, optimizing the time spent in each step. However, most of the pipelines are not developed for differential expression analyses of plant miRNAs. To calculate the level of differential expression in plants a pipeline was developed, pipeMIRSEQ. It was developed in five modules: quality control, reads mapping, quantification of miRNAs, and differential expression calculation and a review step. In the step of quality control, the tools FastQC, Minion, Trimmomatic and Prinseq Lite were used. At the mapping step, three aligners were compared - Bowtie 1, Bowtie 2 and BWA. The quantification step was made considering the homolog miRNAs, and for differential expression calculation two packages were used, edgeR and DESEq2. At the end, the review step aimed at synthesizing the results from the previous steps. For validation, a miRNA library of *Coffea arabica* at G5 stage of floral development from two cultivars was used. The pipeline presented a good quality control of the miRNAs. Two aligners were chosen, Bowtie 1 and BWA, due to several aspects. In general, the pipeline achieved its aims and executed the data in a reasonable time.

Keywords: *RNA-seq*. Mapping. Expression level

1. INTRODUÇÃO

Devido aos custos menores de sequenciamento e a possibilidade da multiplexação de várias bibliotecas em paralelo, obtendo assim um grande volume de dados (METZKER, 2010; NIELSEN et al., 2011), a utilização das técnicas conhecidas como sequenciamento de segunda geração (NGS) vem crescendo nas últimas décadas (SHENDURE, 2008; MUIR et al., 2016). Tais métodos além de fazer a identificação das sequências de nucleotídeos de DNA e RNA (MOREIRA, 2015), apresentam elevada sensibilidade de detecção, proporcionando estudos com novos transcritos, como fusões de genes, expressão específica de alelos, transcritos novos alternativos e monitoramento do perfil de expressão de microRNAs (GIT et al., 2010; JOHNSON et al., 2016).

MicroRNAs são pequenos RNAs não codantes com aproximadamente 19-24 nucleotídeos, que apresentam a capacidade de regular a expressão de genes a nível pós-transcriptional, interferindo na estabilidade dos RNAs e/ou silenciando a expressão de gênica (BARTEL, 2004). Atualmente, a utilização da tecnologia chamada *miRNA-seq*, a qual é baseada no método de *RNA-seq*, permite a identificação de microRNAs, e determinação do perfil de expressão de microRNAs (WEN et al., 2012; TAM et al., 2015), devido à sua alta sensibilidade (GIT et al., 2010). A utilização da técnica de sequenciamento de nova geração abriu novos caminhos para uma compreensão mais profunda dos processos biológicos e seus mecanismos de regulação (D'ANTONIO et al., 2015).

As metodologias e ferramentas utilizadas para a análise de mi*RNA- seq* variam de acordo com o objetivo final do projeto, mas apresenta algumas etapas em comum para com estudos envolvendo outros tipos de RNA ou mesmo DNA, como: controle de qualidade e alinhamento dos *reads* (sequência de nucleotídeos de RNA ou DNA geradas a partir das análises de *miRNA-seq, RNA-seq* ou *CHIP-seq*) (TAM et al., 2015). Entretanto, a identificação de perfis de expressão de RNAs

envolve algumas etapas adicionais, como a quantificação dos transcritos e o cálculo da diferença de expressão (GIT et al., 2010).

Análises de dados provenientes de sequenciamento em larga escala apresentam grandes volumes de dados, que podem variar na grandeza de gigabytes (HODKINSON & GRICE, 2015), gerando assim grandes desafios computacionais para análises desses dados, o que requer do conhecimento de bioinformática para a obtenção e compreensão do resultado obtido que requerem maior atenção (AUER & DOERGE, 2010). Devido a estes fatores, o desenvolvimento de plataformas, *pipelines* (integração de ferramentas computacionais para análises de bioinformática) e softwares que analisam os dados gerados pelas tecnologias de NGS são essenciais (NEVADO & PEREZ-ENCISO, 2015). Tais plataformas possuem distintos objetivos, como identificação de novos *miRNAs* (AN et al., 2014), detecção de polimorfismo de um único nucleotídeo (SUN et al., 2014), predição de estruturas secundárias (AN et al., 2013) e detecção de perfil de expressão de *miRNAs* (WANG et al., 2009).

Os microRNAs estão presentes em plantas e animais, entretanto se diferem em vários aspectos (AXTELL et al., 2011). Para a detecção do perfil de expressão, a maioria das ferramentas são desenvolvidas com o objetivo de analisar *miRNAs* animais (FRIEDLANDER et al., 2012; SUN et al., 2014; MINIANDRÉS-LEÓN et al., 2016). Já em plantas, poucas ferramentas foram desenvolvidas com o objetivo de avaliar esses ácidos nucléicos em específico (WANG et al., 2009). Neste contexto, este estudo teve como objetivo a criação de um pipeline automatizado, otimizado, robusto e rápido para análise de expressão diferencial em dados de *miRNA-seq* em plantas, o qual possa ser utilizado por usuários com pouco ou nenhum conhecimento técnico em programação de computadores e computação.

2. MATERIAL E MÉTODOS

2.1. Análise de requisitos

Para se garantir uma análise de requisitos de sucesso deve-se trabalhar com pessoas que tem influências e entendimento das premissas de um sistema, conhecidas como *stakeholders* (SOMMERVILLE, 2003). A análise de requisitos teve como objetivo conhecer requisitos relevantes e estabelecer um consenso entre os *stakeholders* a respeito dos mesmos. O sucesso desta etapa reflete diretamente no sucesso e na qualidade do *software* a ser desenvolvido (KLAUS POHL, 2011).

Para garantir o conhecimento necessário foram abordadas três etapas. A primeira etapa foi a elicitação, onde foi aplicada a técnica de *brainstorming*; etapa de levantamento de dados com o objetivo de definir as ferramentas utilizadas para análise de *miRNA-seq*; e por fim a etapa de documentação dos requisitos, que definiu a compilação dos resultados das etapas anteriores.

2.1.1. *Brainstroming*

Brainstorming é um método que consiste em várias reuniões com *stakeholders* com o objetivo de dar origem ao levantamento de requisitos do sistema (SOMMERVILLE, 2003).

Foi promovido um total de quatro *brainstormings*, que após rodadas de discussão alcançou respostas para as seguintes questões:

a) Quais plataformas de *NGSs* mais utilizadas pelo grupo de pesquisa?
b) Quais os protocolos utilizados na análise de expressão de dados de *RNA-seq* pelo grupo de pesquisa?

2.1.2. Levantamento de dados

Na etapa de levantamento de dados foi levado em consideração sete plataformas de análises de *miRNA*: CAP-miRSeq (SUN et al., 2014) e miARma-Seq (MINIANDRÉS-LEÓN et al., 2016), miRDeep* (AN et al., 2013) as quais são utilizadas para *miRNAs* de animais, e miRPlant (AN et al., 2014) e o miRDeep-P (YANG & LI, 2011) as quais são utilizadas para *miRNAs* de plantas, e miREvo (WEN et al., 2012) e miRExpress (WANG et al., 2009) utilizados tanto em plantas como em animais.

2.1.3. Documentação dos requisitos

Através da etapa de documentação, foram definidos os módulos e as ferramentas que o *pipeline* seria desenvolvido levando em consideração a etapa de *brainstroming* e levantamento de dados.

Através das sete plataformas foram selecionadas as ferramentas que seriam utilizadas para compor o *pipeline*. Na etapa de controle de qualidade: FastqQC versão 0.11.5 (ANDREWS, 2010), Minion (DAVIS et al., 2013), Trimmomatic versão 0.36 (BOLGER et al., 2014) e Prinseq Lite versão 0.20.4 (SCHMIEDER & EDWARDS, 2011). Na etapa de mapeamento dos *reads* foram feitos testes com três ferramentas, Bowtie 1 versão 1.2 (LANGMEAD et al., 2009), Bowtie 2 versão 2.3 (LANGMEAD & SALZBERG, 2012) e BWA versão 0.7.15 (LI & DURBIN, 2009). Na etapa de quantificação dos *miRNAs* a ferramenta SAMTools versão 1.3.1 (LI et al., 2009) foi utilizada para extrair informações de arquivos de alinhamento, e com essa saída foi desenvolvido uma estratégia para a quantificação dos reads alinhados. E por fim, na etapa de cálculo para diferença de expressão foi utilizado a linguagem de programação estatística R (R DEVELOPMENT CORE TEAM, 2009) usando os pacotes DESeq2

(ANDERS & HUBER, 2010) e edgeR (ROBINSON et al., 2010) interpretado pela linguagem de programação Perl.

