JACIANE COELHO GONÇALVES

INFLUÊNCIA DO NÚMERO DE REPETIÇÕES NA IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS EM EXPERIMENTOS DE RNA-SEQ

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

VIÇOSA
MINAS GERAIS-BRASIL
2013

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
G635i
2013

Gonçalves, Jaciane Coelho, 1981-
Influência do número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq / Jaciane Coelho Gonçalves. – Viçosa, MG, 2013.
vi, 33f. : il. (algumas color.) ; 29cm.

Inclui apêndice.
Orientador: Luiz Alexandre Peternelli
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 27-31

1. Estatística. 2. Biometria. 3. Sequënciamento de nucleotídeo. 4. Regulação de expressão gênica. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

JACIANE COELHO GONÇALVES

INFLUÊNCIA DO NÚMERO DE REPETIÇÕES NA IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS EM EXPERIMENTOS DE RNA-SEQ

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

APROVADA: 16 de janeiro de 2013

Moysés Nascimento
(Co-orientador)

Fernanda Miquelitto Figueira da Silva

Luiz Alexandre Peternelli
(Orientador)

AGRADECIMENTOS

Um sonho feito de ideais foi lapidado pelas dificuldades, impulsionado pelo desejo de acertar e fortalecido pelo medo de errar.
Por tudo isso...Agradeço.

A Deus, por conceder-me força, saúde e persistência nessa jornada.
À minha Mãe, seu amor, ideais e caráter sempre nortearam meus passos e são exemplos de conduta, dos quais sempre serei seguidora. Vencemos juntas!
À minha irmã Geiza, obrigada pelas orações, conselhos, amizade, carinho, você é meu espelho, meu anjo da guarda. Compartilho com você esta vitória!
Ao Roberto pelo carinho e respeito.
Aos meus anjinhos Ceceia e Juju, pelo sorriso no momento de que precisava, por escrever "Jaci eu te amo" em meus livros, apostilas e rascunhos, que no momento de desespero me deixava feliz...AMO VOCÊS....
À Jane, estimuladora suprema dos meus sonhos nesta caminhada, construída pela amizade e fortalecida pelo carinho.. Nos méritos dessa conquista, há muito da sua presença. Caminhamos juntas!
À Rosangela que, com apoio, incentivo, consideração e carinho impulsionou a conquista desse ideal. Divido com você este momento!
À melhor amiga que alguém pode ter, Geise, que com carinho me fez forte, com amizade me fez feliz, com incentivo me deu coragem. Você é um presente de Deus, um tesouro em minha vida...Amigas para sempre!
À Cris pela amizade, por entender minha ausência e por me socorrer nos momentos de desespero. Obrigada pela amizade sincera e verdadeira.
Ao meu orientador Luiz Alexandre Peternelli, primeiramente, preciso agradecê-lo por ter me aceitado como sua orientada e, principalmente por ter executado seu papel com excelência. Seu respeito por mim e sua dedicação estarão sempre presentes em minha memória. Preciso muito agradecer sua paciência e compreensão frente aqueles momentos em que necessitei dedicar a outras atividades de minha vida. Fico muito feliz e honrada por ter compartilhado comigo o seu precioso tempo, sua inteligência. Hoje posso dizer que tenho em quem me espelhar profissionalmente. Obrigada por fazer parte dessa conquista.

Ao Prof. Moysés, por não medir esforços em ajudar. Obrigada pela paciência, por me “co-orientar”, por não deixar faltar nada para atingir o resultado . Por estar sempre pronto em me receber quando precisei. Muito obrigada!!!

Ao Prof. Gustavo, sempre que precisei não mediu esforços para me ajudar.

A Fernanda Miquellito pela sua disponibilidade e colaboração que trouxe grandes melhorias para nosso trabalho.

Ao Talles, um presente divino!! Mesmo durante suas tribulações, não mediu esforços para me ajudar, sua paciência, seus conselhos e seus ensinamentos foram de grande importância para meu trabalho. Tudo que agradecer ainda será pouco...Obrigada por tudo!!

São muitos os responsáveis por essa vitória. É imensa e justa a minha gratidão.

MUITO OBRIGADA!

RESUMO

GONÇALVES, Jaciane Coelho, M.Sc., Universidade Federal de Viçosa, Janeiro de 2013. **Influência do número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq**. Orientador: Luiz Alexandre Peternelli. Coorientadores: Moysés Nascimento e Gustavo Costa Bressan.

Um dos objetivos atuais da biologia molecular é medir e avaliar os perfis de expressão gênica em diferentes tipos de tecidos biológicos, para entender os mecanismos de transformação molecular sob determinadas condições. Tecnologias de sequenciamento de Nova Geração (NGS) promovem o sequenciamento de DNA em plataformas capazes de gerar informações sobre milhões de pares de bases em uma única etapa. Porém essas tecnologias ainda apresentam custo elevado, dificultando a obtenção de elevado número de repetições de dados amostrais. Assim, torna-se necessária a descoberta e o aprimoramento de metodologias estatísticas eficientes para a otimização das análises de dados gerados em plataformas de sequenciamento de genomas. O objetivo geral desse trabalho consistiu em avaliar o efeito do número de repetições na identificação de genes diferencialmente expressos, em experimentos de RNA-Seq, contribuindo para o esclarecimento de pesquisadores que venham a auxiliar nas análises de dados em experimentos de RNA-Seq. De forma específica, avaliamos empiricamente o efeito do número de repetições na análise estatística da expressão gênica em experimentos de RNA-Seq. Para a realização das análises foi utilizado um conjunto de dados definido em Li *et al.* (2008), o qual comparou células cancerígenas tratadas e não tratadas. Naquele estudo havia quatro repetições biológicas para o grupo controle (células não tratadas) e três repetições biológicas para grupo de tratamento (células que receberam o tratamento). Os dados foram analisados utilizando o pacote DESeq do Programa computacional R. Um total de 2566 genes foram considerados diferencialmente expressos (DE) quando avaliamos o conjunto de dados original completo. Quando analisamos três repetições do controle e do tratamento, nós encontramos, em média, 2153 genes DE. A partir do momento em que apenas duas repetições para ambos os tratamentos foram utilizadas, foram identificadas, em média, 1241 genes DE. A grande alteração no número de genes DE foi observada quando repetições não foram utilizadas. Nesse caso identificamos em torno de 44 genes diferencialmente expressos. De acordo com os resultados gerados nas análises, foi possível verificar que o número de repetições é um fator essencial para se obter um número significativo de genes diferencialmente expressos.

ABSTRACT

GONÇALVES, Jaciane Coelho, M.Sc., Universidade Federal de Viçosa, January, 2013. **Influence of the number of repetitions in the identification of differentially expressed genes in RNA-Seq experiments**. Advisor: Luiz Alexandre Peternelli. Co-advisors: Moysés Nascimento and Gustavo Costa Bressan.

One of the current objectives of molecular biology is to measure and assess the gene expression profiles in different types of biological tissues, to understand the mechanisms of molecular transformation under certain conditions. Next-generation sequencing (NGS) technologies promote DNA sequencing in platforms capable of generating information about millions of base pairs in a single step. However these technologies still have high cost, making it difficult to obtain large number of repetitions of sample data. Therefore, it becomes necessary the discovery and the improvement of efficient statistical methodologies for optimizing analysis of data generated in genome sequencing platforms. The overall objective of this work was to evaluate the effect of the number of repetitions in the identification of differentially expressed genes, in RNA-Seq experiments, contributing to the clarification of the statistic that researchers will assist in data analysis in RNA-Seq experiments. Specifically, we evaluate empirically the effect of the number of repetitions in the statistical analysis of gene expression in RNA-Seq experiments. To carry out the analyses we used a dataset defined in Li *et al*. (2008), which compared treated and non-treated cancer cells. That work had four biological replications for the control group (non-treated) and three replications for biological treatment group (cells that have received the treatment). The data was analyzed using the package DESeq from the statistical environment R. A total of 2566 genes were considered differentially expressed (DE) when we evaluate the original and complete dataset. When we analyzed three replications of the control and treatment, we found, on average, 2153 genes DE. From the moment in which only two reps for both treatments were used, were identified, on average, 1241 genes DE. The major change in the number of genes DE was observed when replications were not used. In this case we identified around 44 differentially expressed genes. According to the results generated in the analysis, it was possible to verify that the number of repetitions is an essential factor in order to obtain a significant number of differentially expressed genes.

SUMÁRIO

1- INTRODUÇÃO

As informações genéticas necessárias para a manutenção da vida de um ser vivo são armazenadas em estruturas celulares conhecidas como DNA (ácido desoxirribonucleico). Essas informações são codificadas em pequenas porções de DNA, conhecidas como genes. Estes são transcritos em moléculas conhecidas como RNA (ácido ribonucleico) (LEHNINGER *et al.*, 2006; JUNQUEIRA, *et al.*, 1997) em um processo denominado expressão gênica (ESTEVES, 2007). Os RNAs são, então, traduzidos em proteínas que são as biomoléculas funcionais das células e que trabalham na manutenção de um ambiente favorável à sustentação do ciclo de vida celular.

A expressão adequada de genes é de grande importância para a manutenção dos processos vitais das células. Genes expressos em momentos indesejados podem levar ao surgimento de diversas doenças, como por exemplo, o câncer. Assim, um dos principais desafios da biologia molecular é medir e avaliar os perfis de expressão gênica em diferentes tecidos biológicos com o objetivo de entender os mecanismos de transformação molecular (ESTEVES, 2007).

