Sequenciamento de RNA em larga escala como ferramenta para identificação e caracterização de genes em culturas de importância agronômica

David Gabriel dos Santos Fagundes[1], Alexandro Cagliari[2]

[1] Núcleo de Bioinformática e Biotecnologia (NBB). Universidade Estadual do Rio Grande do Sul (UERGS).
E-mail: david-fagundes@uergs.edu.br
[2] Unidade em Santa Cruz do Sul. Universidade Estadual do Rio Grande do Sul (UERGS).
E-mail: alexandro-cagliari@uergs.edu.br

Submetido em: 2 ago. 2019. Aceito: 14 out. 2019.
DOI: http://dx.doi.org/10.21674/2448-0479.53.271-279

Resumo

A transcritômica permite catalogar todas as diferentes classes de transcritos presentes nas células, possibilitando a quantificação dos níveis de expressão variáveis de cada transcrito durante o processo de desenvolvimento e sob diferentes condições fisiológicas. À tecnologia que se utiliza do sequenciamento de nova geração para analisar o transcritoma dá-se o nome de Sequenciamento de RNA (RNA-Seq). Na agricultura, o RNA-Seq permite, através do estudo das mudanças nos níveis de expressão gênica, explicar os efeitos biológicos causados por alterações ambientais, ocorridas quando uma perturbação externa é inserida no sistema, como por exemplo, uma infecção por um patógeno ou parasita, mudanças nutricionais, restrição hídrica e outros tipos de estresses que as plantas podem sofrer. Elucidando as alterações nos níveis de expressão gênica é possível compreender melhor a relação entre os genes e seus produtos. A descoberta e o estudo de genes envolvidos em características fenotípicas economicamente importantes podem contribuir para o fornecimento de matéria-prima para programas de melhoramento genético em culturas de importância agronômica.

Palavras-chave: RNA-Seq. Transcritômica. Melhoramento Vegetal.

Abstract

Large-scale RNA sequencing as a tool for identification and characterization of genes in crops of agronomic importance

The transcription allows to catalog all the different classes of transcripts present in cells, making possible the quantification of the variable levels of expression of each transcript during the development process and under different physiological conditions. The technology that uses new generation sequencing to analyze the transcritoma is called RNA Sequencing (RNA-Seq). In agriculture, RNA-Seq allows, through the study of changes in the levels of gene expression, to explain the biological effects caused by environmental changes that occur when an external disturbance is inserted into the system, such as infection by a pathogen or parasite, nutritional changes, water restriction and other types of stresses that plants may suffer. By elucidating the changes in gene expression levels it is possible to better understand the relationship between genes and their products. The discovery and study of genes involved in economically important phenotypic characteristics can contribute to supply raw material for breeding programs in crops of agronomic importance.

Keywords: RNA-Seq. Transcriptomic. Plant Breeding.

Introdução

Transcritoma é o conjunto completo de transcritos em uma célula em um determinado estágio de desenvolvimento ou condição fisiológica. A transcritômica busca determinar a estrutura transcricional dos genes e quantificar os níveis de expressão de cada transcrito durante o desenvolvimento e sob diferentes condições ambientais ou fisiológicas (LINDBERG; LUNDEBERG, 2010). Entender o transcritoma é essencial para interpretar os elementos funcionais do genoma e compreender o desenvolvimento dos organismos vivos (WANG; GERSTEIN; SNYDER, 2009).

A transcritômica busca catalogar todas as classes de transcritos, incluindo RNA mensageiro codificador de proteínas (mRNA) e RNA não codificador [ncRNA: RNA ribossômico (rRNA), RNA transportador (tRNA), pequenos RNAs (miRNAs, siRNA, snRNA, etc) dentre outros].

Sequenciamento de próxima geração (NGS), sequenciamento paralelo massivo ou sequenciamento profundo são termos sinônimos que descrevem uma tecnologia de sequenciamento de DNA que revolucionou a pesquisa genômica. Através das plataformas de NGS é possível executar o sequenciamento de milhões de pequenos fragmentos de DNA em paralelo (BEHJATI; TARPEY, 2013).

O uso do NGS não se restringe apenas ao estudo do genoma estático. Essa tecnologia pode ser explorada para analisar o transcritoma dinâmico, fornecendo um novo método para mapeamento e quantificação de transcritomas, denominado sequenciamento de RNA (RNA-Seq) (QIAN *et al.*, 2014).

Enquanto o genoma é relativamente estável, o transcritoma varia com o estágio de desenvolvimento, condição fisiológica e ambiente externo. Nesse contexto, RNA-Seq pode ser usado para construir um mapa completo do transcritoma em cada uma dessas condições (QIAN *et al*., 2014).