2.2. Base de dados

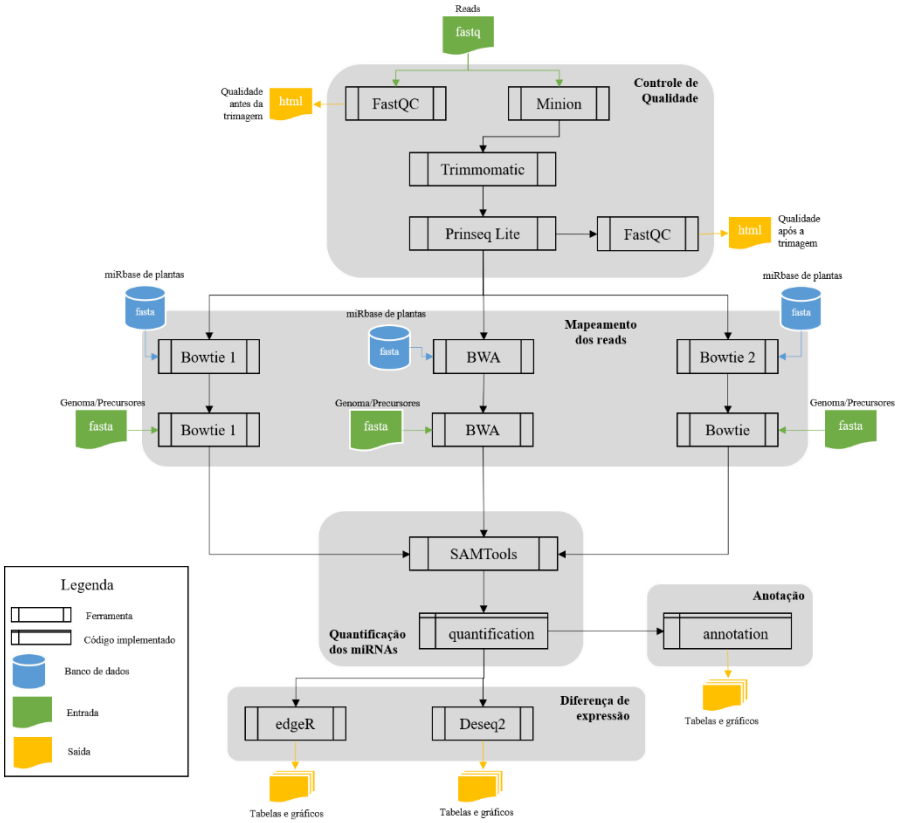
Foi selecionado a base de dados de microRNAs maduros pertencentes ao miRBase (KOZOMARA & GRIFFITHS-JONES, 2013)), versão 21 (última versão). Posteriormente, essa base de dados foi passada por uma curadoria manual, afim de retirar todos os maduros que não pertenciam a plantas, resultando em 8.509 *miRNAs* maduros. O sequenciamento de dados da plataforma Illumina resulta em cDNAs que possuem o nucleotídeo timina, entretanto, a base de dados de microRNAs maduros do miRBase possui o nucleotídeo Uracila, logo, todos os nucleotídeos de Uracila foram substituídos por nucleotídeos de Timina, para permitir um alinhamento preciso ao se usar o pipeline.

2.3. Implementação

Para a implementação do *pipeline* foi utilizado como linguagem principal a linguagem Perl (TOM CHRISTIANSEN, 2012). Em combinação a linguagem Perl foi utilizado pacotes em R e Shell em um ambiente GNU/Linux, utilizando o sistema operacional Ubuntu versão 14.04.

Para a implementação do código foi utilizado o Git (GIT et al., 2010), um sistema colaborativo de controle de versão de arquivos, que possibilita minimizar os riscos de perdas alterações sobrescritas no código. Toda a implementação e suas versões foram alocadas no Bitbucked (ATLASSIAN, 2008), um serviço de hospedagem de projetos e suas versões. O *pipeline* não necessita ser instalado, mas para rodá-lo deve se cumprir alguns requisitos, instalação do interpretador de Perl e algumas bibliotecas. Toda a implementação, os módulos, base de dados, arquivos de entrada e saída do *pipeline* estão descritos na Figura 2.2.

Figura 2.2 – Fluxograma do design do pipeMIRSEQ, com os principais módulos representados em cinza, arquivos de entrada representados por verde e de saída por amarelo.



2.3.1. Controle de qualidade

A primeira etapa do módulo de controle de qualidade tratou da geração de análises de qualidade através do software FastQC, com o objetivo de comparar dados brutos e dados *trimados*. Os parâmetros utilizados foram: –k a 3 que define um tamanho menor de *kmer* devido ao tamanho dos *reads* no sequenciamento, e –t que corresponde a quantidade de processadores no qual a ferramenta foi executada. Para a remoção dos adaptadores, foi utilizado a ferramenta Minion,

que faz uma previsão de adaptadores em um arquivo de formato FASTA a partir de uma biblioteca de entrada.

Comumente o sequenciamento de *miRNAs* utilizando plataformas de NGS é realizado em um sentido único (*single-end*), logo toda a estratégia e utilização das ferramentas seguem o modelo *single-end*. Após a predição dos adaptadores, os *reads* foram *trimados* utilizando o *software* Trimmomatic na seguinte ordem: retirada dos adaptadores com *mismatch* de 2 e *threshold* de 10, que corresponde a precisão entre o adaptador e retirada dos *reads* com o tamanho menor do que 17 pb. Optou-se pela não retirada das bases da cabeça e cauda da sequência dos *reads* que possuíam baixa qualidade, pois poderiam comprometer etapas posteriores do protocolo, como o mapeamento dos *reads*. Os *miRNAs* de plantas, em sua maioria, apresentam tamanho entre 20 a 24pb (DEBAT & DUCASSE, 2014), e, dessa forma, *reads* com o tamanho maior que 30 bases foram descartados através da ferramenta *Prinseq Lite*, evitando também a caraterização de outros *smRNAs*. Ao final de cada etapa da *trimagem*, a biblioteca foi submetida novamente à ferramenta FastQC, com os mesmos parâmetros usados anteriormente.

2.3.2. Mapeamento dos reads

Para mapear as bibliotecas de *miRNA-seq* foram confrontados três *unspliced read aligners*; Bowtie 1, Bowtie 2 e BWA. Primeiramente, os *reads trimados* foram alinhados contra uma base de dados curada de *miRNAs* de plantas, o *miRBase*, com o objetivo de caracterizar na espécie de estudo os ortólogos de *microRNAs*. A próxima etapa, opcional para o usuário, foi utilizar os *reads* não mapeados no *miRBase*, e alinhá-los em um genoma ou banco de precursores como entrada. Esta etapa tem como objetivo identificar microRNAs específicos de uma espécie e também a identificação de novos *miRNAs*. Essas estratégias foram desenvolvidas paralelamente para os três alinhadores utilizados. Para todos os três alinhadores é preciso construir um arquivo index com a base de dados que será

alinhada, o qual foi utilizado para otimizar o alinhamento (LANGMEAD et al., 2009; LI & DURBIN, 2009; LANGMEAD & SALZBERG, 2012).

O alinhador Bowtie 1 foi utilizado para o primeiro alinhamento, seguindo os seguintes parâmetros: -p número de processadores; -n 2, aceita até dois *mismatches;* -l 8, considera um *seed* (tamanho de comparação de alinhamento) menor que o *default* devido ao tamanho dos *miRNAs*; os parâmetros –a --best –strata trabalham em conjunto, o --best e –a reporta os melhores alinhamentos para um *read* que tem menor número de incompatibilidade, induzindo a importância do *seed* e o *mismatch*, que acoplados ao –strata retornam o melhor alinhamento do *read* levando em consideração as incompatibilidades, se a pontuação dos alinhamentos empatarem ele retorna todos os melhores alinhamentos; foi adicionado o parâmetro –M 1 para retornar apenas um melhor alinhamento, a escolha do melhor é feito aleatoriamente em caso de empate. No segundo alinhamento, alinhando contra os *reads* não alinhados no miRBase, o único parâmetro modificado em relação ao primeiro alinhamento foi –n 1 modificado para 1 *mismatch*, pois este segundo alinhamento tem o objetivo de utilizar uma base de dados da mesma espécie do *miRNA-seq* ou mais próxima.

BWA é um pacote de alinhadores, com o objetivo de alinhar pequenos *reads* em grandes referências (LI & DURBIN, 2009). O algoritmo do pacote BWA utilizado foi o *BWA-backtrack*, por ser desenhado para *reads* provenientes da plataforma Illumina. Os seguintes parâmetros foram utilizados foram: -t, para número de processadores, -n 1, permitindo até 1 mismatch; -o 0 e –e 0, permitindo nenhum gap no alinhamento; -k 1, distância máxima de edição (soma de *mismatch* e *gap*) por *seed* (TAM et al., 2015).

Bowtie 2 é um alinhador de sequências rápido e apresenta melhor alinhamentos utilizando *reads* acima de 50pb (LANGMEAD & SALZBERG, 2012). Os parâmetros utilizados no Bowtie 2 foram: -p, número de processadores; --local, Bowtie 2 executa o alinhamento do *read* local, podendo assim omitir as

bases das pontas para maximizar a pontuação do alinhamento; *--very-sensitive-local*, para aumentar a sensibilidade de um alinhamento.

2.3.3. Quantificação dos *reads*

A etapa de quantificação é baseada na contagem de *reads* (CONESA et al., 2016). A contagem dos alinhamentos foi feita com auxílio da ferramenta SAMTools, por meio de comandos básicos do Linux e utilizando o pacote Bioperl (STAJICH et al., 2002). O SAMtools, auxilia na manipulação de arquivos de saída de alinhamentos, formato SAM e BAM, os quais possuem um formato de texto onde os campos são separados por *tab ("\t")*. Juntamente com comandos como *grep, cut, wc* e outros foi possível a quantificar os *reads* sem utilização custosa da memória. Para auxiliar na contagem, foi utilizado o pacote Bioperl, que é um conjunto de módulos na linguagem Perl que facilita o desenvolvimento de *scripts*, sendo utilizado para ler as bases de dados em formato fasta, tanto do miRBase quanto de qualquer outra base utilizada pelo usuário como entrada.