Uma técnica utilizada para estudar os níveis de expressão gênica é a análise de microarranjos de DNA (*microarrays*), também conhecidos como *DNA chips*. Essa técnica permite verificar, de forma rápida e simultânea, os níveis de expressão de milhares de genes (NEVES, 2010; DIOGENES FILHO, 2009; ESTEVES, 2007). Porém, apesar de útil, as análises por microarranjo de DNA apresentam limitações tais como: alto índice de ruído devido à hibridização cruzada, limitação na faixa de detecção, devido ao próprio ruído e saturação de sinais, detecção apenas dos transcritos representados nos *spots* e a não garantia da cobertura total dos transcritos. Além disso, as análises de expressão comparativa entre diferentes experimentos são difíceis e requerem complicados métodos de normalização baseados em cálculos estatísticos muito complexos (VAN VLIET, 2010).

Atualmente, novas tecnologias de sequenciamento de ácido nucleico, denominadas tecnologias de sequenciamento de nova geração (NGS, or *New Generation Sequencing*) (AUER & DOERGE, 2010), têm sido utilizadas como alternativa aos microarranjos nos estudos de genes diferencialmente expressos. A NGS promove o sequenciamento de DNA em plataformas capazes de gerar informações sobre milhões de pares de bases em uma única etapa.

Dentre as novas plataformas de sequenciamento, duas são amplamente utilizadas. A plataforma 454 FLX da Roche®, primeira plataforma de sequenciamento de nova geração a

ser comercializada, e a Solexa da Illumina® (CARVALHO, 2010) que conseguiu reduzir os custos e aumentar a capacidade de sequenciamento (FARIAS *et al*., 2012). Outros dois sistemas de sequenciamento utilizados são a plataforma da *Applied Biosystems*, denominada *SOLiD System* e o *Heliscope True Single Molecule Sequencing* (tSMS), da Helicos®.

Essas novas plataformas possuem como características comuns o poder de sequenciar milhares de bases em uma só corrida, gerando muito mais informações que o método tradicional de Sanger, possibilitando assim uma grande economia de tempo e custo por base sequenciada (CARVALHO, 2010).

O sequenciamento direto do RNA (RNA-Seq) ainda não é possível com as plataformas de sequenciamento existentes. No entanto, é possível sequenciar indiretamente o RNA pela transcrição reversa em cDNA, ou seja, a fração que corresponde a região codificadora do genoma (genes). Dessa forma, é possível identicar os RNAs que estão sendo expressos em um dado organismo ou tecido. Diferentemente dos métodos de *microarray*, o RNA-Seq não necessita de conhecimento *a priori* do transcriptoma e não se limita apenas a avaliação de genes para os quais existam sondas. Assim, novos transcritos e novas variantes de *splicing* podem ser identificadas. Além disso, é possível determinar polimorfismos em regiões transcritas com resolução de um nucleotídeo (WANG *et al.*, 2009). Outra vantagem em relação aos microarranjos de DNA é que o RNA-Seq é pouco influenciado por sinal de fundo (*background*), comum em microarranjo, visto que as sequências geradas normalmente são mapeadas a uma única região no genoma (KESSLER, 2010).

Como o RNA-Seq utiliza uma medida absoluta (quantitativa), ele pode ser usado para determinar o nível de expressão gênica de forma mais acurada do que em microarranjos. Em princípio, é possível determinar a quantidade absoluta de cada molécula de mRNA numa determinada condição (quantificação absoluta). Assim, é possível comparar os resultados entre experimentos independentes, diferentemente de microarranjos nos quais normalmente é realizada uma quantificação relativa entre duas amostras (WANG *et al*., 2009).

Para determinar o nível de expressão de cada gene, o número de leituras de RNA-Seq deve ser convertido num valor quantitativo, normalmente este valor é obtido mapeando os *reads* sequenciados numa referência. Para possibilitar a subsequente comparação entre bibliotecas, também é feita uma normalização em relação ao número total de sequências obtidas por biblioteca (WILHELM E LANDRY, 2009).

Nas análises de RNA-Seq não há um limite máximo de detecção de transcritos. Esse valor se correlaciona com o número de sequências obtidas. Assim, os limites do alcance

dinâmico (*dynamic range*) são determinados apenas pela quantidade de leituras obtidas. Isso significa que através do sequenciamento contínuo de uma determinada biblioteca, seria possível medir a expressão de qualquer transcrito presente e, portanto, o alcance dinâmico só representaria a diversidade biológica do transcriptoma analisado (KESLLER, 2011).

Apesar das vantagens mencionadas acima, a metodologia de RNA-Seq enfrenta alguns desafios, tais como o armazenamento, a recuperação e o processamento das grandes quantidades de dados gerados. Outro desafio é o custo do sequenciamento, pois quanto maior a cobertura desejada, mais fragmentos devem ser sequenciadas, o que ocorre comumente quando se sequencia genomas novos, ou ainda pouco conhecidos (SILVA, 2012). O custo do sequenciamento é ainda um fator limitante em análises de RNA-Seq, reduzindo, assim, o número de repetições técnicas e biológicas nas análises.

Nesse trabalho queremos demonstrar que a variabilidade (que pode ser inerente aos organismos, a fatores humano ou devido aos fatores biológicos sob investigação), pode ser maior do que o esperado se não for estimada corretamente, o que poderá afetar os resultados dos estudos. A repetição não contribui, necessariamente, para o incremento da precisão do experimento, mas é extremamente importante para o aumento da precisão das estimativas de médias e de outras funções das variáveis respostas. A ampliação do número de repetições contribui, substancialmente, para o aumento da confiabilidade dessas estimativas e da sensibilidade do experimento para detectar pequenas, mas importantes, diferenças de efeitos de tratamento (KVAM & LIU, 2012).

Devido a curta história do RNA-Seq e seu desenvolvimento, não existem ainda um referencial teórico e aplicado para se detectar genes diferencialmente expressos com base em tais dados (KVAM & LIU, 2012). Assim, acredita-se que a elaboração de uma dissertação nesta área dará valiosas contribuições à ciência ajudando os pesquisadores no planejamento de novos experimentos, tanto no sentido acadêmico quanto financeiro, já que será avaliado o efeito no número de repetições em experimentos de RNA-Seq, pois essas tecnologias ainda apresentam custo elevado, dificultando a obtenção de amplo número de repetições de dados amostrais.

Diante do exposto, tem-se como objetivo geral avaliar o efeito do número de repetições na identificação de genes diferencialmente expressos, em experimentos de RNA-Seq, contribuindo para o esclarecimento de pesquisadores da área da Estatística que venham a auxiliar no planejamento e análise de dados de experimentos de RNA-Seq.

2-REVISÃO DE LITERATURA

2.1-Tecnologias de sequenciamento de nova geração para o sequenciamento indireto do RNA

O *microarray* era a técnica mais utilizada para determinação dos níveis de expressão gênica, porém alguns fatores dificultam a análise dos experimentos, tais como a necessidade do conhecimento prévio do genoma, as variações provocadas pelo denominado efeito *background* e a utilização de métodos normalizadores complexos (NEVES, 2010). O surgimento das tecnologias de sequenciamento de nova geração (NGS- *New Generation Sequencing*) foi uma alternativa para solucionar esses problemas e para suprir dados transcriptômicos, independentemente da necessidade de uma sequência genômica previamente conhecida.

A NGS emergiu como uma ferramenta revolucionária para estudos do transcriptoma, com geração de dados altamente reprodutíveis e com precisão na quantificação de transcritos. Essas tecnologias de nova geração permitem ainda o estudo de vários fenômenos biológicos, incluindo polimorfismo de nucleotídeo único, eventos epigenéticos, *splicing* alternativo e o estudo de interações proteína-DNA (WANG *et. al.*, 2009)

Essa eficiência em relação às demais técnicas advém do uso da clonagem *in vitro* e de sistemas de suporte sólido para as unidades de sequenciamento, não necessitando de grande trabalho laboratorial. A clonagem *in vitro* em suporte sólido permite que milhares de leituras possam ser sequenciadas de uma só vez gerando informações sobre milhões de pares de bases em uma única corrida. Dentre as novas plataformas de sequenciamento, as mais utilizadas são a plataforma 454 FLX da Roche, a Solexa da Illumina, *Solid System* da *Applied Biosystems* e o *Heliscope True Single Molecule Sequencing (tSMS)* da Helicos (CARVALHO, 2010).

O Sequenciamento indireto de RNA por NGS permite mapear os *reads* obtidos numa determinada referência e consequentemente quantificar uma população de transcritos (MORTAZAVI *et al*., 2008). De forma resumida podemos falar que o RNA é isolado a partir de fragmentos aleatórios de células e transcritos em cDNA o qual será sequenciado. Fragmentos de tamanho apropriado (por exemplo 200-300 pares de bases de comprimento) são selecionados para amplificação utilizando reação em cadeia da polimerase (PCR). Após a amplificação, o cDNA é sequenciado utilizando alguma tecnologia NGS. As leituras resultantes (*reads)* são mapeadas num genoma de referência, permitindo a obtenção de uma tabela que será o arquivo de entrada para análise da expressão gênica diferencial (AUER e

DOERGE, 2010; MOROZOVA *et al.* 2009). A Figura 1 apresenta uma visão simplificada das tecnologias de sequenciamento de nova geração.