Criar um mapa de todos os genes, juntamente com suas isoformas alternativas e sua expressão em diversos tipos de células, é fundamental para compreender o metabolismo de uma célula. Até recentemente, a produção desses tipos de dados era muito cara e trabalhosa. Devido ao alto custo e ao limitado rendimento, as técnicas tradicionais forneciam apenas um vislumbre da verdadeira complexidade do transcritoma estudado. A análise desses dados exigia ferramentas computacionais sofisticadas, muitas das quais forneceram a base para os programas usados hoje para análise de dados de RNA-Seq (GARBER *et al*., 2011)

A tecnologia de RNA-Seq utiliza poderosas ferramentas computacionais e técnicas de bioinformática para o tratamento e estudo dos dados gerados experimentalmente (GARBER *et al*, 2011). As aplicações do RNA-Seq visam resolver problemas biológicos específicos, incluindo a quantificação de *splicing* alternativo, descoberta de novos genes relacionados a câncer, melhoria da montagem do genoma, quantificação da expressão gênica, descoberta de novos transcritos, detecção de polimorfismos de nucleotídeo único (*Single Nucleotide Polimorphism - SNPs*), edição de RNA, detecção de fusão gênica, entre outras (GARBER *et al.*, 2011; CONESA *et al*., 2016).

A tecnologia de RNA-Seq já possibilitou descobertas importantes em vários campos da ciência, melhorando a compreensão da prevalência e significado funcional de RNAs não-codificantes (GRIFFITH *et al*., 2015).

Na agricultura, o RNA-Seq pode permitir, através do estudo das mudanças nos níveis de expressão gênica, explicar os efeitos biológicos causados por alterações ambientais, ocorridas quando uma perturbação externa é inserida no sistema. Estas perturbações incluem, por exemplo, uma infecção por um patógeno ou parasita, mudanças nutricionais, restrição hídrica e outros tipos de estresses que as plantas podem sofrer (GIACHETTO; HIGA, 2014).

O estudo das alterações nos níveis de expressão gênica permite compreender o comportamento das plantas nas diferentes fases de desenvolvimento e em diferentes ambientes. Além disso, estudos transcritômicos podem facilitar a descoberta de genes envolvidos em características fenotípicas economicamente importantes em espécies vegetais de importância agronômica (GIACHETTO; HIGA, 2014; MARTIN *et al*., 2013).

A identificação de transcritos e a quantificação da expressão gênica eram, tradicionalmente, análises separadas na biologia molecular. O advento da técnica de RNA-seq passou a permitir que a descoberta e a quantificação possam ser combinadas em um único ensaio de sequenciamento (CONESA *et al.*, 2016).

Comparado com as tecnologias de microarranjos, baseadas em hibridização, que vinham sendo a abordagem dominante para estudar a expressão gênica, a tecnologia de RNA-Seq oferece várias vantagens, incluindo uma maior faixa de níveis de expressão, maior sensibilidade na detecção da expressão de alelos específicos, promotores e isoformas, menor ruído de processamento e maior rendimento (WANG *et al.*, 2009)

Entre as principais vantagens da técnica de RNA-Seq podemos citar: (i) permitir uma medição mais precisa dos níveis de transcritos e suas isoformas, (ii) apresentar potencial para o desenvolvimento de *SNPs* usados para detectar expressão específica de alelo, (iii) capacidade de identificar leituras com modificações pós-transcricionais ou sequências rearranjadas, (iv) permitir a identificação de genes específicos de determinadas espécies e (v) não requer conhecimento prévio do transcritoma em consideração (ROBLES *et al*., 2012).

O RNA-Seq demonstrou ser altamente preciso para quantificar os níveis de expressão, além de demonstrar altos níveis de reprodutibilidade, tanto para replicatas técnicas quanto biológicas. A eficiência, a resolução, a reprodutibilidade do RNA-Seq como ferramenta para criar perfis de expressão diferencial levaram muitos cientistas a abandonar os microarranjos em favor desta nova tecnologia (ROBLES *et al*., 2012).

Tendo em vista o grande potencial de uso da tecnologia de RNA-Seq como ferramenta para identificação e caracterização de genes em culturas de importância agronômica, o presente trabalho tem como objetivo apresentar uma revisão bibliográfica sobre as principais estratégias para o delineamento experimental, montagem e análise um experimento de RNA-Seq, bem como apresentar exemplos de trabalhos que utilizaram tal tecnologia para estudos em plantas de interesse agronômico. Para tanto, foi realizada uma ampla revisão bibliográfica em artigos científicos, publicados à partir do ano de 2010, em periódicos nacionais e internacionais disponíveis na base de dados Pubmed (www.ncbi.nlm.nih.gov/pubmed).

Revisão Bibliográfica

Preparo da biblioteca de cDNA

Um bom desenho experimental consiste em escolher o tipo de biblioteca, a profundidade de sequenciamento e número de réplicas apropriadas para o sistema biológico em estudo, além de assegurar que a aquisição dos dados não esteja contaminada com vieses desnecessários (CONESA *et al*., 2016).

Ainda não há um *pipeline* padrão para a variedade de diferentes aplicações e cenários de análise nos quais o RNA-Seq pode ser usado. Os experimentos são planejados e adotam diferentes estratégias de análise, dependendo do organismo e dos objetivos da pesquisa (CONESA *et al*., 2016).