Vários *loci* de *microRNAs* expressam sequências maduras não idênticas, mas muito semelhantes (homologia), sendo que suas nomenclaturas possuem o mesmo número, mas se diferem com a colocação de uma letra minúscula no final. Para quantificar esses microRNAs homólogos, foi traçado uma estratégia que faz a contabilidade dos alinhamentos a partir da expressão regular "miR[0-9]*". Esta lógica de expressão foi retirada do banco de dados miRBase observando a anotação de microRNAs maduros de plantas (KOZOMARA & GRIFFITHS-JONES, 2013), logo, a quantificação final é dada de acordo com a homologia do *microRNA*.

2.3.4. Diferença de expressão

Como o objetivo da ferramenta é a identificação dos *microRNAs* diferencialmente expressos, foram utilizados dois pacotes desenvolvido em R,

edgeR (ROBINSON et al., 2010) e DESeq 2 (LOVE et al., 2014), para a normalização e os cálculos de diferença de expressão, tais como *p-value*, FDR, e $\log_2$ *fold change*. Para a utilização do R na linguagem Perl foi utilizado o pacote Statistics::R (P., 2004). Os dois pacotes utilizam a distribuição estatística binomial negativa (KHANG & LAU, 2015). Há diferenças entre os dois pacotes, o edgeR demonstra maior sensibilidade para identificar DEG (genes diferencialmente expressos) ao contrário do DESeq 2, que demonstra maior especificidade para identificar DEG (RAJKUMAR et al., 2015).

A tabela de quantificação dos *reads* gerada na etapa anterior e o arquivo gerado pelo *pipeline* com o design do experimento foram carregados primeiramente no edgeR. Foi feito o cálculo de normalização entre as bibliotecas, o qual leva em consideração a composição de RNA total por célula e, através deste valor de normalização, os genes DE são classificados. A classificação de DE em edgeR segue os seguintes critérios *p-value* < 0.01 e log_2 *fold change* ≥ 2 como *up-regulated*, e *p-value* < 0.01 e log_2 *fold change* ≤ 0.5 *down-regulated*.

O DEseq2, assim como o edgeR, requer a entrada dos dados não normalizados, apresentando uma normalização interna em relação ao tamanho da biblioteca (LOVE et al., 2014). Depois de realizada a normalização, os *counts* são submetidos a análise de diferença de expressão com os critérios *p-value* < 0.01 e log_2 *fold change* ≥ 2 como *up-regulated*, e *p-value* < 0.01 e log_2 *fold change* ≤ 0.5 *down-regulated*, as tabelas de DE dos dois pacotes contendo informações como log_2 *fold change*, *p-value* e FDR foram geradas, facilitando assim, análises pontuais dos DEG.

2.3.5. *Review*

Com a etapa de *review* objetivou-se retornar ao usuário informações para cada biblioteca de entrada: quantidade de *reads trimados*, quantidade de *reads* mapeados nas diferentes bases de dados e quantidade de *miRNAs* homólogos

alinhados, além de informar quais *reads* foram alinhados e suas sequências, facilitando assim análises posteriores. Nesta etapa, nenhuma ferramenta ou pacote exterior ao *pipeline* foi utilizada. Todo o código foi implementado utilizando as bases de dados, os arquivos de saída das *trimagens* e os alinhamentos.

3. RESULTADOS E DISCUSSÃO

3.1. Análise de requisitos

Na etapa de análise de requisitos foram levantadas todas as informações necessárias para o desenvolvimento da ferramenta. Foram selecionadas duas pessoas como *stakeholders* que possuem o conhecimento prático e técnico de sequenciamento e análises de *microRNAs*. Estas pessoas foram submetidas a reuniões, movidas com a técnica *brainstorming,* e como resultado final definiu-se as questões abordadas nas reuniões:

a) Escolha da plataforma Illumina de NGSs, levando em consideração que ela é uma das plataformas mais utilizadas e a que possui maior quantidade de algoritmos e protocolos para análise de *miRNA-seq*;
b) Nenhum protocolo de análise de *miRNA-seq* visando a DE foi estabelecido anteriormente pelo grupo de pesquisa;

Através da escolha da plataforma Illumina foram estudadas sete plataformas de análises de *miRNAs*, as quais apresentam diferentes objetivos. Destas sete, apenas quatro apresentam resultados testados em base de dados voltadas para plantas, miRDeep-P, miREvo, miRExpress e miRPlant, entretanto, a plataforma miRPlant não apresenta resultados de diferença de expressão em seu *pipeline,* como demonstrado na Tabela 3.1.

Tabela 3.1 – Informações de plataformas para análises de microRNAs. A coluna interface define se a plataforma foi desenvolvida com interface gráfica (GUI) ou linha de comando (CLI).

Plataformas	Organismos	Entrada	Saída	Linguagem	Interface	Referência
CAP-miRSeq	Animais	Reads, Genoma, miRBase	miRNA, Outros RNAs, DE	Shell, Perl, Python, R	CLI	Sun et al. (2014)
miARma-Seq	Animais	Reads	miRNA alvos, Outros *smRNAs*, DE, GO	Perl, R	CLI	Andrés-leon et al. (2016)
miRDeep*	Animais	Reads/Genoma/miRbase	pre-miRNA, miRNA, DE	Java	GUI	An et al. (2014)
miRDeep-P	Plantas	Reads, Genoma	pre-miRNA, miRNA, DE	C++	CLI	Wang et al. (2009)
miREvo	Animais, Plantas	Reads, Genoma	pre-miRNA, miRNA, DE	Perl	GUI	Yang; Li (2011)
miRExpress	Animais, Plantas	Reads, miRBase	pre-miRNA, miRNA, DE	Perl	CLI, GUI	Wen et al. (2012)
miRPlant	Plantas	Reads, Genoma	pre-miRNA, miRNA	Perl	CLI	An et al. (2013)
pipeMIRSEQ	Plantas	Reads/Genoma/miRbase	miRNA, DE	Shell, Perl, R	CLI	-

Tabela 3.2 – Ferramentas utilizadas em cada plataforma separadas por etapas de análises.

Plataformas	Controle de qualidade	Mapeamento	Quantificação	Diferença de expressão	Referência
CAP-miRSeq	FastQC, Cutadapt	Bowtie 1	HTSEq	edgeR	Sun et al. (2014)
miARma-Seq	FastQC, Minion, Cutadapt, Reaper	Bowtie 1, Bowtie 2, BWA	featureCounts, CIRI	edgeR, NOISeq	Andrés-leon et al. (2016)
miRDeep*	-	Bowtie 1	-	-	An et al. (2014)
miRDeep-P	-	Bowtie 1	-	-	Wang et al. (2009)
miREvo	-	Bowtie 1, Blat	-	-	Yang; Li (2011)
miRExpress	-	Smith-Waterman	-	-	Wen et al. (2012)
miRPlant	-	Bowtie 1	-	-	An et al. (2013)
pipeMIRSEQ	FastQC, Minion, Trimmomatic, Prinseq Lite	Bowtie 1	quantification, Samtools, Bioperl	DESeq2, edgeR	-

De modo geral, todas as plataformas definem quatro módulos principais: controle de qualidade, mapeamento dos *reads*, quantificação dos *reads* e por fim a etapa de cálculo da diferença de expressão. De acordo com as análises de dados provenientes do sequenciamento de *mRNA*, após o módulo de mapeamento dos *reads*, deve-se fazer um montagem dos transcritos (GRIFFITH et al., 2015; CONESA et al., 2016), entretanto, essa etapa não é necessária para análise de *miRNA-seq*, pois eles já possuem seu tamanho total na saída do próprio sequenciamento. Foram definidos então para o *pipeMIRSEQ* cinco módulos de implementação, controle de qualidade, mapeamento dos *reads*, quantificação dos *reads*, diferença de expressão e *review*.

3.2. Validação do *pipeline*

Para validação do *pipeline* foram utilizadas bibliotecas de *miRNA-seq* da espécie de *Coffea arabica L.* de duas cultivares, Catuaí e Siriema, na fase reprodutiva G5, caracterizado por gemas de 6,1 a 10 mm (coloração verde claro) (MORAIS, 2008). Por questão de eficiência, apenas 1% da representação total de cada biblioteca foi utilizado, com duas repetições biológicas de cada cultivar, a cultivar Catuaí foi considerada como controle (Tabela 3.3).

Tabela 3.3 – Descrição das bibliotecas utilizadas para validação do *pipeline* pipeMIRSEQ.

Biblioteca	Descrição	Design	Reads
G5_Cat_green_r1.fastq	G5 Catuaí	Controle	201845
G5_Cat_green_r2.fastq	G5 Catuaí	Controle	242382
G5_Sir_green_r1_S29_L006_R1_001.fastq	G5 Siriema	Tratamento	411303
G5_Sir_green_r2_S27_L006_R1_001.fastq	G5 Siriema	Tratamento	264122

Legenda: Nomes dos arquivos das bibliotecas, descrição de cada uma, design utilizado para o cálculo da DE, e quantidade de reads em cada biblioteca.

Além do alinhamento contra o miRBase, o *pipeline* utiliza a opção de alinhamento contra uma base de dados de entrada dada pelo usuário; a base de dados utilizada foi uma base de 467 precursores de *miRNAs* preditos a partir do genoma de *Coffea canephora L* (dados não publicados).

Foi dada a cada biblioteca um identificador, como pode ser visto na Tabela 3.4 e, para trazer maior clareza ao apresentar os resultados das análises, cada biblioteca será designada pelo seu identificador.

Tabela 3.4 – Identificador de cada biblioteca.