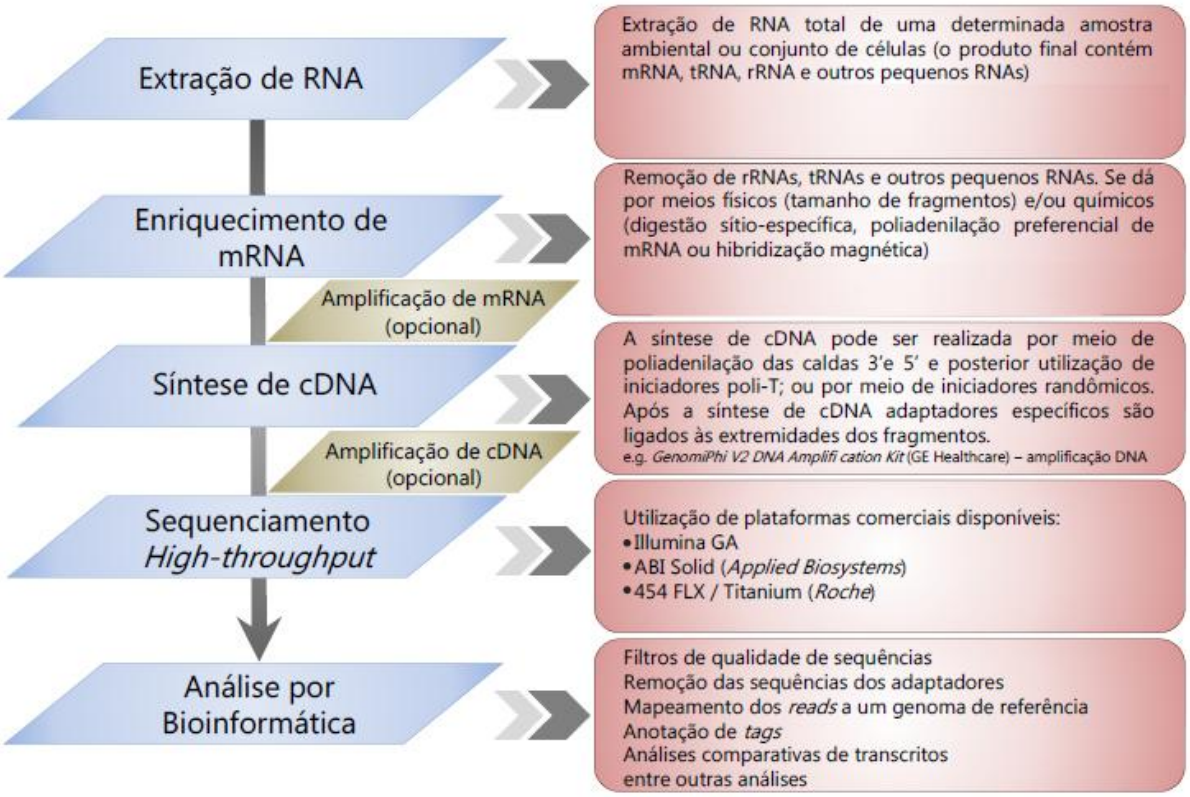

Figura1 – Fluxo de trabalho necessário para o sequenciamento indireto de um determinado transcriptoma por RNA-Seq. Os quadrantes em azul representam o fluxograma experimental, e os vermelhos uma descrição detalhada de cada processo com as opções metodológicas disponíveis no mercado. Fonte: adaptado de ANDREOTE (2011).

Embora existam muitas etapas nesse processo experimental que podem induzir a erros e distorções, o RNA-Seq tem sido aclamado como o futuro em pesquisa de transcriptoma, pois possui vantagens claras sobre as abordagens atuais, uma vez que proporciona uma maior sensibilidade do que os *microarrays.* Além disso, o método é capaz de discriminar regiões homólogas, não necessitando de conhecimento a *priori* do transcriptoma e não se limita apenas a avaliação de genes para os quais existam sondas (BULLAR *et al*., 2010, MARIONI *et al*., 2008). O RNA-Seq não possuiu um limite superior para a quantificação. Consequentemente, tem uma grande gama dinâmica dos níveis de expressão através da qual podem ser detectados transcritos.

2.2-Etapas para análises da expressão gênica utilizando RNA-Seq

2.2.1 - Mapeamento dos *reads*

Para comparar a expressão diferencial entre duas condições por RNA-Seq é necessário transformar milhões de leituras em uma quantificação de expressão. O primeiro passo desse procedimento é o mapeamento dos *reads* (sequência provenientes das plataformas de sequenciamento), numa determinada referência. Estes *reads*, às vezes podem alinhar a vários locais prejudicando a análise de expressão gênica. Portanto, a verdadeira tarefa é encontrar o local onde cada *read* melhor se alinha à referência (OSHLACK *et al*., 2010).

2.2.2 - Normalização

O objetivo da normalização é tornar os valores de expressão gênica comparáveis. No entanto, o procedimento para gerar dados do RNA-Seq é diferente daquele para os dados de microarranjos. Assim, a normalização não é aplicada exatamente da mesma maneira. No RNA-Seq normaliza-se levando em consideração o número de *reads* gerados por biblioteca e o tamanho da referência (ROBINSON & OSHLACK, 2010).

Tem sido demonstrado que a normalização é um passo essencial na análise de genes diferencialmente expressos (DGE) a partir de dados de RNA-Seq. Diversos métodos de normalização podem ser encontrados nos diferentes pacotes estatísticos disponíveis para análise (BULLARD *et al.*, 2010). Dentre eles podemos citar os pacotes DESeq, DEGSeq, edgeR e baySeq implementados no R. Além dos processos de normalização implementados no R, a normalização através do valor de RPKM (*Reads Per Kilobase of exon model per Million mapped sequence reads*) é bastante utilizada nas análises de RNA-Seq.

2.2.3 - Análise da expressão gênica diferencial

Em análises de dados de RNA-Seq existe um grande interesse na identificação dos genes diferencialmente expressos, ou seja, aqueles que mudaram seus níveis de expressão nas diferentes condições experimentais. Uma possibilidade para atingir esse objetivo consiste em organizar os dados em uma tabela de dados de contagem sintetizada para cada biblioteca e a realização de testes estatísticos entre as amostras de interesse que irá permitir alcançar a distinção entre genes que apresentam expressão diferencial.

Vários métodos têm sido desenvolvidos para a análise de expressão diferencial utilizando os dados de *microarray*. Entretanto, em análises de RNA-Seq se dá uma medição discreta para cada gene, enquanto que as intensidades geradas no *microarray* têm uma distribuição de intensidade contínua. Apesar das intensidades do *microarray* são tipicamente log-transformadas e analisadas como variáveis aleatórias normalmente distribuídas, a transformação dos dados de contagem não é bem aproximada por distribuições contínuas. Portanto, modelos estatísticos apropriados para dados de contagem são vitais para extrair o máximo de informações a partir de dados de RNA-Seq.

Distribuições discretas de probabilidade têm sido propostas para modelar os dados de contagem a partir de experimentos de RNA-Seq: Poisson e binomial negativa. Em geral, em alguns estudos a distribuição de Poisson constitui a base para os dados de modelagem de contagem de RNA-Seq. Entretanto, a variabilidade biológica não é bem captada pela distribuição de Poisson, devido ao fato de possuir um único parâmetro, o qual é exclusivamente determinado pela sua média, tornando assim a variância igual a sua média. (LANGMEAD *et al*., 2010). Assim, a distribuição de Poisson baseada em análises de conjuntos de dados com repetições biológicas estará propensa a altas taxas de falsos positivos resultantes da subestimação do erro experimental (ANDERS & HÜBER 2010). Apesar do baixo ruído e da alta sensibilidade da plataforma de RNA-Seq, projetar experimentos com a repetição biológica é ainda fundamental para identificar alterações na abundância de RNA que generalizam para a população a ser amostrada.

Para explicar a variabilidade biológica, a distribuição binomial negativa tem sido usada como uma extensão natural da distribuição de Poisson, requerendo um parâmetro de dispersão adicional a ser estimado (BULLARD, 2010).

2.2.3.1 - Binomial Negativa

A distribuição binomial conta o número de sucessos em um número fixo de provas de Bernoulli. Suponha que, em vez disto, contamos o número de provas de Bernoulli necessário para obter um número fixo de sucessos. Esta última formulação leva à distribuição binomial negativa.

Considerando a variável X como sendo igual ao número de fracassos requeridos para que ocorra o r-ésimo sucesso, então, X se distribui de acordo com a binomial negativa. Verifica-se que ocorreram x fracassos e $r-1$ sucessos antes do r-ésimo sucesso no último

ensaio. Cada configuração de x fracassos, com probabilidade $1-p$, com $r-1$ sucessos, com probabilidade p cada, tem probabilidade igual a $(1-p)^x p^{r-1}$. O número de combinações possíveis para organizar X fracassos em um total $x+r-1$ configurações é dado por:

$$\binom{x+r-1}{x} = \binom{x+r-1}{r-1}$$

Sendo assim, a função de probabilidade da distribuição binomial negativa pode ser obtida multiplicando-se esse número de combinações das configurações possíveis de fracassos e sucessos nos $x+r-1$ ensaios anteriores ao r-ésimo sucesso pela probabilidade de cada configuração $(1-p)^x p^{r-1}$, e ainda, pela probabilidade p da r-ésimo sucesso obtido independentemente no último ensaio. A função de probabilidade assim obtida é dada por:

$$P(X=x) = \binom{x+r-1}{x} p^r (1-p)^x$$

em que x = 0,1,2,3,...representa o número de fracassos até a ocorrência do r-ésimo sucesso, r=1,2,...e $0 < p \leq 1$. .