Quando o organismo estudado já possui o genoma e transcritoma sequenciados, a análise de RNA-Seq normalmente envolverá o mapeamento das leituras contra esse genoma ou transcritoma de referência para inferir quais transcritos são expressos. Quando apenas o transcritoma está disponível, o mapeamento impede a descoberta de novos transcritos não anotados e foca a análise apenas na quantificação. Neste caso, o caminho de análise é primeiro montar as leituras em *contigs* mais longos e depois tratar esses *contigs* como um transcritoma expresso para o qual as leituras serão mapeadas novamente para quantificação. Em ambos os casos, a cobertura de leitura pode ser usada para quantificar o nível de expressão do transcrito (CONESA *et al*., 2016).

A estratégia de montagem de transcritoma baseada em um genoma de referência tem várias vantagens. Primeiramente, ela é computacionalmente mais econômica. Além disso, os artefatos de contaminação ou sequenciamento não são uma preocupação importante para essa estratégia, porque não se espera que eles se alinhem com o genoma de referência. Como a sequência do genoma já é conhecida, pequenas lacunas dentro do transcrito que foram causadas pela falta de cobertura de leitura podem ser preenchidas usando essa sequência de referência. Além disso, esta estratégia baseada em referência é muito sensível e pode revelar transcritos de baixa abundância (MARTIN; WANG, 2011).

No desenho experimental de extração de RNA é importante utilizar-se de estratégias de remoção do RNA ribossômico (rRNA) que é altamente abundante. Do total de RNA contido na célula, 90% é rRNA enquanto apenas 1–2% é RNA mensageiro (mRNA), o qual estamos interessados. Em eucariotos a remoção do rRNA pode ser feita principalmente por duas vias, enriquecendo o mRNA usando seleção de poli(A) ou a depleção do rRNA (CONESA *et al*., 2016).

A preparação da biblioteca é um passo fundamental para o RNA-Seq, pois determina quão precisamente os dados de sequenciamento refletem o transcritoma original (QIAN *et al*., 2014).

Para a realização do RNA-seq, primeiramente o RNA é extraído e tratado com DNAse para eliminar qualquer possível contaminação da amostra com DNA (MARTIN; WANG, 2011). Posteriormente, uma biblioteca de cDNA (DNA complementar) é preparada via fragmentação do mRNA (hidrólise ou nebulização) e posterior transcrição reversa (WANG *et al*., 2009). O objetivo da fragmentação é atingir o tamanho desejado de fragmentos para serem utilizados nas tecnologias de NGS (QIAN *et al*., 2014).

A biblioteca de cDNA é então sequenciada por sequenciadores de nova geração gerando milhões à bilhões de leituras curtas (*reads*) a partir de uma ou de ambas as extremidades dos fragmentos de cDNA (MARTIN; WANG, 2011).

No que diz respeito ao sequenciamento em si, importantes decisões de projeto experimental incluem o número de replicatas técnicas e/ou biológicas a serem usadas e a escolha da profundidade do sequenciamento (ROBLES *et al*., 2012). Embora a questão de quantas réplicas são necessárias ainda esteja aberta, em geral, quanto mais repetições melhor. Com os kits atualmente disponíveis, o sequenciamento de cada condição em triplicata já é muito viável (TRAPNELL *et al*., 2012).

O grau de variação técnica presente nesses conjuntos de dados parece originar-se principalmente na fase de preparação da biblioteca. A replicação biológica pode contrabalançar a variação técnica aleatória como parte da preparação de amostras independentes. Já foi demonstrado, por exemplo, que o poder de detecção de transcritos diferencialmente expressos melhora quando o número de réplicas biológicas aumenta de 2 para 5 (ROBLES *et al.*, 2012; YOON; NAM, 2017).

Outra opção de design também importante para a precisão dos dados é que os fragmentos de biblioteca podem ser sequenciados a partir de uma ou de ambas as extremidades. Embora as leituras de finalização pareada possam custar até duas vezes o custo das leituras de extremidade única, recomenda-se fortemente o sequenciamento emparelhado sempre que possível. Além disso, o comprimento da leitura também é uma consideração importante. Leituras mais longas (maiores que 75 pb) são geralmente preferíveis às curtas. No entanto, leituras mais longas podem aumentar substancialmente o custo de um experimento de RNA-Seq. Portanto, muitos pesquisadores preferem sequenciar mais amostras (ou mais réplicas das mesmas amostras) com leituras mais curtas (TRAPNELL *et al*., 2012).

Leituras mais curtas geralmente são suficientes para estudos de níveis de expressão gênica em organismos com genoma bem anotados, enquanto leituras mais longas são preferíveis para caracterizar transcritomas pouco anotados (CONESA *et al*., 2016).

Uma característica distinta das moléculas de RNA que afeta a análise é que elas ocorrem em uma ampla gama de tamanhos. RNAs muito pequenos (<100pb), como microRNAs (miRNA), devem geralmente ser capturados e sequenciados por uma estratégia independente, pois as estratégias de seleção de tamanho acabam os excluindo da análise geral (GRIFFITH *et al*., 2015).

Parâmetros de sequenciamento em larga escala

Em geral, qualquer tecnologia de sequenciamento de nova geração pode ser usada para o RNA-Seq.

As plataformas NGS geram milhões de sequências curtas, denominadas leituras ou *reads*, cujo comprimento varia de 25 a 450 pb, dependendo do tipo de plataforma NGS utilizado (QIAN *et al*., 2014).