Identificador	**Biblioteca**
G5_ Cat_r1	G5_Cat_green_r1.fastq
G5_ Cat_r2	G5_Cat_green_r2.fastq
G5_ Sir_r1	G5_Sir_green_r1_S29_L006_R1_001.fastq
G5_ Sir_r2	G5_Sir_green_r2_S27_L006_R1_001.fastq

O *pipeline* apresentou como parâmetros as opções: -seqGen <STRING>, caminho e nome do banco de dados utilizado para o segundo alinhamento, parâmetro opcional ao usuário; -label <STRING>,<STRING>, identificador para as bibliotecas separados por vírgula; -locLib <STRING>, diretório onde se encontra todas as bibliotecas; -seqLib <STRING>,<STRING> , local e nomes das bibliotecas separados por vírgula; -threads <N>, número de processadores que serão utilizados pelo *pipeline*; -numRep <N>, número de repetições biológicas; -version, mostra a versão do *pipeline*, parâmetro opcional; -help, mostra o manual de ajuda, parâmetro opcional.

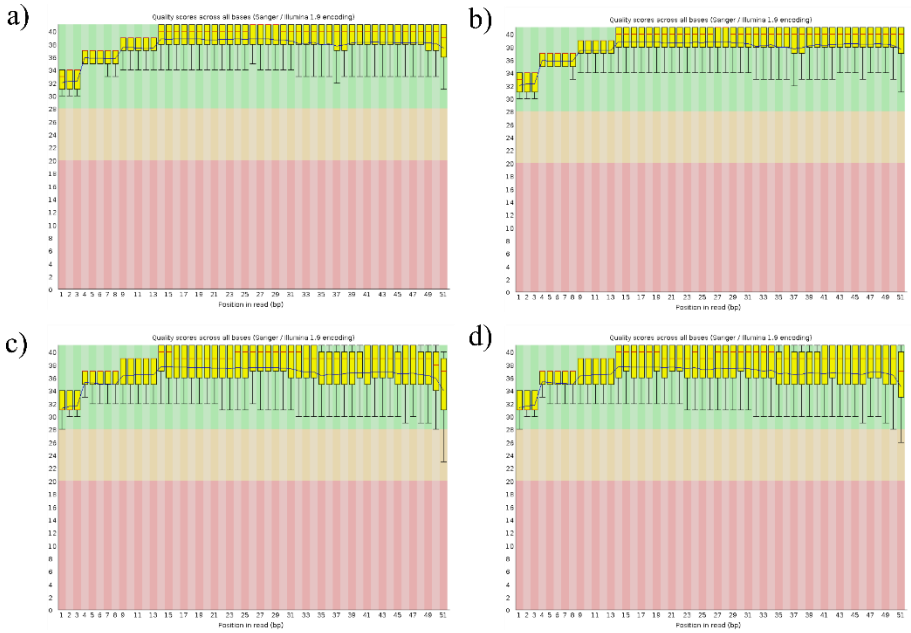
Todas as análises foram executadas em um computador com processador Intel(R) Core(TM) i5-4210U CPU 2.40GHz, com 8GB de memória RAM e 4 núcleos, utilizando Ubuntu 14.04. O comando para execução do pipeMIRSEQ com os parâmetros para as bibliotecas de validação foram:

```
perl pipeMIRSEQ.pl –seqLib ../dados/G5_Cat_green_r1.fastq.gz,
../dados/G5_Cat_green_r2.fastq.gz,../dados/G5_Sir_green_r1_S29_L006_R
1_001.fastq.gz,../dados/G5_Sir_green_r2_S27_L006_R1_001.fastq.gz -
seqGen precursor/precursores_final.fa -threads 4 -locLib ../dados/ -numRep
2 -label G5_Cat,G5_Sir
```

3.2.1. Controle de qualidade

Na análise de qualidade, o *pipeline* trabalhou com diferentes ferramentas: FastQC, para visualização antes e depois da *trimagem*, Minion para predição de adaptadores e Trimmomatic e Prinseq Lite, para o processo de limpeza.

Figura 3. 1 – Resumo dos valores de qualidade em todas as bases em cada posição no arquivo com extensão *fastq* de cada biblioteca antes do processo de *trimagem*.



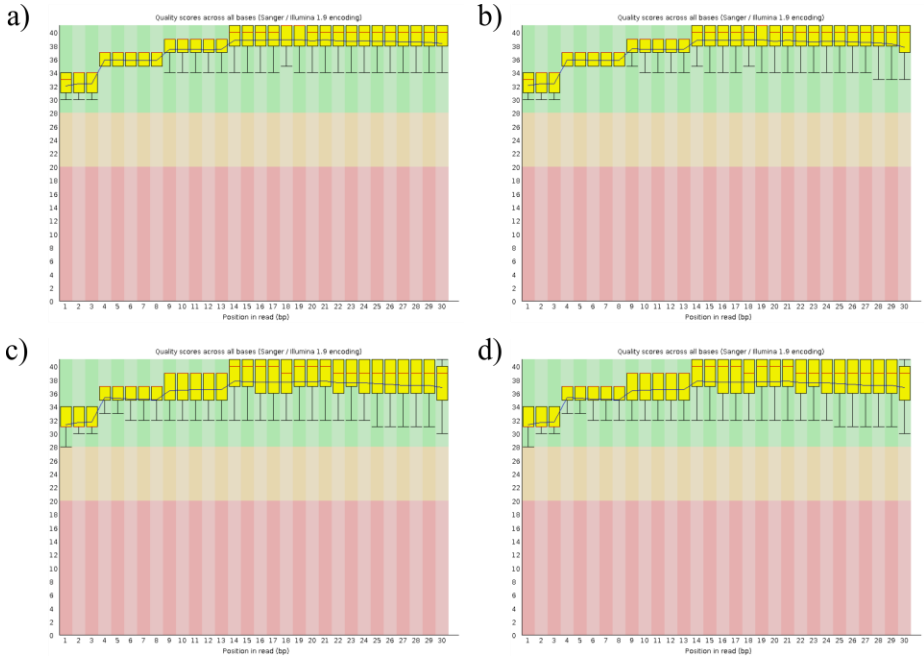
Legenda: No eixo x a representação das posições da base e no eixo y o valor de qualidade. **a)** Gráfico representando a biblioteca G5_ Cat_r1. **b)** Gráfico representando a biblioteca G5_ Cat_r2. **c)** Gráfico representando a biblioteca G5_ Sir_r1. **d)** Gráfico representando a biblioteca G5_ Sir_r2

Ao fazer o processo de *trimagem*, a primeira questão a ser visualizada é como está a qualidade de sequenciamento das bases. Na

Figura 3. 1 foram gerados gráficos de BoxWhisker que representam a qualidade das bases das sequências de uma biblioteca de modo geral, sendo esta

análise foi feita antes da limpeza dos dados possibilitando a observação de que todas as bibliotecas possuem, no geral, boa qualidade, pois os elementos do gráfico se encontram na cor verde do eixo y, que é definido como a qualidade esperada. A Figura 3.2 foi gerada pela ferramenta FastQC, após o processo de *trimagem*, e ela demonstra que o processo de limpeza dos *reads* não afetou a qualidade, apenas restringiu a utilização de *reads* que possuem entre 17 e 30pb. Lembrando que este corte só foi feito após a retirada dos possíveis adaptadores.

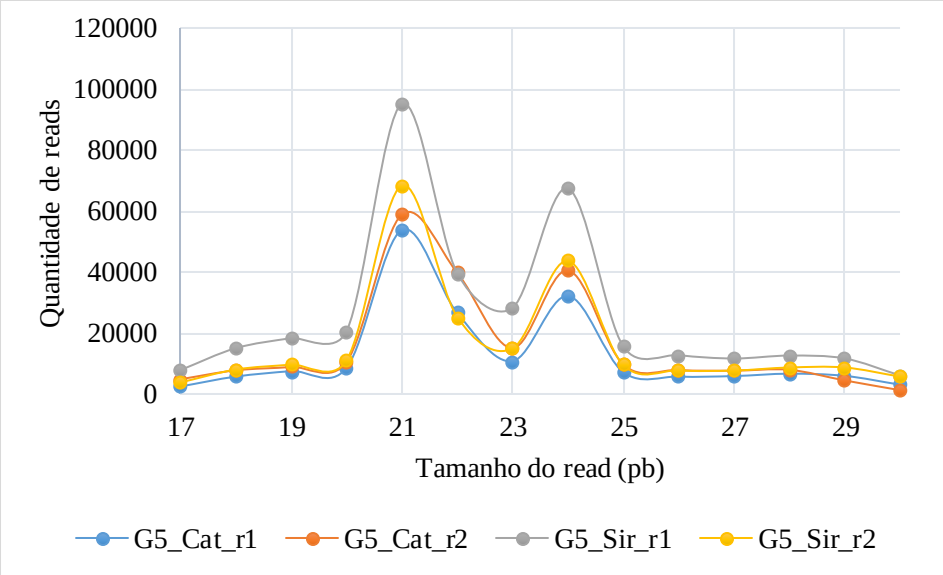
Figura 3.2 – Resumo dos valores de qualidade em todas as bases em cada posição no arquivo com extensão *fastq* de cada biblioteca.