A média e a variância da distribuição binomial negativa são

$$\mu_x = \frac{r(1-p)}{p} \text{ e } \sigma^2{}_x = \frac{r(1-p)}{p^2}$$

2.3 - Seleção dos genes diferencialmente expressos

Vários pacotes, tais como o baySeq, DESeq, DEGSeq e edgeR (KVAM *et al.*, 2012) podem ser utilizados na busca de genes diferencialmente expressos. No entanto, vamos, neste trabalho, nos ater apenas ao pacote DESeq implementados no software livre R versão 2.15.1 (R Development Core Team, 2012), disponível em http://www.R-project.org. O pacote DESeq foi escolhido por permitir análise de dados oriundos de experimentos sem repetição (ANDERS, 2012; KVAM *et al.*, 2012), uma vez que um dos objetivos do trabalho é avaliar o efeito de nenhuma ou de pelo menos uma repetição na análise estatística de expressão gênica.

2.3.1- DESeq

O DESeq é um pacote desenvolvido para analisar os dados de contagem a partir de ensaios de alto rendimento de sequenciamento, tais como RNA-Seq.

A contagem de dados segue uma distribuição de Poisson. No entanto, uma limitação com essa distribuição é que ela assume média igual à variância, ou dispersão (ou seja, $\mu = \sigma^2$). Os testes de expressão diferencial entre duas condições experimentais devem levar em conta tanto a variabilidade técnica quanto a biológica. No entanto, os testes com base na distribuição de Poisson ignoraram a variação de amostragem biológica, ou seja para dados em que apresentam superdispersão, análises baseadas em Poisson será propensa a alta taxa de falsos positivos resultantes de uma subestimação do erro de amostragem (KVAM & LIU, 2012).

Quando existem repetições biológicas, dados de RNA-Seq podem apresentar maior variabilidade, ou seja, a variância é susceptível de ultrapassar o valor da média consideravelmente para muitos genes, que é descrito na literatura como problema de superdispersão (KVAM & LIU, 2012).

Uma solução para esse problema mencionado acima, é assumir que os dados seguem uma distribuição binomial negativa que é normalmente usada para tratar superdispersão. Sendo assim o DESeq utiliza um modelo baseado na distribuição binomial negativa que controla o problema da variabilidade da amostra. Assim, um ensaio com base na distribuição binomial negativa, que pode refletir essas propriedades, tem um potencial muito mais elevado para detectar genes diferencialmente expressos, permitindo modelar a variância biológica corretamente.

2.4- Correção de testes múltiplos

Um dos principais objetivos em análises de RNA-Seq é a identificação de genes que apresentam expressão gênica diferencial, cujo nível de expressão está associado a uma resposta ou variável de interesse. Na grande maioria das vezes é testada uma hipótese para cada gene e, assim, tem-se um grande número de hipóteses sendo testadas simultaneamente. Como consequência ocorre o problema de multiplicidade.

Quando várias hipóteses são testadas, a probabilidade conjunta de que o erro tipo I seja cometido aumenta expressivamente com o número de hipóteses. O erro tipo I, também

denominado de falsos positivos (CASELLA, 2010), é o erro que se comete ao rejeitar a hipótese nula quando a mesma é verdadeira (BENJAMINI & HOCHBER, 1995). O erro tipo I ocorre ao afirmar que um gene apresenta expressão gênica diferencial quando na realidade, isso não ocorre.

Para controlar a taxa de erro global pode-se utilizar o método de Bonferroni ou o ajuste de FDR (*False Discovery Rate*). Esses procedimentos ajustam os níveis descritivos individuais, garantindo o controle da taxa de falsas descobertas.

Adotando-se um nível de significância α para cada teste, tem-se que o nível de significância conjunto (N.S.C) do teste, α^*, considerando os *t* testes independentes será:

$$\alpha^* = 1-(1-\alpha)^t$$

À medida que aumenta o número de testes o nível significância conjunta aumenta de forma significativa (Figura 2), justificando a necessidade da correção dos testes múltiplos.

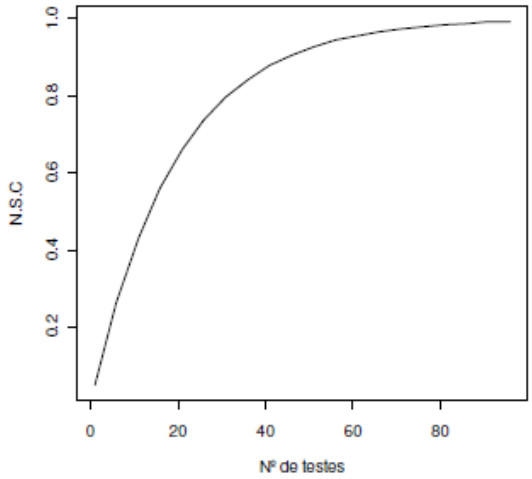

Figura 2–Comportamento do Erro Tipo I para diferentes quantidades de testes (NEVES, 2010).

2.4.1- Proteção de Bonferroni

A utilização da proteção de Bonferroni visa manter um nível de significância conjunta α_T para o experimento. Para isso deve-se estimar o nível de significância α para cada teste, que proporcione o nível de significância α_T para o experimento.

A correção de Bonferroni é dada por (SCHUSTER & CRUZ, 2008):

$$\alpha = -\exp\left(\frac{\ln(1-\alpha_T)}{N}\right)+1$$

Essa correção pode ser aproximada por (SCHUSTER & CRUZ, 2008)

$$\alpha \cong \frac{\alpha_T}{N}$$

Quando utilizada em experimentos com RNA-Seq, devido a grande quantidade de genes estudados, a correção de Bonferroni passa a ser muito conservadora, diminuindo de maneira drástica o poder do teste para a detecção de genes diferencialmente expressos. Além disso, num experimento com tantos genes e testes considerados, o erro tipo I na maioria das vezes, não possui tanta relevância em análises de RNA-Seq, visto que em análises de expressão gênica, em geral os pesquisadores aceitam maiores riscos em termos de falsos positivos para diminuir a chance de que genes com expressão importantes não sejam detectados (ROSA *et al.*, 2007).

Um critério menos restritivo para testes múltiplos refere-se à taxa de falsos positivos (FDR), denominada como a proporção esperada de falsos positivos entre todos os testes significativos.

2.4.2- FDR - False Discovery Rate

Benjamim & Hochberg (1995) propuseram controlar a FDR, definida como a proporção de hipóteses nulas H_0 verdadeiras, entre as hipóteses nulas rejeitadas, ou seja, a proporção de erros devido à rejeição incorreta de H_0. Diferentemente do nível de significância, o qual é pré-estabelecido antes de iniciar as análises. O FDR é calculado após a realização dos múltiplos testes de hipóteses, sendo calculada a partir das informações presentes nos dados.

Um procedimento bastante utilizado para calcular o FDR é o *Linear Step-Up* proposto por Benjamim & Hochberg (1995). Esse procedimento ordena os p-valores $p_{(1)} \leq ... \leq p_{(m)}$ resultantes das m hipóteses $H_{(1)}...,H_{(m)}$, testadas de forma simultânea. Sejam $p_{(1)}, p_{(2)}, ..., p_{(m)}$ os p-valores ordenados, define-se

$$q^* \geq \frac{mP_{(i)}}{i}$$

Portanto para controlar a FDR a um nível q^* = 5% o ponto de corte será o $P_{(i)}$ com maior i que satisfaça a condição: $5\% \geq \frac{mP_{(i)}}{i}$, ou seja, serão rejeitadas as hipóteses com p-valores menores ou iguais a $P_{(i)}$.

A título de ilustração do procedimento, considere-se o exemplo apresentado por Benjamini e Hochberg (1995), em que foram realizados 15 testes de hipóteses, e que os p valores ordenados, tenham sido: 0,0001; 0,0004; 0,0019; 0,0095; 0,0201; 0, 02780, 0298; 0,0344; 0,0495; 0,3240; 0,4262; 0,5719; 0,6528; 0,7590; 1,0000.

Considerando-se, ainda, que se deseje obter um nível de significância conjunto:

$$\alpha^{*} = 0,05 \Rightarrow 0,05 \geq \frac{15P_{i}}{i}$$

Assim, para $P_{(4)}$= 0,0095 ➔ 0,05 ≥15(0,0095)/4≥0,0356; $P_{(5)}$= 0,0201 ➔ 0,05 ≤15(0,0201)/5≤0,0603, portanto, devem-se rejeitar todas as hipóteses com p-valores menores ou iguais a 0,0095.

3 - MATERIAL E MÉTODOS

3.1- Dados experimentais

Para a realização das análises utilizou-se um conjunto de dados de RNA-Seq definido no artigo de Li *et al*. (2008), intitulado "*Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model*". O estudo comparou células cancerígenas tratadas e não tratadas, existindo quatro repetições biológicas para o grupo controle (não tratadas) e três repetições biológicas para grupo de tratamento (células que receberam o tratamento). Os dados foram analisados com o auxílio do pacote DESeq implementado no bioconductor (GENTLEMAN, *et al*., 2004) utilizando o programa computacional R (R Development Core Team, 2012). O módulo utilizado reconhece um arquivo organizado conforme exemplificado na Tabela 1.