Mesmo pequenos experimentos de RNA-Seq, envolvendo apenas uma única amostra, podem produzir enormes volumes de sequenciamento bruto (TRAPNELL *et al*., 2012).

A profundidade de sequenciamento é a cobertura média esperada em todos os *loci* ao longo da(s) sequência(s) alvo. Sem o benefício de estudos prévios, na maioria dos casos, antes da geração de dados é difícil estimar a profundidade ideal de sequenciamento ou a quantidade de dados necessários para alimentar adequadamente a detecção de expressão diferencial no transcritoma de interesse (ROBLES *et al*., 2012).

De uma maneira geral, a profundidade de sequenciamento de RNA-Seq é escolhida baseando-se em uma estimativa do comprimento total do transcritoma e na faixa dinâmica esperada de abundância dos transcritos. Mas como o transcritoma é dinâmico, a adequação dessas estimativas pode variar bastante entre organismos, tecidos, fase fisiológica e outros contextos biológicos. Com isso, transcritos com níveis de expressão baixos a moderados permanecerão difíceis de quantificar com boa precisão usando os atuais protocolos de RNA-Seq, mesmo em maiores profundidades de leitura (ROBLES *et al*., 2012).

Mapeamento de *reads*

Após obter *reads* de alta qualidade, a primeira tarefa da análise de dados é mapear essas *reads* em um genoma de referência (quando houver), ou montá-las em *contigs* antes de alinhar ao genoma para revelar a estrutura dos transcritos. No entanto, nesse conjunto de *reads*, existem leituras que abrangem junções exon-exon ou que contêm extremidades poli (A), que não podem ser analisadas da mesma maneira. As caudas poli(A) são identificadas simplesmente pela presença de múltiplos A's ou T's no final da leitura. Já as junções

exon-exon podem ser identificadas pela presença de sequências específicas, geralmente dinucleotídeos GT--AG que flanqueiam os locais de *splicing*, e confirmadas pela baixa expressão de sequências intrônicas, que são removidas durante o *splicing* (WANG *et al.*, 2009).

No processo de pré-processamento das *reads* também é realizada a remoção das leituras e artefatos de baixa qualidade, como sequências de adaptadores, DNA contaminante e duplicatas de PCR (MARTIN; WANG, 2011).

Um indicador global da precisão geral do sequenciamento e da presença de DNA contaminante é a porcentagem de leituras mapeadas no genoma/transcritoma de referência. Outros parâmetros importantes são a uniformidade da cobertura de leitura nos exons e na cadeia mapeada. Se as leituras se acumulam principalmente na extremidade 3' dos transcritos em amostras selecionadas com poli(A), isso pode indicar baixa qualidade de RNA no material de partida (CONESA *et al*., 2016).

A filtragem de genes expressos em níveis baixos antes da análise de expressão diferencial melhora o poder de detecção da expressão diferencial. O aumento da profundidade de sequenciamento também pode melhorar o poder estatístico para genes de baixa expressão, e para qualquer amostra existe um nível de sequenciamento no qual a melhoria de poder é alcançada aumentando o número de réplicas (CONESA *et al.*, 2016).

Os processos de controle de qualidade devem ser aplicados em todos os estágios da análise para garantir a reprodutibilidade e a confiabilidade dos resultados (CONESA *et al.*, 2016).

Análises de bioinformática

Os experimentos de RNA-Seq são analisados com algoritmos robustos, eficientes e baseados em métodos de estatística poderosos. Ferramentas de análise de RNA-Seq geralmente são subdivididas em três categorias, de acordo com seus objetivos específicos: (i) alinhamento de leitura; (ii) montagem de transcritos ou anotação do genoma; e (iii) transcrição e quantificação genética (TRAPNELL *et al*., 2012).

A análise de RNA-Seq é uma das tarefas mais exigentes computacionalmente na bioinformática. A análise de grandes conjuntos de dados requer uma estação de trabalho ou um servidor poderoso com amplo espaço em disco (TRAPNELL *et al*., 2012).

Geralmente os algoritmos e programas utilizados nas análises de RNA-Seq se baseiam em linhas de comando. No entanto, já existem produtos comerciais e interfaces de código aberto para ferramentas de análise de RNA-Seq, que são mais amigáveis para o usuário. O Galaxy Project, por exemplo, usa uma interface web e recursos de computação em nuvem para levar ferramentas direcionadas por linha de comando a usuários sem habilidades UNIX (TRAPNELL *et al*., 2012). Essas ferramentas baseadas em interface web facilitam conversões de formato e a extração de resultados relevantes (CONESA *et al*., 2016).

Análise de expressão diferencial

A expressão gênica é um processo amplamente estudado e uma das principais áreas de foco para a genômica funcional. A expressão gênica está relacionada ao fluxo de informações genéticas do modelo de DNA genômico para produtos proteicos funcionais. O RNA-seq tornou-se um ensaio de expressão gênica padrão, particularmente para analisar a abundância e diversidade de transcritos relativos (GRIFFITH *et al.*, 2015).