Legenda*:* No eixo x a representação das posições da base e no eixo y o valor de qualidade após o processo de *trimagem*. **a)** Gráfico representando a biblioteca G5_ Cat_r1. **b)** Gráfico representando a biblioteca G5_ Cat_r2. **c)** Gráfico representando a biblioteca G5_ Sir_r1. **d)** Gráfico representando a biblioteca G5_ Sir_r2

O sequenciamento das bibliotecas apresentou todos os *reads* com 51 pb, após a retirada dos adaptadores, e de sequências menores que 17 pb e menores que 30 pb, sendo que todas as bibliotecas apresentaram maior concentração de *reads* com 21 a 24pb (Figura 3.3), corroborando assim com a literatura, que mostra que plantas processam classes de pequenos RNAs de 21-24 pb (SUNKAR, 2010).

Figura 3.3 – Quantidade de reads (eixo x) em relação ao tamanho do mesmo em pares de base (eixo y) nas quatro bibliotecas.



A etapa de análise de qualidade retirou, de cada biblioteca, cerca de 7-12% dos *reads* totais (Tabela 3.5). Optou-se pelo não tratamento de bases com baixa qualidade, pois sequenciamentos de pequenos RNAs, na sua maioria, não possuem baixa qualidade relatados devido ao tamanho dos *reads* sequenciados, além disso, a retirada de bases com baixa qualidade poderia comprometer significativamente a etapa de alinhamento devido ao tamanho dos *reads*.

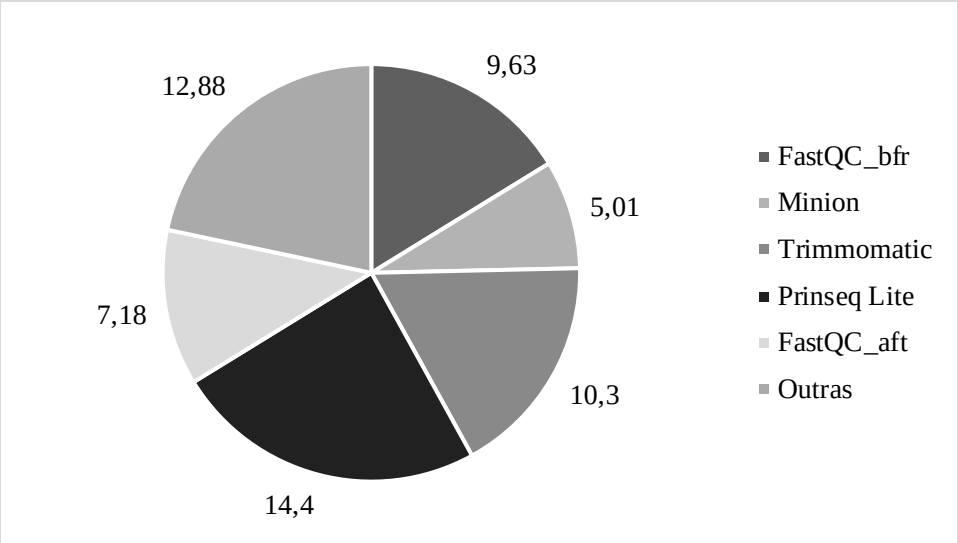
Tabela 3.5 – Tabela demonstrando a quantidade de *reads trimados* em relação a quantidade de *reads* totais em cada biblioteca

Biblioteca	**Reads**	**Reads após a *trimagem***	**Reads após a *trimagem* (%)**
G5_ Cat_r1	201845	182913	91
G5_ Cat_r2	242382	226131	93
G5_ Sir_r1	411303	363875	88
G5_ Sir_r2	264122	233788	89

O tempo de processamento total do módulo da análise de qualidade foi em média de 59,4 segundos para as quatro bibliotecas, que possuem um total de 1.119.652 de *reads*. Entre todas as etapas internas, a que apresentou maior tempo de processamento foi a ferramenta Prinseq Lite (

Figura *3.4*) que retira as sequências que são maiores que 30pb, entretanto a ferramenta Prinseq Lite não apresenta a opção de paralelização por multiprocessamento, encontrado nas ferramentas, Trimmomatic e FastQC, que foram executadas com a opção de quatro processadores. Levando em consideração que as bibliotecas de entrada apresentam apenas 1% de uma biblioteca real, a utilização da ferramenta Prinseq Lite pode acarretar no atraso da análises devido a não paralelização da mesma.

Figura 3.4 – Comparativo interno das etapas da análise de qualidade, mostrando o tempo de execução em segundos de cada uma das etapas.



Legenda: FastQC_bfr determina a execução da ferramenta antes dos dados trimados. FastQC_aft define a execução da ferramenta depois dos dados trimados.

3.2.2. Mapeamento dos *reads* e quantificação

As bibliotecas de Catuaí G5_Cat_ r1 e G5_Cat_ r2, ao se alinharem na base de dados do miRBase, utilizando o Bowtie 1, obtiveram respectivamente 21% e 17% dos *reads* totais, enquanto que as bibliotecas de G5_Sir_ r1e G5_Sir_ r2, da cultivar Siriema, alinharam 16% e 20% dos *reads*. Em relação ao alinhamento dos *reads* não alinhados ao miRBase, contra a base de dados de precursores, a quantidade de *reads* alinhados diminuiu significativamente, G5_Cat_ r1 0,89% e 0,87% dos reads, G5_Sir_ r1 e G5_Sir_ r2 0,74% e 0,80% dos reads (Tabela 3.6).

Tabela 3.6 – Quantidade de *reads* alinhados através do Bowtie 1 contra a base de dados do mirBase e a base de dados de precursores.

Bowtie 1			
Biblioteca	**Reads**	**miRBase**	**Precursores**
G5_Cat_ r1	201845	38134	1290
G5_Cat_ r2	242382	38441	1631
G5_Sir_ r1	411303	58728	2269
G5_Sir_ r2	264122	46706	1498

No alinhador Bowtie 2, a primeira etapa contra a base de dados de plantas no miRBase, na biblioteca G5_Cat_r1, 19% dos *reads* foram alinhados, G5_Cat_r1 16% dos reads alinhados, G5_Sir_ r1 15% dos reads alinhados e na biblioteca G5_Sir_ r2 18% dos reads alinhados Os reads não alinhados contra a base de dados do miRBase, foram alinhados contra a base de dados dos precursores, apresentando 0,86 e 0,85% dos reads alinhados às bibliotecas G5_Cat_ r1 e G5_Cat_ r2, respectivamente. Já para as bibliotecas G5_Sir_ r1 e G5_Sir_ r1, 0,72% e 0,73% dos reads foram alinhados, respectivamente (Tabela 3.7).

Tabela 3.7 - Quantidade de *reads* alinhados segundo o Bowtie 2 ao mirBase e a base de dados do usuário.

Bowtie 2			
Biblioteca	**Reads**	**miRBase**	**Precursores**
G5_Cat_ r1	201845	38855	1413
G5_Cat_ r2	242382	39447	1738
G5_Sir_ r1	411303	60534	2530
G5_Sir_ r2	264122	48102	1592

Na Tabela 3.8, é possível visualizar a quantidade de *reads* alinhados contra a base de dados de plantas do miRBase e a base de precursores através do

algoritmo BWA. As bibliotecas da cultivar Catuaí (G5_Cat_ r1 e G5_Cat_ r2) obtiveram em média cada um 14% de *reads* alinhados contra o miRBase e, contra os precursores, por volta de 4% de *reads* alinhados. Já nas bibliotecas de Siriema, em média 12% dos *reads* foram alinhados contra o miRBase e 4% dos *reads* foram alinhados contra os precursores.

Tabela 3.8 – *Reads* alinhados segundo o BWA contra o mirBase e a base de dados do usuário.

	BWA		
Biblioteca	**Reads**	**miRBase**	**Precursores**
G5_Cat_ r1	201845	31364	7106
G5_Cat_ r2	242382	32240	6683
G5_Sir_ r1	411303	45423	13453
G5_Sir_ r2	264122	36839	9967

Bowtie 1 apresentou um maior tempo de execução 327 segundos para alinhar as quatro bibliotecas, entretanto, o Bowtie 2 apresentou maior eficiência comparando o tempo de execução e os *reads* alinhados, em relação as ferramentas Bowtie 1 e BWA (

Tabela *3.9*), demonstrando assim um resultado similar a outros estudos que também apresentaram o Bowtie 2 com o maior número de *reads* e contagem de *miRNAs* (TAM et al., 2015). O Bowtie 2 foi implementando utilizando o parâmetro de alinhamento local, que possibilita que *reads* possam ser aparados nas extremidades para otimizar a pontuação do alinhamento, o que pode ter possibilitado o maior número de *reads* alinhados.

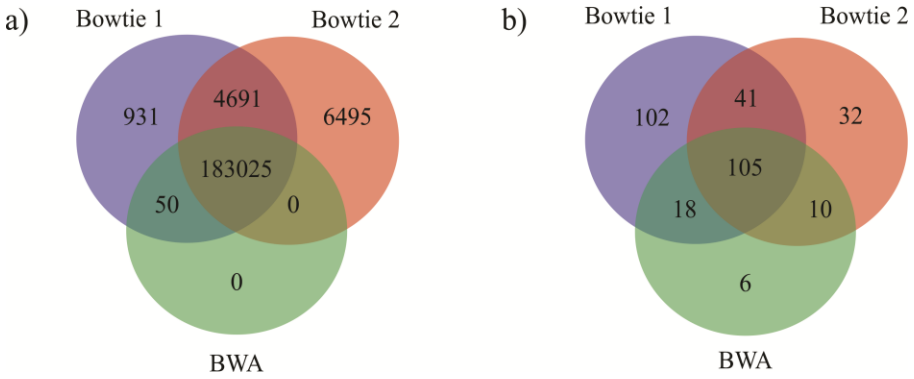
Tabela 3.9 – Tempo de execução em segundos, a quantidade de *reads* alinhados e a quantidade de reads maduros de cada alinhador, usando todas as quatro bibliotecas.