Tabela 1. Exemplo (com 5 transcritos) de tabela necessária para análise estatística pelo DESeq

Identificação do transcrito	***reads* da condição 1 mapeados**	***reads* da condição 2 mapeados**
transcrito 1	2	2
transcrito 2	0	0
transcrito 3	35	3
transcrito 4	0	3
transcrito 5	5	0

Foram desconsideradas da análise as linhas da tabela que não tiveram *read* mapeado em nenhum dos grupos comparados, pois o valor zero influencia a estimativa da variância, uma vez que o método utiliza os valores de todos os tratamentos que estão sendo comparados para o cálculo desta estimativa.

Com o intuito de avaliar o efeito do número de repetições foram primeiramente analisadas todas as repetições (n) num único experimento. Posteriormente, implantou-se uma sequência de análises com redução do número de repetições para controle e/ou tratamento, ou seja, foram eliminadas k (k = 1 a $n-1$) repetições do conjunto de dados original de modo a ser possível avaliar o efeito da redução do número de repetições nas análises. Para cada

número de repetições k eliminado foram analisados C_n^{n-k} experimentos, totalizando 35 análises. Os cenários de combinações encontram-se na Tabela 2.

Tabela 2 - Cenários de combinações adaptadas do experimento de Li *et al.* (2008)

Cenário	Controle (C)	Tratamento (T)
1	C1,C2,C3,C4	T1,T2,T3
2	C1,C2,C3	T1,T2,T3
3	C1,C2,C4	T1,T2,T3
4	C1,C3,C4	T1,T2,T3
5	C2,C3,C4	T1,T2,T3
6	C1, C2	T1,T2
7	C1, C3	T1,T2
8	C1, C4	T1,T2
9	C2, C3	T1,T2
10	C2, C4	T1,T2
11	C3, C4	T1,T2
12	C1, C2	T1,T3
13	C1, C4	T1,T3
14	C2, C3	T1,T3
15	C2, C4	T1,T3
16	C1, C3	T1,T3
17	C3, C4	T1,T3
18	C1, C2	T2,T3
19	C1, C4	T2,T3
20	C2, C3	T2,T3
21	C2, C4	T2,T3
22	C1, C3	T2,T3
23	C3, C4	T2,T3
24	C1	T1
25	C2	T1
26	C3	T1
27	C4	T1
28	C1	T2
29	C2	T2
30	C3	T2
31	C4	T2
32	C1	T3
33	C2	T3
34	C3	T3
35	C4	T3

Criado todos os cenários de interseção foram realizadas as seguintes análises utilizando o pacote DESeq implementado no software R.

3.2 Análise da expressão diferencial

3.2.1 Descrição do modelo

Considere que o número de *reads* da *j*–ésima amostra, referente ao gene *i*, segue uma distribuição binomial negativa (ANDERS & HUBER, 2010)

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$$

em que μ_{ij} é a média e σ_{ij}^2 é a variância

3.2.2 Estimação dos parâmetros

Na prática os parâmetros média e variância são desconhecidos. Assim é necessário estimá-los a partir dos dados. Devido ao pequeno número de repetições em experimentos de RNASeq, Anders e Huber (2010), visando obter estimativas confiaveis, propuseram alguns pressupostos para essa modelagem. São elas:

- A média μ_{ij}, isto é, o valor esperado das contagens observadas para o gene *i* na amostra *j*, é o produto de uma condição dependente de valor por gene $q_{i,\rho(j)}$ (onde $\rho(j)$ é a condição experimental da amostra *j)* e um fator (fator de correção) de tamanho s_j,

$$\mu_{ij} = q_{i,\rho(j)} s_j.$$

- A variância é definida como uma função da média e uma quantidade ajustada da variância amostral suavizada

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)}.$$

- A variância é uma função suavizada de q_i e a condição experimental ρ

$$v_{i,\rho(j)} = v_p(q_{i,\rho(j)})$$

A terceira suposição é relacionada com fato do número de repetições ser tipicamente pequeno. Assim, para ser obter uma estimativa precisa da variação dos genes i , esta hipotese se torna necessária.

Para estimar os fatores de tamanho referentes a j-esima repetição, utiliza-se como estimativa a mediana da razão dos valores observados, não padronizando os *reads*, mas sim as médias dos *reads*. Utiliza-se a mediana, visto que a ocorrência de valores muito altos influenciam na padronização. O denominador dessa expressão pode ser interpretado como uma amostra obtida tomando a média geométrica entre as amostras. Assim, cada estimativa do fator tamanho é calculado como a mediana das razões das contagens da j-ésima amostra aos da pseudo-referência.

$$\hat{s}_j = \underset{i}{median} \frac{k_{ij}}{\prod_{v=1}^{m} k_{ij}^{1/m}}$$

Para estimar a quantidade $q_{i\rho}$ usamos a média das contagens a partir de j amostras correspondente a condição ρ transformada para a escala comum:

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{K_{ij}}{\hat{s}_j},$$

onde m_ρ é o número de repetições na condição ρ e o somatório sobre essas repetições. Observa-se que os valores de K são divididos pelo fator tamanho. Essa estratégia serve para tornar possível a comparação entre diferentes amostras.

Para obtenção da quantidade $\nu_{i\rho}$, primeiramente calcula-se as variâncias amostrais em escala comum por meio da seguinte expressão:

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{K_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2,$$

em que

$$z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j}.$$

Anders e Huber (2010) mostraram que $w_{i\rho}$ - $z_{i\rho}$ é um estimador não viciado de $\nu_{i\rho}$.

No entanto, para um pequeno número de repetições m_ρ, como é tipicamente o caso nas aplicações, os valores de $w_{i\rho}$ são altamente variáveis, e $w_{i\rho}$ - $z_{i\rho}$ não seriam um bom estimador da variância $\nu_{i\rho}$. Para contornar tal situação, ajusta-se um modelo de regressão local (LOADER, 1999) considerando os pares $(\hat{q}_{i\rho}, w_{i\rho})$ obtendo assim uma função suavizada $\tilde{w}_\rho(\hat{q})$. Posteriormente obtem-se uma estimativa para a variância por meio de: realiza-se

$$\hat{v}_\rho(\hat{q}_{i\rho}) = w_\rho(\hat{q}_{i\rho}) - z_{i\rho}.$$

Após a obtenção desta quantidade podemos realizar um teste de hipótese para verificar expressão diferencial.

3.3 - O teste estatístico para genes diferencialmente expressos

Os *p-valores* são calculados através de um método que é análogo ao Teste Exato de Fisher, aplicado a uma tabela de contingência 2x2. Entretanto, ao invés de assumir que os dados sigam uma distribuição hipergeométrica, eles seguem uma distribuição binomial negativa parametrizada a partir da média e da dispersão estimada (ANDERS, 2012).

3.4 - Análise de *Fold-Change*

O pacote DESeq informa também o valor da razão de expressão (*fold-change*) em logaritmo de base 2 ($\log_2$ *fold-change*) de uma condição em relação à outra. Especificamente, o gene é considerado diferencialmente expresso, se o valor de *fold-change* exceder um valor de corte arbitrário pré-estabelecido ou determinado empiricamente (NEVES, 2010; ESTEVES 2007; CRISTO, 2003). Para visualização dos genes diferencialmente expressos, pode ser utilizado o gráfico de MA-plot, conforme ilustrado na Figura 3.

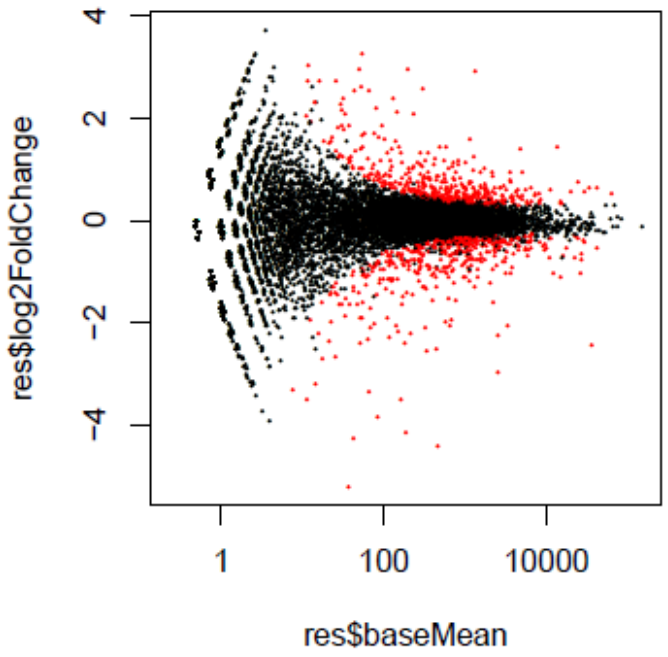

Figura 3- Diagrama de dispersão do *foldchange* transformado versus a média, evidenciando os genes diferencialmente expressos (cor vermelha). Acima do 0 no eixo das coordenadas, estão os genes mais expressos no tratamento e abaixo os mais expressos no controle (Anders – 2012)

Através do MA-plot conseguimos representar a dispersão da variabilidade (M), dos valores de expressão gênica na escala logarítmica (base 2) entre as amostras, pela média (A) do conjunto de todos os fragmentos genéticos.