A identificação de transcritos e a quantificação da expressão gênica eram atividades separadas na biologia molecular desde a descoberta do papel do RNA como intermediário-chave entre o genoma e o proteoma. O poder do RNA-seq reside no fato de que descoberta e quantificação podem ser combinados em um único ensaio de sequenciamento de alto rendimento (CONESA *et al*., 2016).

A análise de RNA-Seq permite não apenas quantificar os níveis de expressão gênica dentro de uma única amostra de RNA, mas também detectar a expressão diferencial em diferentes tratamentos ou condições (KVAM *et al*., 2012; OSHLACK *et al*., 2010; QIAN *et al.*, 2014).

Ao mapear as *reads* geradas no sequenciamento no genoma ou transcritoma de referência, os níveis de expressão dos genes relativos à condição de interesse ou níveis absolutos podem ser quantificados através da análise diferencial (KOGENARU *et al.*, 2012).

Para a análise de expressão diferencial dos dados de RNA-Seq, a normalização deve ser realizada para ajustar as diferenças entre as amostras, como tamanho da biblioteca e características específicas do gene dentro da amostra quanto ao conteúdo GC e comprimento do gene (KVAM *et al*., 2012; QIAN *et al*., 2014).

Existem duas fontes principais de variabilidade sistemática que requerem normalização. Primeiro, a

fragmentação do RNA durante a construção da biblioteca faz com que transcritos mais longos gerem mais leituras em comparação com transcritos mais curtos presentes na mesma abundância na amostra. Segundo, a variabilidade no número de leituras produzidas para cada execução causa flutuações no número de fragmentos mapeados nas amostras. Para considerar esses problemas, a leitura por kilobase de transcrição por milhão de leituras mapeadas (RPKM) normaliza a contagem de leitura de uma transcrição tanto pelo tamanho quanto pelo número total de leituras mapeadas na amostra (GARBER *et al.*, 2011).

A abundância de cada transcrição é estimada utilizando um modelo probabilístico de máxima verossimilhança que faz uso de informações como distribuição do comprimento do fragmento, tamanho do gene, conteúdo GC, número de leituras multimapeadas e número e estrutura das isoformas previstas (GRIFFITH *et al.*, 2015).

Alguns autores argumentam que cinco milhões de leituras mapeadas são suficientes para quantificar com precisão genes altamente expressos na maioria dos transcritomas eucarióticos. Outros sequenciam até 100 milhões de leituras para quantificar precisamente genes e transcritos com baixos níveis de expressão (CONESA *et al.*, 2016; SIMS *et al.*, 2014).

A seguir, é apresentado um fluxograma que ilustra os passos gerais comumente utilizados para uma análise de expressão diferencial utilizando a tecnologia de RNA-Seq (Figura 1).

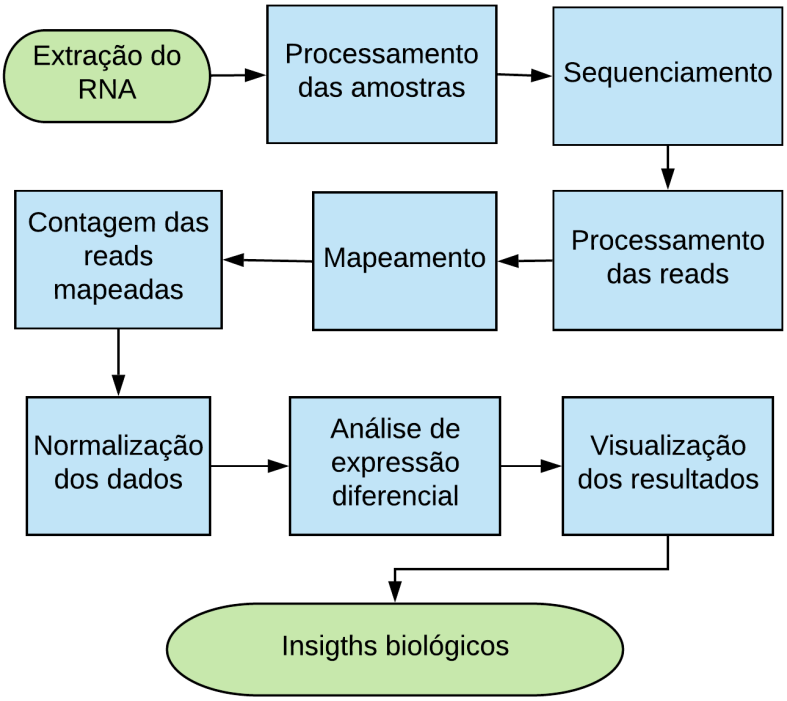

Figura 1
Fluxograma do processo de análise de expressão diferencial utilizando a tecnologia de RNA-Seq. O processo se inicia com a extração e preparação das amostras de RNA. Posteriormente, as amostras são sequenciadas, gerando um conjunto de *reads* que são processadas para verificar a qualidade do sequenciamento. As *reads* que passam no controle de qualidade são mapeadas contra um genoma ou transcritoma de referência, após isso, são contadas e normalizadas para ajustar as diferenças entre as amostras. Com as *reads* normalizadas é possível realizar a análise de expressão diferencial visualizando os resultados em forma de tabelas ou *heatmaps* onde então é possível obter *insigths* sobre a função de cada gene expresso.