Alinhador	**Tempo (s)**	**Reads alinhados**	***miRNAs* maduros**
Bowtie 1	327	188697	266
Bowtie 2	91	194211	188
BWA	27,7	183075	139

Comparando os três alinhadores em relação aos *reads* alinhados das quatro bibliotecas, todos os *reads* alinhados pelo BWA foram mapeados também pelos outros alinhadores Bowtie 1 e Bowtie 2, como pode ser visto na Figura 3.5a. Mais de 97% do reads mapeados, foram mapeados nos três alinhadores, e 187.716 reads foram mapeados no Bowtie 1 e Bowtie 2 (Figura 3.5a). Em relação aos *miRNAs* quantificados, o alinhador Bowtie 1 não apresentou maior número de reads mapeados, porém, apresentou maior número de *miRNAs* alinhados e 102 *miRNAs* encontrados a partir dele, mostrando que os reads se alinharam com mais dispersão entre os *miRNAs* homólogos em relação ao Bowtie 2 (Figura 3.5b).

Figura 3.5 – **a)** Comparação entre os reads totais das quatro bibliotecas alinhados entre os três alinhamentos. **b)** Comparação entre os *miRNAs*

quantificados totais das quatro bibliotecas alinhados entre os três alinhamentos.



3.2.3. Diferença de expressão

Esta etapa foi implementada utilizando dois pacotes do R, edgeR e DEseq2, para cada ferramenta de alinhamento. Pela Tabela 3.10 é possível visualizar que o Bowtie 1, utilizando o edgeR, obteve 67% dos *miRNAs*. Já o Bowtie 2 obteve 65% dos *miRNAs*, e o BWA apresentou maior número de *miRNAs* diferencialmente expressos, aproximadamente 80% do total. O Bowtie 1 apresentou maior número de *miRNAs* diferencialmente expressos conhecidos e específicos Tabela 3.10.

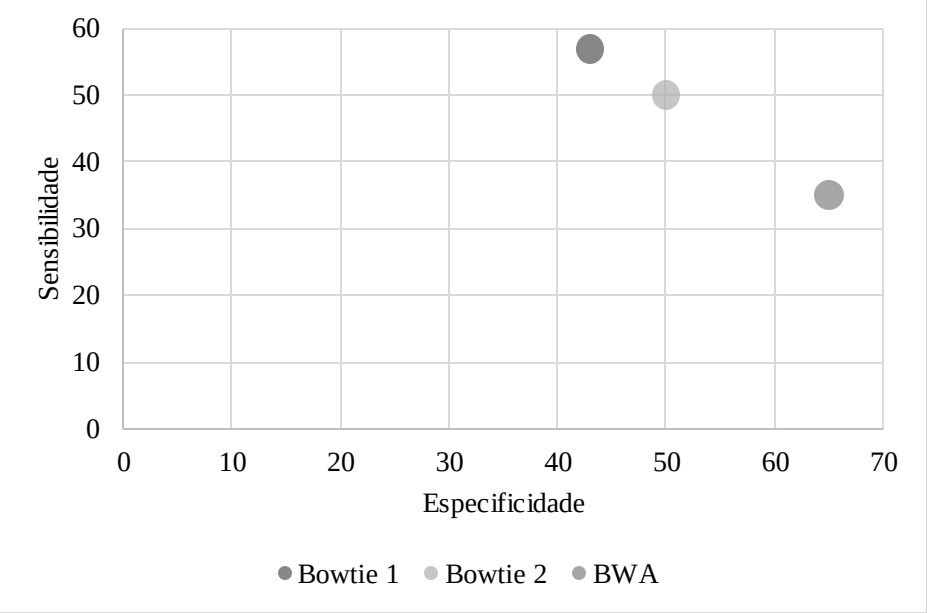
Tabela 3.10 – Comparação entre os microRNAs diferencialmente expressos pelo edgeR entre os três alinhamentos.

Alinhador	***miRNAs* maduros**	***miRNAs* DE**	***miRNAs* conhecidos DE**	***miRNAs* específicos DE**
Bowtie 1	266	179	102	77
Bowtie 2	188	123	62	61
BWA	139	111	39	72

Legenda: *miRNAs* conhecidos correspondem a microRNAs já caracterizados tanto no miRBase quanto no banco de precursores. *miRNAs* específicos correspondem a microRNAs não caraterizados na base de dados de precursores.

O Bowtie 1 apresentou maior sensibilidade (57%), que corresponde a razão entre *miRNAs* conhecidos e o total de *miRNAs* diferencialmente expressos, enquanto o alinhador BWA apresentou maior especificidade (65%), satisfaz a razão entre *miRNAs* específicos (não caracterizados) e total *miRNAs* DE. No entanto, o Bowtie 2 apresentou um equilíbrio entre sensibilidade e especificidade 50% (Figura 3.6).

Figura 3.6 – Comparativo entre os alinhamentos levando em consideração a sensibilidade (*miRNAs* conhecidos/*miRNAs* DE) e especificidade (*miRNAs* específicos /*miRNAs* DE) do edegR.



Através da análise dos resultados, foi possível identificar que o Bowtie 1 apresentou maior número de *miRNAs* maduros identificados e diferencialmente expressos pelo edgeR, segundo o seu manual ele também apresenta maior sensibilidade para o mapeamento com reads menores que 50pb ao comparado ao Bowtie 2 (LANGMEAD et al., 2009; LANGMEAD & SALZBERG, 2012),

entretanto, o seu resultado de processamento foi muito mais elevado ao ser comparado com o Bowtie 2 e BWA.

Usando o pacote DEseq2 para analisar os *miRNAs* diferencialmente expressos, somente o alinhador BWA apresentou apenas um *miRNAs* DE (Tabela 3.11), o *miR482*, que também foi identificado como DE pelo edgeR. É possível visualizar com esse comparativo, que o DEseq2 é restringente para detecção de diferença de expressão, assim como foi demonstrado por outros estudos de comparação entre métodos para cálculo e detecção de diferença de expressão (KHANG & LAU, 2015; RAJKUMAR et al., 2015). Além da restringência do DEseq2, a quantidade de detecção de *miRNAs* DE pode estar relacionada ao tamanho das bibliotecas usadas para teste, que correspondem a 1% do total de bibliotecas reais, diminuindo assim a possibilidade de detecção de microRNAs diferencialmente expressos.

Tabela 3.11 - Comparação entre os diferencialmente expressos pelo DEseq2 entre os três alinhamentos.

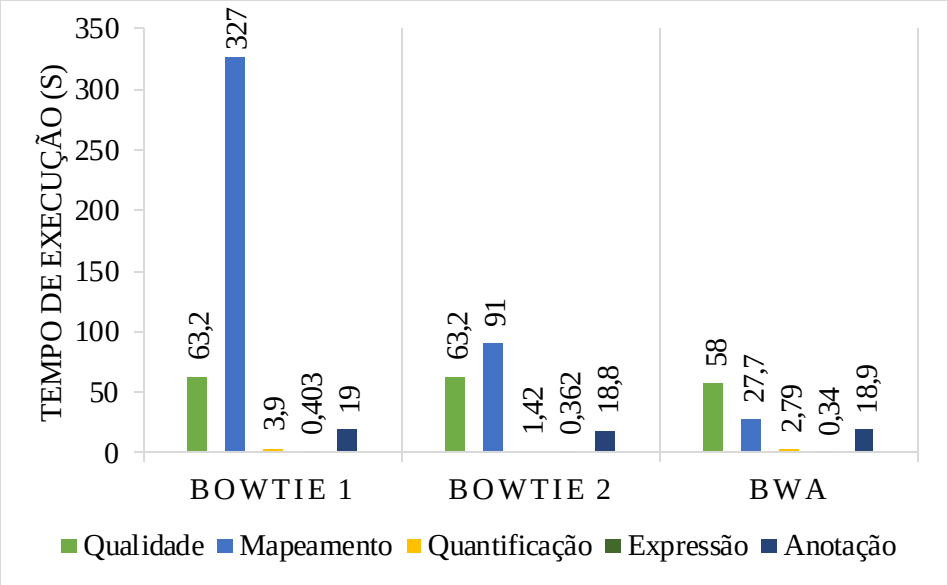
Alinhador	***miRNAs* maduros**	***miRNAs* DE**
Bowtie 1	266	0
Bowtie 2	188	0
BWA	139	1

Legenda: *miRNAs* maduros correspondem aos alinhados e *miRNAs* DE são os *miRNAs* detectados como diferencialmente expressos pelo DEseq2 .