Onde define-se $M = \log_2 C_1 - \log_2 C_2$ e $A = \frac{\log_2 C_1 + \log_2 C_2}{2}$, assumindo que C_1 e C_2 são independentes e seja C_1 e C_2 denotarem os valores dos dados normalizados para um gene específico obtido a partir de duas amostras (NEVES, 2010).

3.7 - Correção de testes múltiplos

Para controlar a taxa de erro global utilizou o ajuste de FDR com o nível de significância (α) igual a 5%.

3.8 - Organização e interpretação dos resultados

Realizadas todas as análises descritas acima e com o objetivo de avaliar a robustez do teste quanto ao número de repetições, foram feitas as análises em todos os cenários e posteriormente foram analisados os percentuais de interseções de todos os cenários, estando os genes ordenados em ordem de expressão gênica e comparados com o cenário 1 (em que possuímos todas as repetições tanto para o controle, quanto para o tratamento).

Posteriormente foram criados novos limites de interseção do cenário 1 com os demais cenários. Tendo como objetivo avaliar o percentual de interseção com os genes mais expressos, tendo como limite os cem mais expressos (1:100) os 28 mais expressos (1:28), valor este estipulado, devido ao fato do menor número de genes encontrados quando não se trabalhava com repetição, e os 10 mais expressos (1:10).

4 - RESULTADOS E DISCUSSÃO

Realizadas as análises pelo pacote DESeq, foram encontrados 2566 genes considerados diferencialmente expressos. Para melhor visualização utilizamos a Figura 4, em que se compara o nível da expressão gênica entre tratamento e controle.

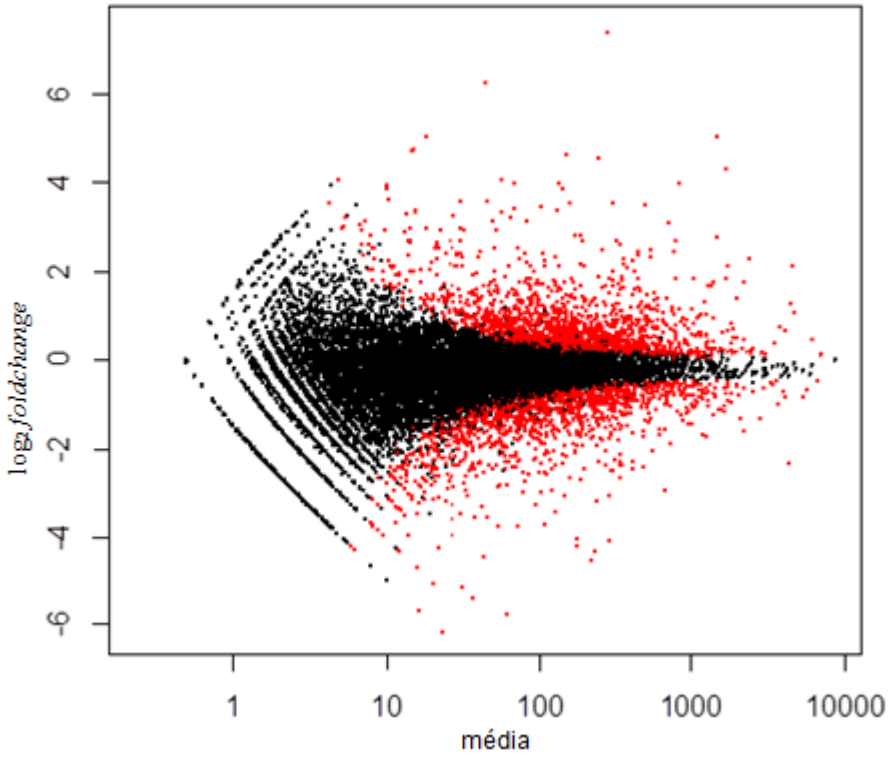

Figura 4. Diagrama de dispersão do *foldchange* transformado versus a média evidenciando os genes diferencialmente expressos (cor vermelha). Acima do 0 no eixo das ordenadas, estão os genes mais expressos no tratamento e abaixo os mais expressos no controle.

Representados em vermelho, acima do valor zero, no eixo das ordenadas, estão os genes mais expressos no tratamento de acordo com o método empírico de *fold-change*, onde definimos o ponto de corte através do *p-valor*. Abaixo, também em vermelho, estão representados os genes mais expressos no controle.

Após determinado o número de genes diferencialmente expressos utilizando as 4 repetições para controle e as 3 repetições para tratamento foram investigados o números de genes diferencialmente expressos e os percentuais de interseção em cada cenário conforme apresentado na tabela 3.

Tabela 3 - Números de genes diferencialmente expressos em cada cenário e percentuais de interseção

Cenário	Controle	Tratamento	genes DE	% interseção com o cenário 1	% Interseção 1:28	% Interseção 1:100	% Interseção 1:10
1	C1,C2,C3,C4	T1,T2,T3	2566				
2	C1,C2,C3	T1,T2,T3	2083	97,22	96,43	94	90
3	C1,C2,C4	T1,T2,T3	2037	97,94	96,43	94	90
4	C1,C3,C4	T1,T2,T3	2189	97,40	92,85	98	90
5	C2,C3,C4	T1,T2,T3	2304	96,53	100	96	90
6	C1, C2	T1,T2	1403	95,51	89,29	87	80
7	C1, C3	T1,T2	1488	95,77	85,71	84	70
8	C1, C4	T1,T2	1453	96,42	89,29	88	80
9	C2, C3	T1,T2	1714	94,34	92,85	86	70
10	C2, C4	T1,T2	1605	95,58	89,29	87	80
11	C3, C4	T1,T2	1924	93,24	89,29	89	70
12	C1, C2	T1,T3	1009	97,72	85,71	88	90
13	C1, C4	T1,T3	1012	97,63	89,29	92	90
14	C2, C3	T1,T3	1116	96,59	85,71	90	90
15	C2, C4	T1,T3	1083	96,86	92,85	90	90
16	C1, C3	T1,T3	1047	96,85	85,71	91	90
17	C3, C4	T1,T3	1196	96,49	85,71	91	90
18	C1, C2	T2,T3	984	97,66	92,85	90	90
19	C1, C4	T2,T3	996	97,59	89,29	92	90
20	C2, C3	T2,T3	1071	97,48	92,85	88	90
21	C2, C4	T2,T3	1032	98,06	96,43	89	80
22	C1, C3	T2,T3	1019	97,84	92,85	90	100
23	C3, C4	T2,T3	1187	96,97	92,85	91	80
24	C1	T1	46	97,83	35,71	43,48*	20
25	C2	T1	57	89,47	32,14	43,84*	20
26	C3	T1	45	95,56	32,14	35,56*	20
27	C4	T1	55	90,91	32,14	43,64*	20
28	C1	T2	49	91,84	35,71	44,9*	30
29	C2	T2	54	87,04	32,14	48,15*	30
30	C3	T2	47	95,74	32,14	40,43*	30
31	C4	T2	51	96,08	32,14	49,01*	30
32	C1	T3	33	90,91	35,71	42,43*	40
33	C2	T3	34	91,18	32,14	38,24*	40
34	C3	T3	31	90,32	32,14	35,48*	30
35	C4	T3	28	89,29	32,14	30*	30

* Referem-se as intersecções considerando como limite o número de genes em cada cenário, visto que, quando temos n=1, não obtivemos 100 genes diferencialmente expressos em nenhum tratamento.

Além de identificar os genes diferencialmente expressos e ajustar pelo critério de FDR, foi realizado uma ordenação dos genes, considerando o de maior expressão para o de menor expressão. Desse modo foi possível realizar as interseções e comparações com o conjunto original onde haviam 4 repetições biológicas do controle e 3 repetições biológicas do tratamento.

Na Figura 5, observa-se que o número de genes diferencialmente expressos diminui de maneira drástica quando delimitamos somente para uma amostra no controle e uma amostra no tratamento.

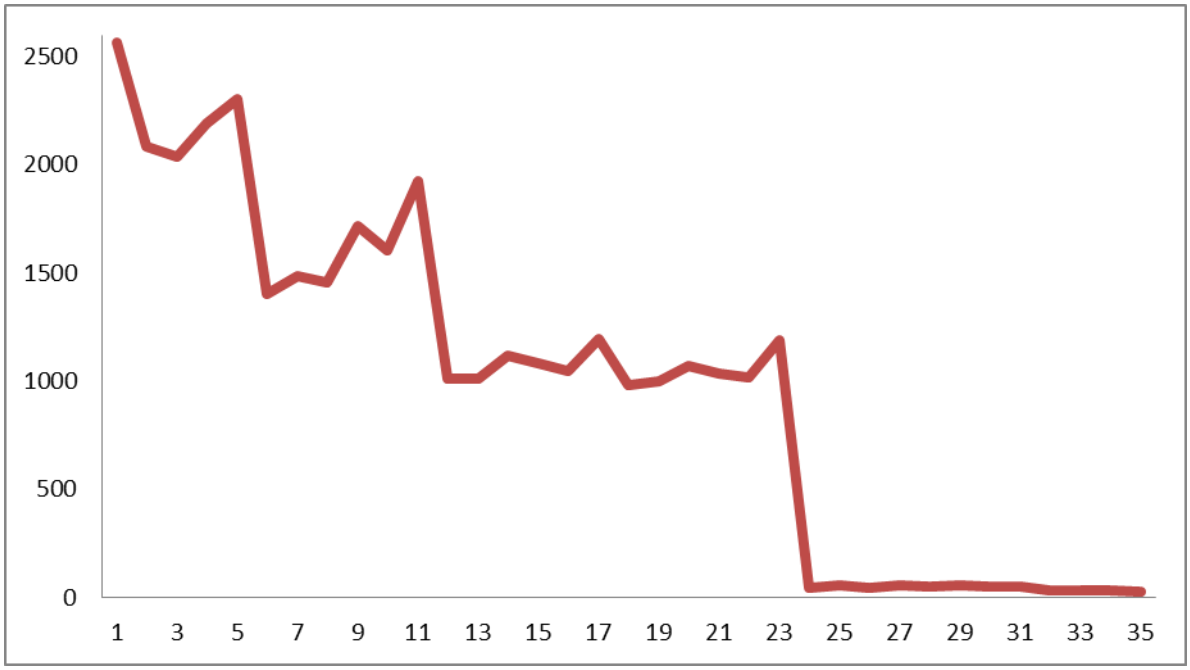

Figura 5 Número de genes diferencialmente expressos em cada cenário.