Fonte: Autor (2019)

Aplicações da tecnologia de RNA-Seq na área agronômica

O sequenciamento completo do primeiro genoma vegetal, Arabidopsis thaliana, no ano 2000, forneceu ímpeto para incursões em investigações moleculares de plantas até os dias atuais. O crescente número de genomas sequenciados impulsiona descobertas biológicas e evolutivas em toda a faixa taxonômica das plantas (MARTIN *et al.*, 2013).

O RNA-Seq provou ser uma ferramenta poderosa com uma variedade notavelmente diversificada de aplicações em estudos do transcritoma de plantas. A relativa facilidade das análises genéticas em muitas espécies de plantas e o valor comercial das espécies cultivadas, tornaram a ciência das plantas uma área especialmente fértil para muitas tecnologias "ômicas" (MARTIN *et al.*, 2013).

A tecnologia de RNA-Seq, foi usada pela primeira vez para estudar plantas há apenas alguns anos (WEBER *et al.*, 2007) e agora fornece acesso pronto a informações transcritômicas de alta resolução. Isso é exemplificado pelo projeto 1KP2, que visa sequenciar os transcritomas de 1.000 espécies vegetais, sendo esta, apenas uma das muitas iniciativas atuais que estão expandindo radicalmente a amplitude e profundidade de nossa compreensão da expressão e evolução dos genes das plantas (MARTIN *et al.*, 2013).

Essa técnica tem sido utilizado em estudos da expressão gênica de várias plantas de importância econômica. Estudos de transcritoma em cana-de-açúcar aumentaram o painel de possíveis marcadores moleculares e informações de sequências disponíveis para programas de melhoramento, podendo resultar em várias melhorias biotecnológicas (CARDOSO-SILVA *et al.,* 2014).

Lopez-Casado *et al.* (2012) demonstraram que o perfil de transcritoma baseado em RNA-Seq pode fornecer um conjunto de dados efetivos para análise proteômica de organismos não-modelos pela montagem de novo de ESTs derivadas do pólen de tomate (*Solanum lycopersicum*) e de dois parentais silvestres. Isso sugere que o RNA-Seq é inestimável para facilitar a identificação de proteínas e que os estudos proteômicos não precisam mais ser taxonomicamente restritos (MARTIN *et al*., 2013).

Uma análise do transcritoma de bagas de uva (*Vitis vinifera*) durante três estágios de desenvolvimento identificou mais de 6.500 genes que foram expressos de maneira específica para cada estágio (ZENONI *et al.,* 2010). Da mesma forma, Wang *et al.* (2012) analisaram o transcritoma de raízes de rabanete (*Raphanus sativum*) em dois estágios de desenvolvimento e encontraram mais de 21.000 genes diferencialmente expressos, incluindo genes relacionados ao desenvolvimento radicular, metabolismo de amido e sacarose e com a biossíntese de fenilpropanoides (MARTIN *et al.,* 2013). Da mesma forma, Severin *et al.* (2010) forneceram um registro da expressão gênica em alta resolução em um conjunto de catorze tecidos diversos de soja (*Glycine max*) (SEVERIN *et al.,* 2010).

Além de estudos com foco em mudanças transcricionais durante o desenvolvimento, o RNA-Seq já se mostrou uma estratégia altamente eficaz para estudar respostas de plantas e adaptações a estresses abióticos e bióticos. Analisando dados de RNA-Seq derivados de plantas de sorgo (*Sorghum bicolor*) tratadas com ácido abscísico (ABA) ou polietilenoglicol, em conjunto com análise de transcritoma publicada para Arabidopsis thaliana, milho e arroz, Dugas et al. (2011) descobriram mais de 50 genes anteriormente desconhecidos sensíveis à seca. O RNA-Seq também foi usado para revelar mudanças maciças no metabolismo e na fisiologia celular da alga verde *Chlamydomonas reinhardtii* quando as células se tornam privadas de enxofre (GONZÁLEZ-BALLESTER *et al.,* 2010). Informações levantadas através de RNA-Seq resultaram em estudos de respostas de plantas a patógenos e da complexidade das vias metabólicas associadas a mecanismos de defesa nas plantas. Os exemplos publicados até o momento incluem uma análise transcritômica da infecção do sorgo pelo fungo *Bipolaris sorghicola* (MIZUNO *et al.,* 2012) e uma investigação sobre os mecanismos de defesa da soja que fornecem resistência a *Xanthomonas axonopodis*, comparando as espécies resistentes e suscetíveis (MARTIN *et al.,* 2013).

Em um estudo sobre o genoma do arroz, o uso do RNA-Seq levou à descoberta de 649 genes que estavam faltando na anotação do genoma de arroz, mas que foram expressos diferencialmente em resposta ao estresse salino (MARTIN; WANG, 2011). Já O'Rourke et al. (2012) utilizaram a tecnologia de RNA-Seq para analisar o perfil de expressão em raízes de tremoço branco (*Lupinus albus L.*) em solos com deficiência de fósforo, identificando um total de 2128 sequências expressas diferencialmente em resposta à essa deficiência, e desse total, 12 foram diferencialmente expressas também em *Arabidopsis thaliana* e *Solanum tuberosum* e sob déficit de fósforo, indicando que essas sequências podem ser candidatos à serem utilizadas para monitorar o nível desse nutrientes em plantas.