3.3. Benckmark

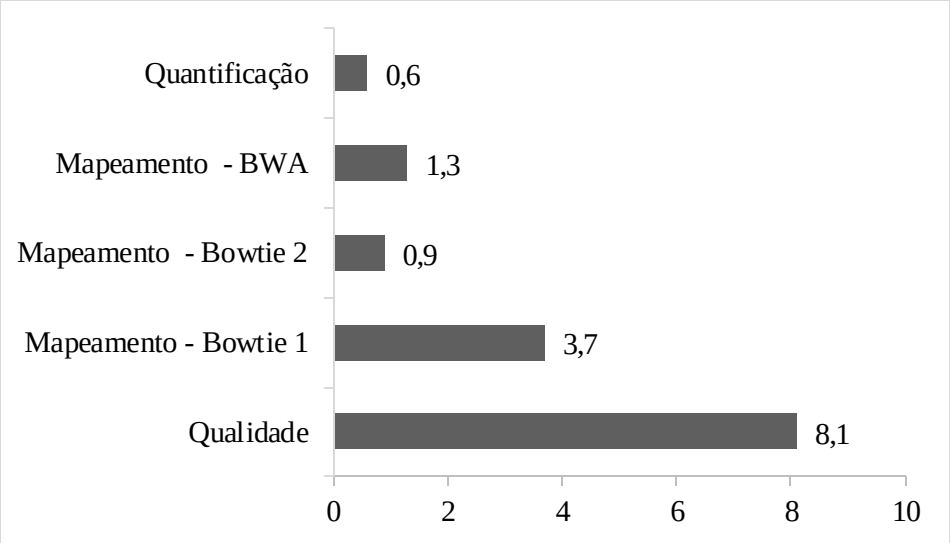
Os três alinhadores utilizados, Bowtie 1, Bowtie 2 e BWA foram implementados de forma separada, com cada um sendo implementado em um *script* diferente onde se tinha todos os módulos implementados, análise de qualidade, quantificação, expressão e *review*. Através da Figura 3.7 é possível visualizar que o maior tempo de execução gasto é com a etapa de mapeamento dos reads.

Figura 3.7 – Comparação entre o tempo de execução dos módulos entre os três alinhadores, Bowtie 1, Bowtie 2 e BWA.



As análises foram executadas em um computador com processador Intel(R) Core(TM) i5-4210U CPU 2.40GHz, em uma máquina de 8GB de memória RAM e 4 núcleos. A média da memória gasta pelos módulos estão representados na Figura 3.8, os outros módulos que não estão na figura, expressão e *review*, não apresentaram uso de memória suficiente para ser apresentada. A análise de qualidade é a que obteve maior gasto de memória, chegando a 8,1%, o que corresponde a 64,8MB dos 8GB que se tinha disponível na máquina utilizada. Esses 8,1% está relacionada com a utilização da ferramenta Trimmomatic, que além de memória, também utiliza dois dos quatro processadores alocados para análise. A pouca utilização de memória do *pipeline* se deu devido a preferência por implementação do código utilizando arquivos, ao oposto de utilizar estruturas de dados, como por exemplo, vetores, que são alocados na memória.

Figura 3.8 - Porcentagem de memória utilizada por cada método.



Legenda: Porcentagem considerando uma máquina de 8GB de memória RAM. Os módulos de expressão e *review* não apresentaram resultados significativos de uso de memória, logo não estão representados neste gráfico.

4. CONCLUSÃO

Este trabalho permitiu o desenvolvimento de um *pipeline* com o objetivo de analisar a expressão diferencial de bibliotecas de *miRNA-seq*, implementado em módulos: análise de qualidade, mapeamento de *reads*, quantificação de *miRNAs*, expressão diferencial e *review*. O módulo que apresentou maior tempo de execução foi o mapeamento, que requer maior processamento devido ao custo computacional gasto. Consequentemente o ajuste e a escolha dos métodos de mapeamento reflete consideravelmente no tempo de execução de todo *pipeline* e também no seu resultado final.

Para uma melhor performance e resultado final é recomendado a utilização do Bowtie 1 contra a base de dados de plantas do miRBase e a utilização do BWA contra a base de dados de entrada pelo usuário, considerando que essa base será específica ou próxima da espécie de estudo. É indicado considerar os microRNAs que apresentam ser diferencialmente expressos nos dois pacotes edgeR e DEseq2, trazendo maior confiabilidade na análise de expressão.

De modo geral, o *pipeline* apresentou o resultado esperado e concluiu seus objetivos para a detecção de perfil de expressão de microRNAs, demonstrou ser eficiente utilizando pouca memória, otimizado por apresentar um protocolo de análise *miRNAs* e robusto, demonstrando confiabilidade para análise de expressão diferencial em dados de *miRNA-seq*, corroborando para conhecimentos relacionados de processos fisiológicos vegetais.

REFERÊNCIAS

AN, J. et al. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. **Nucleic Acids Res,** v. 41, 2013.

AN, J. et al. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. **BMC Bioinformatics,** v. 15, n. 1, p. 275, 2014. ISSN 1471-2105.

ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. **Genome Biol,** v. 11, 2010.

ANDREWS, S. **FastQC:** A quality control tool for high throughput sequence data. p. 2010.

ATLASSIAN. **Bitbucket** 2008.

AUER, P. L.; DOERGE, R. W. Statistical Design and Analysis of RNA Sequencing Data. **Genetics,** v. 185, n. 2, p. 405-416, 2010.

AXTELL, M. J.; WESTHOLM, J. O.; LAI, E. C. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. **Genome Biology,** v. 12, n. 4, p. 221, 2011. ISSN 1474-760X.

BARTEL, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. **Cell,** v. 116, n. 2, p. 281-297, 2004. ISSN 0092-8674.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114-20, 2014. ISSN 1367-4811 (Electronic) 1367-4803 (Linking).

CONESA, A. et al. A survey of best practices for RNA-seq data analysis. **Genome Biol,** v. 17, n. 1, p. 13, 2016. ISSN 1474-760X (Electronic) 1474-7596 (Linking).

D'ANTONIO, M. et al. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. **BMC Genomics,** v. 16, n. 6, p. 1-11, 2015. ISSN 1471-2164.

DAVIS, M. P. A. et al. Kraken: A set of tools for quality control and analysis of high-throughput sequence data(). **Methods (San Diego, Calif.),** v. 63, n. 1, p. 41-49, 2013. ISSN 1046-2023 1095-9130.

DEBAT, H. J.; DUCASSE, D. A. Plant microRNAs: Recent Advances and Future Challenges. **Plant Molecular Biology Reporter,** v. 32, n. 6, p. 1257-1269, 2014. ISSN 0735-9640 1572-9818.

FRIEDLANDER, M. R. et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. **Nucleic Acids Res,** v. 40, 2012.

GARBER, M. et al. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nat Meth,** v. 8, n. 6, p. 469-477, 2011. ISSN 1548-7091.

GIT, A. et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. **Rna,** v. 16, 2010.

GRIFFITH, M. et al. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. **Plos Computational Biology,** v. 11, n. 8, p. 20, 2015. ISSN 1553-734X.

HODKINSON, B. P.; GRICE, E. A. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. **Advances in Wound Care,** v. 4, n. 1, p. 50-58, 2015. ISSN 2162-1918 2162-1934.

JOHNSON, B. K. et al. SPARTA: Simple Program for Automated reference-based bacterial RNA-seq Transcriptome Analysis. **Bmc Bioinformatics,** v. 17, p. 4, 2016. ISSN 1471-2105.

KHANG, T. F.; LAU, C. Y. Getting the most out of RNA-seq data analysis. **PeerJ,** v. 3, p. e1360, 2015.

KLAUS POHL, C. R. **Requirements Engineering Fundamentals: A Study Guide for the Certified Professional for Requirements Engineering Exam**. Rocky Nook; 1 edition (April 28, 2011), 2011. 184 ISBN 978-1933952819.

KOZOMARA, A.; GRIFFITHS-JONES, S. miRBase: annotating high confidence microRNAs using deep sequencing data. **Nucleic Acids Research,** v. 42, n. D1, p. D68-D73, 2013. ISSN 0305-1048.

KUKURBA, K. R.; MONTGOMERY, S. B. RNA Sequencing and Analysis. **Cold Spring Harb Protoc,** v. 2015, n. 11, p. pdb top084970, 2015. ISSN 1559-6095 (Electronic) 1559-6095 (Linking).

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Meth,** v. 9, n. 4, p. 357-359, 2012. ISSN 1548-7091.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology,** v. 10, n. 3, p. 1-10, 2009. ISSN 1474-760X.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics,** v. 25, n. 14, p. 1754-1760, 2009.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, 2009. ISSN 1367-4803 1460-2059.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology,** v. 15, n. 12, p. 550, 2014. ISSN 1465-6906 1465-6914.

METZKER, M. L. Sequencing technologies [mdash] the next generation. **Nat Rev Genet,** v. 11, n. 1, p. 31-46, 2010. ISSN 1471-0056.

MINIANDRÉS-LEÓN, E.; NÚÑEZ-TORRES, R.; ROJAS, A. M. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. **Scientific Reports,** v. 6, p. 25749, 2016.

MORAIS, H. C., PAULO HENRIQUE; KOGUISHI, MIRIAN SEI AND RIBEIRO, ANA MARIA DE ARRUDA. Escala fenológica detalhada da fase reprodutiva de Cooffea arabica. **Bragantia,** v. 67, p. 257-260, 2008. ISSN 0006-8705.

MOREIRA, L. M. **Ciências genômicas : fundamentos e aplicações**. Ribeirão Preto: Sociedade Brasileira de Genética, 2015. 403 ISBN 978-85-89265-22-5.

MUIR, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. **Genome Biology,** v. 17, n. 1, 2016. ISSN 1474-760X.

NEVADO, B.; PEREZ-ENCISO, M. Pipeliner: software to evaluate the performance of bioinformatics pipelines for next-generation resequencing. **Molecular Ecology Resources,** v. 15, n. 1, p. 99-106, 2015. ISSN 1755-098X.