Quando analisamos quatro repetições biológicas para o controle e três repetições biológicas para tratamento foram identificados 2566 genes DE. Quando diminuímos o número de repetições para três no controle e para três no tratamento encontramos, em média, 2153 genes DE. Considerando os cenários 6 a 23 que utilizavam duas réplicas para as duas condições biológicas encontramos, em média, 1241 genes DE. Uma diminuição drástica no número de genes DE ocorreu quando não foram utilizadas repetições para nenhuma das condições biológicas estudadas. Assim nos cenário de 24 a 35, em média, somente 44 genes diferencialmente expressos foram identificados.

Um estudo realizado por Anders e Huber (2010), avaliou a eficiência do pacote DESeq para análises de dados de RNA-Seq sem repetições em células neurais, no qual conseguiram identificar somente 11% dos genes diferencialmente expressos na condição sem repetição

para nenhum dos grupos, quando comparados aos resultados que utilizavam todas as repetições no experimento. Nesse mesmo estudo avaliando o efeito do pacote DESeq para dados sem repetição em moscas conseguiram identificar 75,09% dos genes considerados como diferencialmente expressos quando comparados ao conjunto de dados que trabalhavam com todas as repetições. Os autores afirmam que apesar do pacote DESeq permitir análises de experimentos biológicos sem repetição, em uma ou ambas as condições, não é possível garantir a confiabilidade do experimento, tornando-se assim duvidoso.

De acordo com Anders (2012), o uso apropriado de repetições é essencial para se interpretar um experimento biológico. Afinal, quando se comparam duas condições biológicas e se encontram diferenças entre elas, a falta de repetição não permite distinguir se essa diferença é devido às diferentes condições, ou apenas devido ao ruído experimental biológico. Assim, apesar dos métodos de análises de RNA-Seq contemplarem a falta de repetição, qualquer tentativa de trabalhar sem quaisquer repetições pode levar a conclusões de confiabilidade limitada.

Outro problema identificado neste estudo é devido a amostragem, mesmo quando se trabalha com o mesmo número de repetições. Observa-se nos cenários 11 e 18 (Tabela 3) que existe uma grande diferença no número de genes DE. O número de genes DE no cenário 11 (1924) é quase uma vez maior do que no cenário 18 (984). Podemos verificar que 89,53% dos genes são iguais nos dois cenários. No entanto, quando ordena-se (em ordem decrescente) os valores de expressão gênica nos dois cenários analisados, percebe-se que 73,17% dos genes são os mesmos, ou seja, os genes alteram na posição do ranking, não mantendo assim o mesmo valor de expressão gênica. Hansen e Irizarry (2012) ressaltam que os protocolos de RNA-Seq para preparação da amostra incluem uma multiplicidade de processos que são sensiveis às condições experimentais. Por exemplo, a extração de RNA, a transcrição reversa e a fragmentação que podem introduzir erros, justificando, assim, a importância das repetições para que ocorra uma diminuição do erro experimental.

De acordo com os resultados apresentados na Tabela 3, observamos que o tratamento 3 pode ter influenciado negativamente o experimento, uma vez que em todos os cenários em que estava presente, o número de genes DE diminui. Essa informação pode ser verificada de forma mais clara nos cenários de 24 a 35. Nos tratamentos 1 e 2 presentes nos cenários de 24 a 31 encontramos em média 50 genes DE, já no tratamento 3 inclusos nos cenários de 32 a 35 a média de genes DE esta em torno de 31 genes DE. A influência do tratamento 3 pode ser visualizada também nos cenários de 6 a 23, onde utilizamos duas repetições para controle e

duas para o tratamento. Nos cenários de 6 a 11, que não utilizamos o tratamento 3, encontramos em média 1597genes considerados DE, já nos cenários 12 a 23, em que o tratamento 3 estava presente, uma média de 1062 genes DE. Esses resultados estão de acordo com estudos realizados por Kvam e Liu (2012), que sugerem que a variabilidade devido a fatores biológicos sob investigação, pode ser maior do que a esperada, o que pode afetar os resultados dos estudos. Isso corrobora mais uma vez sobre o efeito da amostragem e a importância da repetição em experimentos de RNA-Seq no número de genes diferencialmente expressos.

Pode-se notar que quando todos os cenários são comparados com o cenário 1 existe, em média, 95,11% de interseção de genes DE entre eles. Quando é considerada apenas uma amostra para o controle e uma amostra para o tratamento a coincidência entre eles é de 92,18%, valor este considerado alto levando em conta o pequeno número de genes diferencialmente expressos encontrados quando não trabalhamos com repetições. Sendo assim, a porcentagem de encontrar os mesmos genes no cenário original de 2566 genes DE é alta, visto que, mesmo em pequeno número, os genes foram mantidos, não gerando assim pela análise um percentual grande de novos genes diferencialmente expressos.

A partir do momento em que é criado um ponto de interseção entre os mais expressos considerando um limite de 1:28, 1:100 e 1:10, notamos que, em média, 90% dos genes são os mesmos quando consideramos os top 100 e os top 28, do cenário 2 ao cenário 23 (cenários que contêm de 3 a 2 repetições). Porém, quando não trabalhamos com repetição, notamos que a ordem dos genes diferencialmente expressos sofre grandes alterações, e que somente 33% em média dos genes se mantem no ranking.

Devido ao fato de não termos como ordenar os 100 genes mais expressos nos cenários que não tiveram repetições e traçar as interseções com o cenário 1, analisamos os limites de cada cenário. Como exemplo, para o cenário 31, detectou-se 51 genes considerados como diferencialmente expressos. Analisamos as interseções juntamente com o primeiro cenário, incluindo como limite de interseção do 1º ao 51º gene (1:51). Traçados os limites de cada cenário, realizamos as interseções como cenário original e verificamos que houve um acréscimo na porcentagem de interseção do número de genes DE em 91,67% dos cenários que não utilizavam repetições, comparados com os pontos de intersecções de 1:28 e 1:10.

A partir do momento que se estendeu os limites de interseção, a porcentagem de encontrar os mesmos genes no cenário original aumentou (Tabela 3). Quando analisamos as intersecções tendo com limite de 1:28 e 1:10 tínhamos respectivamente em média 33% e 28%

de coincidência com o cenário original, considerando posteriormente o limite de cada cenário, tivemos em média 41,3% de interseção.

Apesar disso, uma atenção tem que ser tomada pelo pesquisador no seguinte sentido: os genes diferencialmente expressos encontrados tanto no cenário com todas as repetições, quanto nos cenários sem repetição, se mantém em torno de 92,18%; contudo a ordem em que eles aparecem quando se listam os mais expressos não se mantém.

Ressaltamos ainda que as repetições são essenciais para alcançar resultados biologicamente interessantes. Seria importante, assim, que todos os experimentos com RNA-Seq incluíssem, em seus planejamentos, pelo menos duas realizações do experimento (repetições biológicas).

5 - CONCLUSÃO

A existência de repetições biológica em experimentos de RNA-Seq é um fator essencial para se obter um número significativo de genes diferencialmente expressos.

Mais de 90% dos genes diferencialmente expressos encontrados nos experimentos sem repetição também foram encontrados nos experimentos com repetição, ou seja, esses genes são bem conservados. Porém a ordem de importância relativa não se mantém.

Caso se opte por realizar experimentos sem repetição, recomenda-se que os pesquisadores trabalhem com todos os genes diferencialmente expressos encontrados, e não apenas com os 10 ou 20 mais expressos.

6 – REFERÊNCIA BIBLIOGRÁFICA

ANDREOTE, F.D. **Análises genômica e transcriptômica de *Methylobacteriummesophilicum SR1.6/6* em interação com a planta hospedeira.** 2011**.** Dissertação (Mestrado em Ciências) Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", 2011.

ANDERS, S.; HÜBER, W. **Differential expression analysis for sequence count data**. Genome Biology 2010, 11:R106.

ANDERS, S. **Analysing high-throughput sequencing data with Python**. http://www.bioinformaticslaboratory.nl/twikidata/pub/Education/BioinformaticsII-Seq/DESeq-tutorial/DESeq-tutorial.pdf, 2012.

AUER, P.L., DOERGE, R.W., **Statistical Design an Analysis of RNA Sequencing Data.** Genetics , 185:405-416, 2010.