Considerações Finais

Como toda tecnologia em evolução, o RNA-Seq ainda tem desafios a superar. A concordância com resultados obtidos a partir de diferentes ferramentas já estabelecidas ainda é insatisfatória, pois os resultados são muito afetados por pequenos ajustes de parâmetros, especialmente para genes que são expressos em níveis baixos (CONESA *et al.,* 2016).

A eficiência do RNA-Seq é prejudicada pela grande quantidade de RNA ribossômico (rRNA) nos dados, leituras curtas, menor precisão da base e variação da densidade lida ao longo do comprimento do transcrito (KOGENARU *et al.,* 2012).

Além disso, as análises de RNA-Seq também são vulneráveis aos vieses e erros gerais inerentes às tecnologias de NGS, nas quais se baseia. Esses erros e vieses incluem: erros de sequenciamento (ligações erradas), vieses na qualidade da sequência, composição de nucleotídeos e taxas de erro relativas à posição base na leitura, variabilidade na profundidade da sequência ao longo do transcritoma devido a locais preferenciais

de fragmentação, efeitos de composição de nucleotídeos primários e transcritos variáveis e, finalmente, diferenças na cobertura e composição de dados de sequência bruta gerados a partir de replicatas técnicas e replicatas biológicas (ROBLES *et al.,* 2012).

O RNA-Seq resulta em uma medição discreta para a expressão gênica, diferentemente da medição da intensidade de fluorescência das tecnologias de microarranjos que são tratadas como uma variável contínua. Logo, os métodos estatísticos usados para analisar os dados de microarranjos não são diretamente aplicáveis, e novas abordagens estatísticas apropriadas para manipular os dados de RNA-Seq são necessárias (KVAM *et al.,* 2012).

Há carência na literatura de desenhos experimentais eficientes para a detecção de expressão diferencial utilizando a tecnologia de RNA-Seq. Assim, não existe consenso sobre uma abordagem padrão e abrangente para contornar as muitas fontes de ruído e vieses presentes no RNA-Seq (ROBLES *et al.,* 2012).

Outra limitação importante da tecnologia de RNA-Seq é que ela representa apenas um único instantâneo da expressão de RNAm de uma população de células e esta expressão nem sempre se correlaciona com a expressão proteica, devido à vários eventos pós-transcricionais que podem ocorrer nas células (GRIFFITH *et al.,* 2015; WOLF, 2013).

Os parâmetros de projeto experimental de RNA-Seq permanecem uma área em desenvolvimento e podem ter impactos significativos na estratégia de análise. Estes parâmetros incluem a realização de enriquecimento da cauda poli (A) de RNA total ou estratégias seletivas de redução de RNA ribossômico, como realizar a seleção de tamanho, o uso de amplificação linear para resgatar amostras com RNA disponível limitado, o uso de métodos de construção de bibliotecas em cadeia ou não-unidimensionais e o uso de técnicas de normalização de cDNA (GRIFFITH *et al*., 2015).

Da mesma forma, a escolha da plataforma de sequenciamento (Illumina, Ion Torrent, etc.), instrumento (ION Personal Genome Machine [PGM], MiSeq, HiSeq, etc.), comprimento de leituras geradas e outros parâmetros podem influenciar nas etapas de análise e interpretação dos dados (GRIFFITH *et al.,* 2015; EGAN *et al*., 2012).

O RNA-Seq se apresenta como uma plataforma versátil, aplicado em vários campos de pesquisa em biologia vegetal. O desenvolvimento contínuo das tecnologias de sequenciamento, tais como maior comprimento de leitura, maior número de leituras por execução e ferramentas de bioinformática que facilitem a montagem, análise e integração de sequências irão acelerar ainda mais a amplitude e a frequência de sua adoção por cientistas de plantas (MARTIN *et al.,* 2013).

Através da revisão bibliográfica realizada nesse trabalho, demonstramos a grande aplicabilidade da tecnologia de RNA-Seq para estudo de transcritomas de várias espécies de planta de intesse comercial, contribuindo para a ampliação do conhecimento nessa área que pode ser usada em programas de melhoramento vegetal com o intuito de desenvolver variedades mais produtivas.

Referências

BEHJATI, S.; TARPEY, P. S. What is next generation sequencing? **Archives of Disease in Childhood - Education and Practice**, v. 98, n. 6, 2013.

CARDOSO-SILVA, C. B.; COSTA, E. A.; MANCINI, M. C.; BALSALOBRE, T. W. A.; CANESIN, L. E. C.; PINTO, L. R.; CARNEIRO, M. S.; GARCIA, A. A. F.; SOUZA, A. P.; VICENTINI, R. De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties. **Plos One**, v.9, n. 2, 2014.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; SZCZESNIAK, M. W.; GAFFNEY, D.; ELO, L.; ZHANG, X.; MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v.17, n. 13, 2016.

EGAN, A. N.; SCHLUETER, J.; SPOONER, D. M. Applications of next-generation sequencing in plant biology. **American Journal of Botany**, v. 99, n. 2, 2012.