NIELSEN, R. et al. Genotype and SNP calling from next-generation sequencing data. **Nat Rev Genet,** v. 12, n. 6, p. 443-451, 2011. ISSN 1471-0056.

P., G. M. **Statistics::R** 2004.

RAJKUMAR, A. P. et al. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. **BMC Genomics,** v. 16, p. 548, 2015. ISSN 1471-2164 (Electronic) 1471-2164 (Linking).

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics,** v. 26, n. 1, p. 139-140, 2010. ISSN 1367-4803 1367-4811.

SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics,** v. 27, n. 6, p. 863-864, 2011. ISSN 1367-4803 1367-4811.

SCHORDERET, P. NEAT: a framework for building fully automated NGS pipelines and analyses. **Bmc Bioinformatics,** v. 17, 2016.

SHENDURE, J. The beginning of the end for microarrays? **Nat Meth,** v. 5, n. 7, p. 585-587, 2008. ISSN 1548-7091.

SOMMERVILLE, I. **Engenharia de software**. 9. São Paulo: 2003.

STAJICH, J. E. et al. The Bioperl Toolkit: Perl Modules for the Life Sciences. **Genome Research,** v. 12, n. 10, p. 1611-1618, 2002. ISSN 1088-9051.

SUN, Z. et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. **BMC Genomics,** v. 15, n. 1, p. 423, 2014. ISSN 1471-2164.

SUNKAR, R. MicroRNAs with macro-effects on plant stress responses. **Seminars in Cell & Developmental Biology,** v. 21, n. 8, p. 805-811, 2010. ISSN 1084-9521.

TAM, S.; TSAO, M. S.; MCPHERSON, J. D. Optimization of miRNA-seq data preprocessing. **Brief Bioinform,** v. 16, n. 6, p. 950-63, 2015. ISSN 1477-4054 (Electronic) 1467-5463 (Linking).

TEAM, C. **R: A language and environment for statistical computing**: R Foundation for Statistical Computing 2009.

TOM CHRISTIANSEN, B. D. F., LARRY WALL, JON ORWANT. **Programming Perl**. O'Reilly Media, 2012. ISBN 978-0-596-00492-7.

WANG, W.-C. et al. miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression. **BMC Bioinformatics,** v. 10, n. 1, p. 328, 2009. ISSN 1471-2105.

WEN, M. et al. miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. **BMC Bioinformatics,** v. 13, p. 140-140, 2012. ISSN 1471-2105.

YANG, X.; LI, L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. **Bioinformatics,** v. 27, n. 18, p. 2614-2615, 2011.

CAPÍTULO 3
Considerações finais e perspectivas

RESUMO

O *pipeline* de análise de expressão diferencial em *miRNAs* no geral conseguiu concluir seus objetivos, porém ainda pode ser melhorado em vários aspectos. A utilização de dois pacotes para cálculos de expressão gênica não apresentou ser suficiente, logo, é necessário o acréscimo de um terceiro como critério de desempate. Em relações a outras etapas, o *pipeline* deverá fazer o tratamento de *isomiRs*, variantes de um mesmo *miRNA*, dos quais podem afetar significativamente a etapa de cálculo de quantificação, destacando falsos positivos na etapa de diferença de expressão. Para a validação mais aprofundada outras análises devem ser consideradas: comparação com outras plataformas integrativas que analisam dados de *miRNA-seq* com o objetivo de analisar diferença de expressão e validação com bibliotecas de tamanhos reais de sequenciamento em larga escala de dados de plantas. Finalmente ao finalizar o desenvolvimento e melhorias do mesmo, deve se atentar-se para uma documentação simples e de fácil entendimento, para que o usuário não deixe de utilizar o mesmo.

Palavras-chave: *isomiR*. Validação. Documentação.

ABSTRACT

The pipeline for differential expression analysis in *miRNAs* was able to achieve its aims, but it can be improved in several aspects. The use of two packages for the calculation of differential expression seemed not to be enough, hence, a third package must be added. Regarding the other steps, the pipeline should treat the isomiRs, variants of the same miRNA, which can affect greatly the calculation step, leading to false positives at the differential expression step. For a deeper validation, other analyses might be considered: comparison with other integrative platforms that analyze miRNA-seq data, aiming at analyzing real sized libraries from high throughput data of plants. At last, to finish the improvement of the pipeline, a simple and comprehensive documentation should be presented to prevent the not usage by the users.

Keywords: isomiR. Validation. Documentation

1. CONSIDERAÇÕES FINAIS E PERSPECTIVAS

De forma geral o *pipeline* cumpriu com os objetivos do trabalho, porém, várias melhorias foram observadas durante o desenvolvimento do mesmo. O pacote DESeq é apresentado em vários estudos como um dos pacotes que possui uma menor sensibilidade para a detecção de transcritos diferencialmente expressos, já o edgeR apresentou maior sensibilidade (KHANG & LAU, 2015). É difícil a confirmação de um transcrito que está diferencialmente expresso em apenas um pacote, desta forma, é importante para o presente projeto possuir pelo menos três ferramentas para critério de desempate de transcritos que são considerados DE em somente uma opção, um exemplo seria a utilização da ferramenta Cuffdiff 2 (TRAPNELL et al., 2013), por apresentar em muitos estudos uma detecção mais sensível que o edgeR para transcritos diferencialmente expressos (RAJKUMAR et al., 2015).

O uso de sequenciamento em larga escala aprimorou as análises de *miRNAs* devido ao seu alto grau de sensibilidade em termos de detecção, porém, *miRNAs* gerados pelo mesmo gene apresentam variações no comprimento da sequência, adição ou deleção de um ou mais nucleotídeos na região do 5' ou 3' (MPATH & DIBB, 2015), as quais essas variações são chamadas de *isomiRs* (BLOW et al., 2006). Tais variantes de um mesmo miRNA pode afetar, significativamente, as análises de expressão gênica, sendo importante a implementação, posteriormente no *pipeline*, de um algoritmo como o CPSS (ZHANG et al., 2012), ou até utilizar um banco de dados, como IsomiR Bank (ZHANG et al., 2016), para tratar a presença de *isomiRs*, muitas vezes considerados artefatos de sequenciamento.

Ressalta-se que, embora o *pipeline* proposto para a análise de diferença de expressão em *miRNA-seq* apresentar o resultado final desejados para as bibliotecas aqui utilizadas, ele ainda deve ser testado utilizando bibliotecas de

sequenciamento em larga escala de tamanho real. Além de ser necessário comparar o seu resultado final com outras ferramentas da literatura que são comumente utilizadas, como o miRDeep-P (YANG & LI, 2011) e o mirExpress (WANG et al., 2009), validando assim o *pipeline* desenvolvido. É necessário, ainda, que o pipeMIRSEQ seja validado através de bibliotecas com tamanhos reais, e que a etapa de teste com o usuário utilizando *pipeline* não seja descartada. Para teste com usuário, é recomendado utilizar o grupo de pessoas que participou do *brainstorming* na fase de análise de requisitos com seus dados empregados na ferramenta, atentando para suas funcionalidades, usabilidade e resultado final das análises.

Ainda sobre as melhorias referentes a implementação do *pipeline*, deve se considerar que com a utilização dos dados reais de *miRNA-seq* o tempo de execução aumentará, no entanto, ainda pode-se melhorar o tempo de execução utilizando ferramentas ou códigos multiprocessados. Ao finalizar o desenvolvimento e todas as etapas de validação é importante ter uma documentação completa do *software* desenvolvido. Visto que, muitos usuários, quando deparados com falta de documentação ou documentação inadequada, facilmente desistem de utilizar os *softwares*, reduzindo assim o impacto do mesmo para com a sociedade (KARIMZADEH & HOFFMAN, 2017). Por fim, o código-fonte pronto será registrado no Instituto Nacional de Propriedade Industrial (INPI), garantindo maior segurança, porém, ele ainda será disponibilizado para a utilização livre.

REFERÊNCIAS

BLOW, M. J. et al. RNA editing of human microRNAs. **Genome Biol,** v. 7, 2006.

KARIMZADEH, M.; HOFFMAN, M. M. Top considerations for creating bioinformatics software documentation. **BriefBioinform**, 2017. ISSN 1477-4054 (Electronic) 1467-5463 (Linking).

KHANG, T. F.; LAU, C. Y. Getting the most out of RNA-seq data analysis. **PeerJ,** v. 3, p. e1360, 2015.

MPATH, G. C. T.; DIBB, N. IsomiRs have functional importance. **Malaysian J Pathol,** v. 37, n. 2, p. 73-81, 2015.

RAJKUMAR, A. P. et al. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. **BMC Genomics,** v. 16, p. 548, 2015. ISSN 1471-2164 (Electronic) 1471-2164 (Linking).

TRAPNELL, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. **Nat Biotech,** v. 31, n. 1, p. 46-53, 2013. ISSN 1087-0156.

WANG, W.-C. et al. miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression. **BMC Bioinformatics,** v. 10, n. 1, p. 328, 2009. ISSN 1471-2105.

YANG, X.; LI, L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. **Bioinformatics,** v. 27, n. 18, p. 2614-2615, 2011.

ZHANG, Y. et al. CPSS: a computational platform for the analysis of small RNA deep sequencing data. **Bioinformatics,** v. 28, 2012.

ZHANG, Y. et al. IsomiR Bank: a research resource for tracking IsomiRs. **Bioinformatics,** v. 32, n. 13, p. 2069-2071, 2016. ISSN 1367-4803.