BENJAMINI, Y.; HOCHEBERG, Y. **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing**. Journal of the Royal Statistics Society, London, v.57, n.1, p.289-300, 1995.

BULLARD, J., PURDOM, E., HANSEN, K., DURINCK, S. & DUDOIT, S. **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. BMC Bioinformatics11, 94 (2010).

CARVALHO, I.D.E.; SILVA, J.; NOBRE, L.L.M.N.; SILVA, M.T.;TEIXEIRA, J.S.; SILVA, J.W. **Efeito do tamanho de amostra na resistência em milho a spodoptera.** Anais/Resumos da 64ª Reunião Anual da SBPC (**ISSN nº 2176-1221), 2012**

CARVALHO, M.C.C.G.; SILVA, D.C.G. **Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas.** Ciência Rural, Santa Maria, v.40, n.3, p.735-744, mar, 2010.

CASELLA, G.; BERGER, R. **Inferência Estatística** – Cengage Learning, 2010. (Versão em português da 2nd edição em inglês)

CRISTO, E.B. **Métodos estatísticos na Análise de experimentos de *microarray*.** Dissertação (Mestrado em Estatística) Universidade de São Paulo, 2003. Disponível em <http://www.teses.usp.br/teses/disponiveis/45/45133/tde-06062007-112551/pt-br.php> acesso em 16/11/2011.

DIOGENES, F. **Estudo de expressão gênica em citros utilizando modelos lineares. 2009. Dissertação (**Mestrado em Estatística) Universidade de São Paulo, 2009. Disponível em http://pandora.cisc.usp.br/teses/disponiveis/11/11134/te-16032010-111945/publico/Diogenes_Ferreira_Filho.pdf> acesso em 10/11/2011

ESTEVES, G.H. **Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado.** 2007. Tese (Doutorado em Bionformática) – Bioinformática (IME/IFSC/ESALQ/IQ/IB/ICB/FMVZ/FCFRP) – Universidade de São Paulo, São Paulo, 2007.

FARIAS, D.R.; WOYANN, L.G; MAIA, L.C.; OLIVEIRA, A.C. **Análise comparativa de ferramentas de bioinformática para montagem de genomas com tecnologia de sequenciamento de nova geração.** VIX ENPOS, 2012.

GENTLEMAN, R.C.; CAREY, V.J.; BATES, D.M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J.; HORNIK, K.; HOTHORN, T.; HUBER, W.; IACUS, S.; IRIZARRY, R.; LEISCH, F.; LI, C.; MAECHLER, M.; ROSSINI, A.J.; SAWITZKI, G.; SMITH, C.; SMYTH, G.; TIERNEY, L.; YANG, J.Y.H.; ZHANG, J. **Bioconductor: Open software development for computational biology and bioinformatics**. Genome Biol 2004, 5:R80.

HANSEN, K.D.; IRIZARRY, R.A.; WU, Z. **Removing technical variability in RNA-Seq data using conditional quantile normalization.** Biostatistics, 13, 204-216.

JUNQUEIRA, L.C.; CARNEIRO, J. **Biologia Celular e Molecular** .6 ed. Rio de Janeiro: Guanabara Koogan, 1997. 299p.

KVAM , V. M. ; LIU, P.; Y. SI. **A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data**. American Journal of Botany 99 : 248 – 256, 2012.

KESSLER; R.L.; **Transcriptômica De *Trypanosoma cruzi* em resposta a inibidores da síntese de esteróis.** 2010. Dissertação (Mestre em Biologia Celular e Molecular)- Universidade Federal do Paraná, 2007.

LANGMEAD, B.; HANSEN, K.D.; LEEK, J.T.; **Cloud-scale RNA-sequencing differential expression analysis with Myrna**.Genome Biol. 2010; 11 :R83.

LI, H.; LOVCI, M.T.; KWON, Y.S.; ROSENFELD, M.G.; FU, X.D.; YEO, G.W. **Determination of tag density required for digital transcriptomeanalysis: application to an androgen-sensitive prostate cancer model**. Proc. Natlcad. Sci. USA, 105, 20179–20184, 2008.

LOADER, C.; **Local regression, and Likelihod.** Springer; 1999

MARIONI, J.C.; MASON, C.E.; MANE, S.M; STEPHENS, M.; GILAD. Y. **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. Genome Res 18:1509–1517, 2008

MORTAZAVI, A.; WILLIAMS, B.A., MCCUE, K.; SCHAEFFER, L.; WOLD, B.**Mapping and quantifying mammalian transcriptomes by RNA-Seq.** Nat Methods 5:621–628, 2008.

MOROZOVA, O.; HIRST, M.; MARRA, M. A. **Applications of new sequencing technologies for transcriptome analysis**. Annu. Rev. Genomics Hum. Genet. 10: 135–151, 2009.

MCCARTHY, D. J.; CHEN, Y.; SMYTH, G. K. **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.** Nucleic Acids Research, 40, 4288-4297, 2012.

NEVES, C.E. **Experimentos de *microarrays* e teoria da resposta ao item. 2010.** Dissertação (Mestrado em Estatística) Universidade de São Paulo, 2010. Disponível em <http://www.teses.usp.br/teses/disponiveis/45/45133/tde-24052010-140944/fr.php> acesso em 27/11/2011.

NELSON, D. L.; COX, M. **Lehninger – Princípios de Bioquímica**. 3ed. São Paulo: Sarvier, 2002, 1009p.

OSHLACK, A.; ROBINSON, M.D.; YOUNG, M.D**.; From RNA-seq *reads* to differential expression results**. Genome Biol.;11:220, 2010.

ROBINSON, M. D.; OSHLACK, A. **A scaling normalization method for differential expression analysis of RNA-seq data**. Genome biology, v. 11, p. R25, 2010.

ROSA, G.J.M.; ROCHA, L.B. FURLAN, L.R. **Estudos de expressão gênica utilizando-se microarrays: delineamento, análise, e aplicações na pesquisa zootécnica**. Revista Brasileira Zootecnia, vol.36, suppl., pp. 186-209, 2007.

SCHUSTER, I.; CRUZ, C.D. **Estatística Genômica aplicada a populações derivadas de cruzamento controlados.** 2º ed. Viçosa: Editora UFV, 2008. 568p.

SILVA, H. D.; VENCOVSKY, R**. Poder de detecção de "Quantitative Trait Loci", da análise de marcas simples e da regressão linear múltipla**. Scientia agrícola, Piracicaba, vol.59, n.4, 2002. Disponível em : < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162002000400020 >. Acesso em 10/04/2012.

SILVA, J.T. **Genetic transcript analyzer - ferramenta computacional para Análise de transcrição gênica por RNA-Seq. 2012. Dissertação (**Mestrado Bioinformática) Universidade Federal do Paraná

VAN VLIET, A. H. M. **Next generation sequencing of microbial transcriptomes: challenges and opportunities.** FEMS microbiology letters, v. 302, p. 1-7, 2010.

WANG, Z.; GERSTEIN, M.; SNYDER, M. **RNA-seq: A revolutionary tool for transcriptomics.** Nature, v. 10, p.57-63, 2009. Disponívelem: <http://www.nature.com/nrg/journal/v10/n1/abs/ nrg2484.html>. Acesso em: 10/11/2011

WILHELM, B.T.; LANDRY, J. **RNA-Seq—quantitative measurement of expression through massively parallel RNAsequencing**. Methods 48:249–257, 2009

Apêndice I – Glossário

Background: intensidades luminosas captadas ao redor de um *spot*, não representando expressão gênica;

Bibliotecas: conjunto de clones que abrigam cópias, se possível, de todos os RNAs mensageiros de uma linhagem celular ou órgão na forma de cDNA

Eventos epignéticos: são alterações com caráter reversível que não provocam modificações na sequência de DNA, mas sim no fenótipo, exercendo assim, influência nos mecanismos de expressão gênica. O estudo dos eventos epigenéticos fornece informações importantes para o entendimento da carcinogenes**.**

Éxons: é um segmento de bases nitrogenadas de um determinado gene eucarioto que consiste em DNA que codifica para uma sequência de nucleotídeos no RNA mensageiro. Um éxon pode codificar aminoácidos de uma proteína. Geralmente encontra-se adjacente a um segmento de DNA não codificante chamado íntron.

Número de leitura: contagem de fragmentos sequenciados na plataforma

Plataformas de sequenciamento de DNA: são equipamentos capazes de gerar informação sobre milhões de pares de bases em uma única corrida.

Sequenciamento de DNA: O sequenciamento de DNA são métodos bioquímicos que têm como finalidade determinar a ordem das bases nitrogenadasadenina (A), guanina (G), citosina (C) e timina (T) da molécula de DNA.

Splicing: é um processo que remove os *íntrons* e junta os *éxons* depois da transcrição do RNA. Ele consiste na retirada dos *íntrons* de um mRNA precursor, sendo um dos processos necessários para formar um mRNA maduro funcional. Essa excisão dos *íntrons* do mRNA é um evento muito importante e requer uma extrema precisão das moléculas envolvidas no processo. A exclusão ou o acréscimo de um único nucleotídeo em um *éxon* pode levar a uma

alteração da fase de leitura e à produção de uma proteína completamente diferente da original ou defeituosa.

Transcriptoma: Coleção de RNAs (transcritos) presentes em uma célula/tecido em um dado momento. O transcriptoma corresponde a fração do código genético (DNA) que e transcrita pela RNA polimerase em moléculas de RNA.