GARBER, M.; GRABHERR, M. G.; GUTTMAN, M.; TRAPNELL, C. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nature Methods**, v. 8, n. 6, 2011.

GIACHETTO, P. F.; HIGA, R. H. Bioinformática aplicada à agricultura. *In:* MASSRUHÁ, S. M. F. S.; LEITE, M. A. de A.; LUCHIARI JUNIOR, A.; ROMANI, L. A. S. **Tecnologias da informação e comunicação e suas relações com a agricultura.** Brasília, DF: Embrapa, 2014. Cap. 4. p. 67-83.

GONZÁLEZ-BALLESTER, D.; CASERO, D.; COKUS, S.; PELLEGRINI, M.; MERCHANT, S. S.; GROSSMAN, A. R. RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival. **The Plant cell**, v. 22, n. 6, 2010.

GRIFFITH, M; WALKER, J. R.; SPIES, N. C.; AINSCOUGH, B. J.; GRIFFITH, O. L. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. **Plos Computational Biology**, v. 11, n. 8, 2015.

KOGENARU, S.; YAN, Q.; GUO, Y.; WANG, N. RNA-seq and microarray complement each other in transcriptome profiling. **BMC Genomics**, v.13, n. 629, 2012.

KVAM, V. M. K.; LIU, P.; SI, Y. A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. **American Journal of Botany**, v. 99, n. 2, 2012.

LINDBERG, J.; LUNDEBERG, J. The plasticity of the mammalian transcriptome. **Genomics,** v. 95, n. 1, 2010.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics**, v.12, n. 10, 2011.

MARTIN, L. B. B.; FEI, Z.; GIOVANNONI, J. J.; ROSE, J. K. C. Catalyzing plant science research with RNA-seq. **Frontiers in Plant Science**, v.4, n. 66, 2013.

MIZUNO, H.; KAWAHIGASHI, H.; KAWAHARA, Y.; KANAMORI, H.; OGATA, J.; MINAMI, H.; ITOH, T.; MATSUMOTO, T. Global transcriptome analysis reveals distinct expression among duplicated genes during sorghum - *Bipolaris sorghicola* interaction. **BMC Plant Biology**, v. 12, n. 121, 2012.

O'ROURKE, J. A.; YANG, S.; MILLER, S. S.; BUCCIARELLI, B.; LIU, J.; RYDEEN, A.; BOZSOKI, Z.; UHDE-STONE, C.; JIN TU, Z.; ALLAN, D.; GRONWALD, J. W.; VANCE, C. P. A RNA-Seq Transcriptome Analysis of P i Deficient White Lupin Reveals Novel Insights Into Phosphorus Acclimation in Plants. **Plant Physiology Preview**, v. 161, n. 2, 2012.

QIAN, X.; BA, Y.; ZHUANG, Q.; ZHONG, G. RNA-Seq Technology and Its Application in Fish Transcriptomics. **OMICS A Journal of Integrative Biology**, v.18, n. 2, 2014.

ROBLES, J. A.; QURESHI, S. E.; STEPHEN, S. J.; WILSON, S. R.; BURDEN, C. J.; TAYLOR, J. M. Efficient experimental de sign and analysis strategies for the detection of differential expression using RNA-Sequencing. **BMC Genomics**, v.13, n. 484, 2012.

SEVERIN, A. J.; WOODY, J. L.; BOLON, Y. T.; JOSEPH, B.; DIERS, B. W.; FARMER, A. D.; MUEHLBAUER, G. J.; NELSON, R. T.; GRANT, D.; SPECHT, J. E.; GRAHAM, M. A.; CANNON, S. B.; MAY, G. D.; VANCE, C. P.; SHOEMAKER, R. C. RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. **BMC Plant Biology**, v.10, n. 160, 2010.

SIMS, D.; SUDBERY, I.; IIOTT, N. E.; HEGER, A.; PONTING, C. P. Sequencing depth and coverage: key considerations in genomic analyses. **Nat Rev Genet**, v. 15, n. 3, 2014.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA, G.; KIM, D.; KELLEY, D. R.; PIMENTEL, H.; SALZBERG, S.; RINN, J.; PACHTER, L., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, **Nature Protocols**, v. 7, n. 3, 2012.

WANG, S.; WANG, X.; HE, Q.; LIU, X.; XU, W.; LI, L. Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. **PlantCell Rep.,** v. 31, n. 8, 2012.

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

WEBER, A. P.; WEBER, K. L.; CARR, K.; WILKERSON, C.; OHLROGGE, J. B., Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. **Plant physiology**, v. 144, n. 1, 2007.

WOLF, J. B. W. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. **Molecular Ecology Resources**, v.13, n. 4, 2013.

YOON, S.; NAM, D. Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. **BMC Genomics**, v. 18, n. 408, 2017.

ZENONI, S.; FERRARINI, A.; GIACOMELLI, E.; XUMERLE, L.; FASOLI, M.; MALERBA, G.; DELLEDONNE, M.; BELLIN, D.; PEZZOTTI, M. Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. **Plant physiology,** v. 152, n. 4, 2010.